

# **A Preference-based Relevance Feedback Approach for Polyrepresentative Multimedia Retrieval**

Von der Fakultät für Mathematik, Naturwissenschaften und Informatik  
der Brandenburgischen Technischen Universität Cottbus-Senftenberg

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften  
(Dr.-Ing.)

genehmigte Dissertation  
vorgelegt von

Diplom-Informatiker (FH) Master of Science (U)

David Zellhöfer

geboren am 20.09.1979 in Berlin

Gutachter: Prof. Dr.-Ing. habil. Ingo Schmitt

Gutachter: Prof. Dr. Birger Larsen

Gutachter: Prof. em. Dr. Dr. h.c. mult. Peter Ingwersen

Tag der mündlichen Prüfung: 02.06.2015





# Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Ingo Schmitt for his guidance, support, and the freedom I could experience during my time at the Brandenburg University of Technology. In particular, I would like to thank Prof. Schmitt for his patience during the discussions on philosophical aspects of my work and its relation to mathematics. The resulting conflicts fueled my creativity to revise my ideas and to approach my research problems from different angles. Somehow, these discussions also awakened my interest in statistics.

Particular thanks to Prof. Birger Larsen and Prof. Peter Ingwersen for hosting me at the Royal School of Library and Information Science in Copenhagen and providing valuable and diverse feedback at early stages of my work. Furthermore, I would like to thank Prof. Christina Lioma for her early feedback on evaluation methods and the hospitality that allowed a vivid discussion with other researchers. In addition, the people in Copenhagen have been very supportive, in particular Dr. Toine Bogers.

Regarding my experiments, I am also very much in debt to Dr. Theodora Tsikrika, who never stopped asking critical questions about user-centered evaluations in the scope of ImageCLEF which helped to increase my understanding of my own research questions.

A big thank you to my former colleagues: Thomas Böttcher who was always willing to run “just another small experiment” at impossible times with me and for surviving my rants about the separation of the DB, IR, and information science communities; Christoph Schmidt, Claudius Tillmann, Markus Uhlig, and Maria Bertram whose knowledge in Greek mythology led to the name “Pythia”. I would also like to thank Sebastian Lehrack for being there during the first lonely years in Cottbus and the latter quantum mechanics-supported train rides to Berlin.

I am indebted to the staff at Cottbus, particularly Ekkehard Schwaar who kept the computer labs alive and reseted the servers numerous times during weekends or when I was working remotely from a more or less tropical place.

I would also like to thank some students whose work I supervised: Bianca Böckelmann, Sebastian Zech, Markus Uhlig, and Ralf Janiak as well as all other student workers at my workgroup.

Further thanks go to Prof. Norbert Fuhr whose talk motivated me to read about cognitive retrieval models such as the principle of polyrepresentation, Prof. Michael A. Herzog without whom I would have never met my supervisor, Prof. Debora Weber-Wulff for her constant support and motivation to finish the thesis, and Prof. Kai Uwe Barthel for starting my interest in image retrieval.

Although this text represents my own work, many people have contributed to the completion of this dissertation, directly or indirectly. I wish to thank all my friends,

colleagues, and reviewers, regardless if they rejected or accepted my papers, for their discussion and support.

I would like to thank Dr. Chris Häusler and Martina Baltkalne for proof-reading my dissertation.

Although there is hardly enough space to express my gratefulness, I would like to thank Diana for her sheer endless patience with me and my dissertation. Diana, I owe you so much.

And finally, I would like to thank my mother and Gunther for supporting my studies, once, twice, or maybe three times. I guess, I am finally ready to quit university.

Gunther, letztendlich tragen ein Commodore PC30-III und Du Schuld an dem ganzen. Ich hoffe, dass Dir das bewusst ist.

Danke, Hühnerlord.

# Abstracts

## Zusammenfassung

Die Suche nach textueller Information, z.B. in Form von Webseiten, stellt eine typische Aufgabe im modernen Geschäfts- und Privatleben dar. Aus Nutzersicht sind die hierfür verwendeten Systeme gereift und es haben sich typische Interaktionsmechanismen, wie die Texteingabeaufforderung, welche beinahe jede gerichtete Suche startet, herausgebildet.

Im Vergleich dazu befindet sich die Suche nach multimedialen Dokumenten (wie Bildern oder Videos) aus Nutzersicht noch in den sogenannten Kinderschuhen, obwohl dieser Anwendungsbereich konstant an Bedeutung zunimmt. In diesem Bereich kämpfen gerichtete und explorative Suchansätze nach wie vor um Nutzerakzeptanz.

Ein weiterer Punkt, der Multimedia Information Retrieval (MMIR) vom traditionellen, Text-basierten Information Retrieval (IR) unterscheidet, ist der Umstand, dass Multimediadokumente nicht notwendigerweise mittels des gleichen Datenzugriffsparadigmas gespeichert sind. So können Daten, welche die Dokumente intern repräsentieren, beispielsweise in Datenbank- (DB) oder IR-Systemen vorgehalten werden. Aus technischer Sicht kompliziert dieser Zustand die Suche in solchen Datenbeständen, da das verwendende Retrieval-Modell die jeweiligen Datenzugriffsparadigmen unterstützen muss.

Konsequenterweise müssen die Hauptherausforderungen des MMIR, das Retrieval-Subsystem und die Nutzerinteraktion, ganzheitlich angegangen werden. Dies ist notwendig, da diese Kernbereiche des Retrievals nicht separiert werden können, wenn man die Suche nach Multimediadokumenten als Anwender-zentrierten Prozess begreifen will.

Solch einen ganzheitlichen, theoretischen Blick auf die MMIR/IR-Forschung wirft das Prinzip der Polyrepräsentation (PdP), das eine Hälfte des theoretischen Hintergrunds dieser Dissertation darstellt. Ziel der Arbeit ist es, einen Präferenz-basierten Ansatz für interaktives MMIR zu entwickeln. Vereinfachend gesprochen geht das PdP davon aus, dass verschiedene Repräsentationen, die ein Dokument beschreiben, auf unterschiedlichen, kognitiven Prozessen basieren, die sich z.B. in dessen Titel, dessen Farb- und Formgebung oder Erstellungsdatum manifestieren. Diese Vielzahl an Repräsentationen kann in einer konjunktiven, kognitiven Überlappung (KÜ) zusammengeführt werden. Hierbei wird angenommen, dass relevante Dokumente in diesem Bereich liegen. Diese explizite Annahme unterscheidet das PdP von anderen Feature-Fusionsansätzen, die sonst häufig beim MMIR Verwendung finden.

Das PdP als kognitives Modell trifft jedoch keine Aussagen darüber, wie ein entsprechendes Retrieval-Subsystem implementiert werden kann – eine wesentliche Fragestel-

lung der angewandten Informatik. Eine Möglichkeit der Implementierung des PdP stellen Quantenmechanik-basierte IR-Modelle, wie die Commuting Quantum Query Language (CQQL), dar, welche auch im Rahmen dieser Dissertation verwendet wird.

Die Wahl fällt hierbei auf CQQL, da sie typische Datenzugriffsparadigmen aus den Gebieten DB und IR unterstützt. Dieser Aspekt ist von großer Bedeutung, um Synergieeffekte zwischen verschiedenen Medien, oder – anders gesprochen – die polyrepräsentative Natur von Multimediadokumenten, ausnutzen zu können. Außerdem erlaubt CQQL die Personalisierung von Suchergebnissen mittels des Präferenz-basierten Relevance-Feedback-Ansatzes (RF) PrefCQQL, welcher auf maschinellen Lernverfahren basiert. Diese Funktionalität ist notwendig, um der dynamischen Natur des Suchprozesses und des Informationsbedürfnisses (IB) des Nutzers gerecht zu werden.

Alleinstellungsmerkmale von PrefCQQL stellen dabei die Unterstützung von negativen Query-by-Example-Dokumenten (QBE) sowohl zu Beginn als auch während der Suche dar. Außerdem können induktive, schwache Präferenzen genutzt werden, um die Relevanz von Ergebnisdokumenten feingranular zu bewerten. Ferner können induktive Präferenzen bereits zur Anfrageerstellung genutzt werden, so dass sich im Prinzip sämtliche Nutzerinteraktion während der Suche mit PrefCQQL umsetzen lässt.

Zur Evaluierung des vorgestellten, polyrepräsentativen PrefCQQL-Ansatzes werden zwei Arten von Experimenten durchgeführt: eine Evaluierung der Retrieval-Effektivität von CQQL/PrefCQQL auf Grundlage des Cranfield-Paradigmas, welches um Nutzersimulationen erweitert wurde, um den Anforderungen an die Bewertung von interaktiven MMIR-Systemen gerecht zu werden; und eine Usability-Studie, welche drei funktional unterschiedliche Benutzerschnittstellen eines MMIR-Systems vergleicht. Hierbei wurden sowohl die Arbeitsaufgabe, die initiale Anfrage als auch das Retrieval-Subsystem fixiert, um eine Vergleichbarkeit zwischen den GUI-Varianten zu gewährleisten. Um die Reproduzierbarkeit und Nachvollziehbarkeit der Experimente sicherzustellen, ist der Quellcode sämtlicher Anwendungen im Anhang zu dieser Dissertation zu finden.

Die beschriebenen Experimente sollen im wesentlichen zwei zentrale Fragen beantworten: erstens, ob die Hypothesen des PdP im Anwendungsgebiet des MMIR verifizierbar sind und zweitens, ob ein nutzbares, interaktives MMIR-System auf Grundlage des PdP und von PrefCQQL implementiert werden kann?

Um die erste Frage beantworten zu können, werden verschiedene Matching-Funktionen, die teilweise dem PdP folgen, mit sechs unterschiedlichen Testdatensätzen, sowohl in einem nicht interaktiven, als auch interaktiven QBE-Szenario gegenübergestellt. Die Ergebnisse dieses Experiments sind uneindeutig.

Im nicht interaktiven Fall, d.h. wenn kein RF zur Verfügung steht, zeigt sich keine Überlegenheit der Matching-Funktionen, welche auf dem PdP basieren, gegenüber der Nutzung einzelner Features oder anderen Matching-Funktionen. So übertrifft die Retrieval-Effektivität des arithmetischen Mittels, welches die gemittelte Ähnlichkeit zwischen den Repräsentationen der Anfrage und denen der Dokumente im Testdatensatz berechnet, die der Konjunktion der vorhandenen Repräsentationen, welche der KÜ entspricht. Nichtsdestotrotz zeigt sich, dass Matching-Funktionen, welche dem PdP folgen, stabiler bezüglich ihrer Effektivität sind als Einzelfeatures. Folglich ist die Leis-

tung der PdP-basierten Funktionen verlässlicher. In jedem Fall wird deutlich, dass die Konjunktion besser als alle anderen fusionsbasierten Matching-Funktionen, außer dem angesprochenen arithmetischen Mittel, abschneidet. Außerdem legt die experimentelle Untersuchung nahe, dass konjunktive Matching-Funktionen immer ein höheres Maß an Retrieval-Effektivität erreichen als ihr disjunktives Gegenstück. Dieses Ergebnis entspricht den Voraussagen des PdP.

Im interaktiven Szenario, also während der Verwendung von PrefCQQL, können die Voraussagen des PdP verifiziert werden. Hierbei muss allerdings angemerkt werden, dass ebenfalls die Anzahl an verfügbaren Repräsentationen innerhalb einer Matching-Funktion einen Einfluss auf deren Retrieval-Effektivität hat. Sind zu wenige Repräsentationen zur Modellierung des IB mittels PrefCQQL vorhanden, kommt es zu Unteranpassungseffekten, welche die Retrievalleistung negativ beeinflussen.

Die zweite Frage wird mithilfe eines prototypischen MMIR-Systems beantwortet, dem Pythia-System, welches als Machbarkeitsstudie eines interaktiven Retrieval-Systems auf Basis von CQQL und PrefCQQL dient. Des Weiteren unterstützt das System unterschiedliche Suchstrategien sowie einen nahtlosen Wechsel zwischen diesen, um Nutzer mit verschiedenen IB bestmöglich zu unterstützen. Der Test der drei GUI-Varianten des Systems mithilfe von 59 Probanden attestieren dem Pythia-System ein hohes Maß an Bedienbarkeit.

Letztendlich zielen die beschriebenen Experimente darauf ab, die Aussagen des PdP zu verifizieren. Diese Verifizierung ist notwendig, um abschließend zu bestimmen, ob es sich beim PdP um eine gültige Theorie des MMIR/IR handelt. Die Beantwortung dieser Frage ist insbesondere interessant, da Experimente im Bereich des textuellen IR die Aussagen des Prinzips stützen. Obwohl die Experimente im Rahmen dieser Dissertation so gestaltet wurden, dass sich deren Ergebnisse mit einer hohen Wahrscheinlichkeit generalisieren lassen, kann diese Dissertation die Frage, ob es sich beim PdP um eine Theorie handelt, nicht abschließend klären. Nichtsdestotrotz bieten die präsentierten Forschungsergebnisse genügend Anzeichen dafür, dass das PdP einen generellen Nutzen für das MMIR bietet, da es definitiv Voraussagen über die Entwicklung der Retrieval-Effektivität einer Matching-Funktion ermöglicht. Außerdem zeigen die Experimente deutlich, dass PrefCQQL eine nützliche RF-Technik für das MMIR darstellt.

### **Hinweis zur geschlechterneutralen Formulierung**

Die Nutzung männlicher Wortformen in dieser Dissertation dient ausschließlich der einfacheren Lesbarkeit dieses Texts und beinhaltet sämtliche Geschlechter.

## Abstract

The search for textual information, e.g., in the form of webpages, is a typical task in modern business and private life. From a user's point of view, the commonly used systems have matured and established common interaction design patterns such as the textual input box that starts virtually every directed search process.

In comparison, although constantly gaining importance, the search for multimedia documents (e.g., images or videos) is still, at least from an end-user's perspective, in its early years. In other words, a pre-dominant search strategy has not yet evolved. That is, directed and exploratory search approaches fight for user acceptance.

One further discriminative factor of multimedia information retrieval (MMIR) from traditional text-based information retrieval (IR) is that multimedia documents are not necessarily stored with the help of the same data access paradigm. For instance, data representing multimedia documents can be stored in databases (DB) or IR systems. From a technical point of view, the use of different data access paradigms complicates the retrieval from such collections because the utilized retrieval model has to support these paradigms.

As a consequence, the main challenges in MMIR – the retrieval engine and the user interaction – have to be addressed in a holistic way because they can hardly be separated if the search for multimedia information is recognized as a user-centered process.

A holistic theoretic perspective on MMIR/IR research is taken by principle of polyrepresentation (PoP), which forms one half of the theoretic background of this dissertation aiming at the development of a preference-based approach to interactive MMIR. Roughly speaking, the PoP theorizes that representations describing a document are based on various cognitive processes dealing with it, e.g., a title, its color or shape features, its creator, or its date of creation. This multitude of representations can be fused to form a conjunctive cognitive overlap (CO) in which highly relevant documents are likely to be contained. This explicit recommendation discriminates the PoP from typical feature fusion approaches often used in MMIR.

However, the PoP does not answer how a retrieval model has to be implemented in a technical sense which is of interest in the field of computer science. One possibility to implement the PoP are quantum mechanics-inspired IR models such as the commuting quantum query language (CQQL) which is used in this thesis.

CQQL is particularly interesting because it integrates data access paradigms used in the fields of DB and IR, which is crucial for exploiting synergies between various media, or, in other words, for exploiting the polyrepresentative nature of multimedia documents. In order to respect the dynamic nature of the search process and information need (IN), CQQL allows the personalization of retrieval results using a preference-based relevance feedback (RF) approach called PrefCQQL, which relies on machine-based learning.

Unique features of the PrefCQQL approach range from the support of negative query-by-example (QBE) documents at query formulation time as well as during the interactive retrieval process to the formulation of weak preferences between result documents to express gradual levels of relevance. In addition, the so-called inductive preferences

can be used from query formulation time onward to learn new CQQL queries. As a consequence, the general user interaction can be based, in principle, on inductive preferences alone.

In order to evaluate the presented polyrepresentative PrefCQQL approach, two kinds of experiments are conducted: a Cranfield-inspired evaluation of CQQL/PrefCQQL's retrieval effectiveness, which is extended by the utilization of user simulations to better fit the requirements of the evaluation of an adaptive IR system, and a usability study that examines three alternative MMIR system UI prototypes based on the same simulated work task, query, and retrieval model to ensure comparability between the variants. In order to increase the reproducibility and confirmability of the experiments, the source code to all used programs is made available as a supplement to this dissertation.

The mentioned experiments aim at answering two central questions: first, whether the hypotheses of the PoP can be verified in MMIR, and second, whether a usable interactive MMIR system can be built on the basis of the PoP and PrefCQQL?

To answer the first question, different matching functions that partly follow the recommendations of the PoP are evaluated with six different test collections in both a non-interactive and interactive QBE scenario. The results of this experiment are ambivalent.

In non-interactive MMIR, i.e., when no RF is given, the experimental data does not provide sufficient justification for the statement that PoP-based matching functions will always surpass single features or other matching functions. For instance, the arithmetic mean, which calculates the average similarity between a query's representations and the documents' representations in the collection, surpasses the conjunction and hence the CO of multiple representations in terms of retrieval effectiveness. Nevertheless, the matching function following the PoP is effectiveness stabler than the best performing single representations per collection. Hence, the CO's retrieval performance is more reliable than the usage of single representations. In any case, the conjunction still performs better than any other examined fusion-based matching function, besides the arithmetic mean. Moreover, there is evidence that the conjunctive variant of a matching function always performs better than its disjunctive counterpart. This finding complies with the predictions of the PoP.

In contrast, the predictions of the PoP can be verified in the investigated PrefCQQL-based interactive MMIR scenario. However, it is important to note that also the number of available representations has an impact on the retrieval outcome. That is, if too few representations are present in a matching function, the corresponding IN model in PrefCQQL obviously becomes subject to underfitting eventually lowering its retrieval effectiveness. Unfortunately, when the point of sufficient representations to support PrefCQQL is reached could not be revealed in this dissertation.

The second question is answered with the help of a prototypical MMIR system: the Pythia system, which serves both as proof of concept of the CQQL and the PrefCQQL approach. Furthermore, the system supports different information seeking strategies and a seamless transition between them in order to support users with different kinds of IN. Different UI variants of the system are tested by 59 subjects, which attest the Pythia system a high level of usability.

To come to an end, the objective of the described experimentation is to justify a theoretic model's validity, i.e., the validity of the predictions made by the PoP. This verification process is required to assess the principle's validity and to investigate whether the PoP forms a theory. The answer to this question is particularly interesting in conjunction to the experiments already conducted in textual IR, which support the PoP. However, although the experiments presented in this thesis were conducted in a way that allows a generalization of the resulting conclusions, we believe that this dissertation cannot answer this question exhaustively. Nevertheless, the presented research results provide sufficient evidence of the PoP's general utility in MMIR as it definitely allows predictions of the retrieval effectiveness development of a matching function. Moreover, the experiments clearly show that PrefCQQL is usable as a RF technique in MMIR.

### **Note on Gender Neutrality**

The use of the masculine form of words in this dissertation is only intended to lighten the text and encompasses both genders.



# Table of Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	2
1.2	Research Questions and Contributions . . . . .	3
1.2.1	Can the Hypotheses of the Principle of Polyrepresentation Be Verified in Multimedia Information Retrieval? . . . . .	3
1.2.2	Do the Hypotheses of the Principle of Polyrepresentation Hold in a Preference-based Interactive Multimedia Search Process? . . . . .	5
1.2.3	Can a Usable Multimedia Information Retrieval System Be Built on the Basis of the Principle of Polyrepresentation and PrefCQQL? . . . . .	5
1.2.4	Positioning of the Dissertation . . . . .	6
1.3	Structure of the Dissertation . . . . .	7
<b>I</b>	<b>Foundations and Background</b>	<b>9</b>
<b>2</b>	<b>Multimedia Information Retrieval</b>	<b>11</b>
2.1	Media and Multimedia . . . . .	11
2.2	Information Retrieval . . . . .	12
2.2.1	Principal Concepts in Information Retrieval . . . . .	12
2.2.2	Information Retrieval Models . . . . .	15
2.3	Principles of Multimedia Information Retrieval . . . . .	22
2.3.1	Features and Feature Extraction . . . . .	23
2.3.2	Distance and Similarity Measures . . . . .	26
2.3.3	The Semantic Gap in Multimedia Retrieval . . . . .	26
<b>3</b>	<b>Interactive Multimedia Information Retrieval</b>	<b>31</b>
3.1	Interactive Information Retrieval . . . . .	31
3.1.1	The User-oriented Viewpoint . . . . .	32
3.1.2	The Cognitive Viewpoint . . . . .	33
3.2	The Principle of Polyrepresentation . . . . .	35
3.2.1	The Polyrepresentation Continuum . . . . .	37
3.3	Interactive Support of Information Seeking Strategies . . . . .	38
3.3.1	Directed Search and Query-By Approaches . . . . .	38
3.3.2	Exploratory Search . . . . .	40

<b>II</b>	<b>Learning User-specific Weights in Logic-based Queries</b>	<b>45</b>
<b>4</b>	<b>A Quantum Logic-based Model for Multimedia Information Retrieval</b>	<b>47</b>
4.1	Theoretical Concepts behind the Commuting Quantum Query Language	47
4.2	Construction and Arithmetic Evaluation of CQQL Queries . . . . .	52
4.3	Integration and Evaluation of Weights in CQQL . . . . .	54
4.4	CQQL as a Multimodal Logical Query Language . . . . .	56
4.4.1	Advantages of Logic-based Multimodal Retrieval . . . . .	59
4.4.2	Issues of Logic-based Multimodal Retrieval . . . . .	63
4.5	The Relation of the Quantum Logic-based Approach to other IR Models	64
4.5.1	Vector Space Model . . . . .	64
4.5.2	Probabilistic Models . . . . .	64
4.5.3	Logic-based Models . . . . .	65
4.5.4	Quantum Mechanics-inspired Models . . . . .	68
4.5.5	Classification of CQQL . . . . .	71
4.6	CQQL as an Implementation of the Principle of Polyrepresentation . . .	73
4.7	The Relation of the CQQL Approach to Other Formalizations of Polyrepresentation . . . . .	76
<b>5</b>	<b>Machine-based Learning of Personalized CQQL Queries</b>	<b>79</b>
5.1	Personalization in Information Retrieval . . . . .	80
5.1.1	Explicit Relevance Feedback . . . . .	80
5.1.2	Implicit Relevance Feedback . . . . .	81
5.1.3	Weighting of Query Parts . . . . .	81
5.2	Learning to Rank . . . . .	82
5.3	Preference Models for the Personalization of Queries . . . . .	84
5.3.1	Qualitative Preferences . . . . .	87
5.3.2	Quantitative Preferences . . . . .	92
5.3.3	Comparison of Qualitative and Quantitative Approaches . . . . .	96
5.4	PrefCQQL – Preferences within the CQQL-based Retrieval Model . . . . .	99
5.4.1	Preference Reasoning: Deduction and Induction . . . . .	102
5.4.2	A User-centered Preference Model for MIR . . . . .	105
5.4.3	Interpreting Preferences as Partially Ordered Sets . . . . .	107
5.4.4	The Learning of Weights as a Non-Linear Optimization Problem	111
5.4.5	Adoption of the Nelder-Mead Downhill Simplex Algorithm . . .	115
5.4.6	Runtime of the Learning Algorithm . . . . .	119
5.4.7	Preference Conflicts and Query Incompatibilities . . . . .	120
5.4.8	Reduction of a Rank . . . . .	123
5.5	Learning CQQL Queries as a Special Case of Weight Learning . . . . .	125
5.6	The Relation of the PrefCQQL Approach to Other Personalization Techniques in Information Retrieval . . . . .	128

<b>III</b>	<b>Concept and Implementation</b>	<b>131</b>
<b>6</b>	<b>Design of the Pythia MIR System Prototype</b>	<b>133</b>
6.1	The User-centered Design Principle . . . . .	134
6.1.1	Personas and Scenarios . . . . .	136
6.1.2	User Stories . . . . .	140
6.1.3	Iterative User Feedback . . . . .	142
6.1.4	General User Experience Objectives . . . . .	143
6.1.5	Summary . . . . .	149
6.2	Conceptual Model . . . . .	149
6.2.1	Theoretical Foundations of a Polyrepresentative User Interaction Model . . . . .	150
6.2.2	Conceptual Model of the Pythia MIR System's User Interaction . . . . .	157
6.2.3	Result Visualization and Organization . . . . .	166
6.2.4	Directed and Exploratory Search . . . . .	173
6.2.5	PrefCQQL – Preference Elicitation and Relevance Feedback . . . . .	176
6.2.6	Faceted Navigation . . . . .	182
6.2.7	Search History . . . . .	183
6.2.8	Manual Weight Setting – Advanced Search . . . . .	185
6.2.9	Result Explanation – Advanced Search . . . . .	185
6.3	Prototypical Implementation . . . . .	188
6.3.1	Basic System Architecture . . . . .	189
6.3.2	Overview of the User Interface and Implemented Functions . . . . .	191
6.4	The Relation of Pythia MIR to other Interactive MIR Systems . . . . .	194
<b>IV</b>	<b>Evaluation</b>	<b>199</b>
<b>7</b>	<b>General Considerations</b>	<b>201</b>
7.1	The Traditional Cranfield Approach . . . . .	202
7.2	Evaluation of Adaptive Retrieval Systems . . . . .	204
<b>8</b>	<b>Evaluation of the Retrieval Effectiveness</b>	<b>207</b>
8.1	Comparing the Effectiveness of IR Systems . . . . .	207
8.1.1	System-Centric Retrieval Metrics . . . . .	208
8.1.2	User-Centric Retrieval Metrics . . . . .	209
8.1.3	A Brief Critique of the Rank-based Measures . . . . .	212
8.1.4	Testing Statistical Significance . . . . .	212
8.2	Overview of Cranfield-based Test Collections . . . . .	213
8.2.1	Caltech 101 and 256 . . . . .	213
8.2.2	MSRA-MM . . . . .	214
8.2.3	UCID v2 . . . . .	215
8.2.4	Wang . . . . .	215
8.2.5	Summary . . . . .	215

## Table of Contents

8.3	Design of a Test Collection for Adaptive Retrieval Systems . . . . .	216
8.3.1	Creation and Origin of the Pythia Collection . . . . .	217
8.3.2	Characteristics of the Pythia Collection . . . . .	218
8.3.3	Obtainment of the Ground Truth for the Pythia Collection . . . . .	219
8.3.4	Support for User Simulations . . . . .	221
8.3.5	The Collection in Relation to Adaptive IR Evaluation . . . . .	223
8.4	Retrieval Effectiveness of the CQQL Approach . . . . .	224
8.4.1	Experimental Setup for Retrieval Effectiveness Evaluation . . . . .	225
8.4.2	Evaluation of Single Representations . . . . .	229
8.4.3	Evaluation of Combined Representations – Main Experiment I . . . . .	235
8.5	Retrieval Effectiveness of PrefCQQL . . . . .	254
8.5.1	Experimental Setup for User Simulation and Relevance Feedback . . . . .	254
8.5.2	Typical Relevance Feedback Performance – Main Experiment II . . . . .	256
8.5.3	Long-Enduring Relevance Feedback Performance . . . . .	270
8.5.4	Development of the Weighting Variables . . . . .	272
8.5.5	Query Incompatibilities and Preference Poset Size . . . . .	280
<b>9</b>	<b>Evaluation of the User Experience</b>	<b>285</b>
9.1	Experimental Setup . . . . .	286
9.1.1	Restrictions of the Retrieval Engine . . . . .	287
9.1.2	Introductory Phase . . . . .	288
9.1.3	GUI Variant T2 . . . . .	288
9.1.4	GUI Variant T3 . . . . .	290
9.1.5	GUI Variant T4 . . . . .	290
9.1.6	Summary . . . . .	291
9.2	Demographics and Computer Usage of the Test Persons . . . . .	292
9.3	Results of the Quantitative Usability Study . . . . .	295
9.3.1	Validity of the Results . . . . .	305
9.4	Results of the Qualitative User Study . . . . .	307
9.4.1	Results of the Open Question . . . . .	308
9.5	Supplementary User Observations . . . . .	309
9.6	Discussion of the General User Experience . . . . .	314
<b>V</b>	<b>Conclusions and Outlook</b>	<b>319</b>
<b>10</b>	<b>Discussion and Conclusions</b>	<b>321</b>
10.1	Discussion . . . . .	321
10.1.1	User Experience . . . . .	322
10.1.2	Retrieval Effectiveness and Further Properties of PrefCQQL . . . . .	324
10.2	Conclusions . . . . .	327
10.2.1	Can the Hypotheses of the Principle of Polyrepresentation Be Verified in Multimedia Information Retrieval? . . . . .	327

10.2.2 Do the Hypotheses of the Principle of Polyrepresentation Hold in a Preference-based Interactive Multimedia Search Process? . . . 328

10.2.3 Can a Usable Multimedia Information Retrieval System Be Built on the Basis of the Principle of Polyrepresentation and PrefCQQL? 329

10.3 Future Work . . . . . 330

10.3.1 Evaluation Opportunities . . . . . 330

10.3.2 Usability Improvements . . . . . 332

10.3.3 Functional Extensions . . . . . 332

10.3.4 Technical Improvements . . . . . 333

10.3.5 Future Directions and Application Areas . . . . . 333

**VI Appendices 335**

**A Mathematical Foundations 337**

A.1 Basics . . . . . 337

A.1.1 Partially Ordered Sets (Posets) . . . . . 337

A.1.2 Total Order . . . . . 337

A.1.3 Lattices . . . . . 337

A.1.4 What is Logic? . . . . . 338

A.1.5 Boolean Algebra . . . . . 339

A.1.6 Dirac notation . . . . . 340

A.2 Probability Theory . . . . . 340

A.2.1 Basic Terminology . . . . . 340

A.2.2 Event Algebra . . . . . 341

A.2.3 Material Implication vs. Implication as Conditional Probability . 342

A.2.4 Bayes' Theorem . . . . . 342

**B Evaluation Appendix 345**

B.1 Formulae of the Examined Matching Functions . . . . . 345

B.1.1 Matching Functions Based on CQQL . . . . . 346

B.1.2 Matching Functions Based Standard Aggregations . . . . . 349

B.2 Retrieval Effectiveness Comparisons of Different Matching Functions . . 352

B.3 Weight Development of Different Matching Functions . . . . . 363

B.3.1 Averaged Weighting Variable Development per Collection . . . . . 363

B.3.2 Weighting Variable Development and Distribution during Relevance Feedback . . . . . 369

B.3.3 Correlation Analyses . . . . . 406

B.4 Results of the Usability Study . . . . . 412

B.4.1 Results for the T2 GUI Variant . . . . . 414

B.4.2 Results for the T3 GUI Variant . . . . . 425

B.4.3 Results for the T4 GUI Variant . . . . . 435

B.5 Miscellaneous . . . . . 446

B.5.1 Distribution of the Random Number Generator . . . . . 446

Table of Contents

<b>C</b>	<b>Questionnaires and Interviews</b>	<b>447</b>
C.1	Demographics and Usage Questionnaire . . . . .	447
C.2	Demographics of the Contributors and Assessors . . . . .	448
C.3	Materials Used in the Usability Test . . . . .	451
C.3.1	Demographics and Usage Questionnaire . . . . .	451
C.3.2	Usability Test Instructions . . . . .	453
<b>D</b>	<b>Sample User Interaction with the Pythia MIR System</b>	<b>455</b>
<b>E</b>	<b>Contents of the Enclosed DVD</b>	<b>459</b>
<b>F</b>	<b>List of Publications</b>	<b>461</b>
	<b>Bibliography</b>	<b>465</b>
	<b>List of Figures</b>	<b>495</b>
	<b>List of Tables</b>	<b>503</b>
	<b>List of Definitions</b>	<b>507</b>
	<b>List of Theorems, Lemmata, and Proofs</b>	<b>511</b>
	<b>List of Matching Functions</b>	<b>513</b>
	<b>List of User Stories</b>	<b>515</b>
	<b>Listings</b>	<b>517</b>







# 1 Introduction

The search for information, e.g., in the form of webpages, is a typical task in modern business and private life. The usage of so-called textual information retrieval (IR) systems is pervasive which can easily be shown by the dominant position in the Internet taken by websites such as Google and Bing. From a user's point of view, these systems have matured and established common interaction design patterns such as the textual input box that starts virtually every search process.

In comparison, although constantly gaining importance, the search for multimedia documents (e.g., images or videos) is still, at least from an end-user's perspective, in its infancy – or teenage years if one takes the development of the related scientific field of multimedia information retrieval (MMIR) into account [Smeulders et al. 2000].

Although a famous proverb claims that a picture is worth a thousand words, the search for such documents is still predominantly text-based. That is, tags, annotations, or other texts related to a document are used to find relevant documents. Although some end-user MMIR systems already support techniques such as face detection or alternative content-based information retrieval (CBIR) techniques, other large websites featuring multimedia content, such as the Wikipedia<sup>1</sup>, Wikimedia Commons<sup>2</sup>, or YouTube<sup>3</sup>, retrieve documents primarily on the basis of textual information.

As implied above, the user interfaces of CBIR or MMIR systems are not as mature as the aforementioned prevalent directed search paradigm in end-user IR systems [Hearst 2009, cf. Sec. 12.2]. This effect is most likely due to the different search strategies applied to search in multimedia document collections. In fact, studies indicate that the exploratory search paradigm is highly important in CBIR/MMIR [McDonald & Tait 2003; Rodden & Wood 2003; Cunningham & Masoodian 2006]. In other words, users like to browse the contents of multimedia document collections, e.g., for recreational purposes, in conjunction to submitting directed searches based on keywords, which is typical for IR.

One further discriminative factor of CBIR/MMIR from traditional IR is that multimedia documents are not necessarily stored with the help of the same data access paradigm. Data representing multimedia documents can be stored in databases (DB) or IR systems. For instance, the visual parts of a multimedia document can be stored in a CBIR system, whereas metadata about the document, e.g., copyright information or the document's file size, may be stored in a relational database. The decision to store this data in different systems can be based on technical reasons, e.g., processing

---

<sup>1</sup><http://en.wikipedia.org/>

<sup>2</sup><http://commons.wikimedia.org/>

<sup>3</sup><http://www.youtube.com/>

## 1 Introduction

efficiency, or design considerations such as the need for a distinct search functionality. From a technical point of view, the use of different data access paradigms complicates the retrieval from such collections because the utilized retrieval model has to support these paradigms.

As a consequence, the main challenges in MMIR – the retrieval engine and the user interaction – have to be addressed in a holistic way because they can hardly be separated if the search for multimedia information is recognized as a user-centered process.

### 1.1 Motivation

A holistic theoretic perspective on MMIR/IR research is taken by the *cognitive viewpoint* which “seeks to cover all aspects of IR” [Ingwersen & Järvelin 2005, p. 112]. Originating in the research field of information science, the cognitive viewpoint has not yet attracted the same amount of attention by computer scientists than the system-centric viewpoint on IR primarily focussing on the development of retrieval models and optimizing their parameters. However, by addressing both the retrieval model and the user interaction as a whole, it becomes possible to design usable MMIR systems that do not incorporate logical breaks and inconsistencies between the user interaction and the underlying relevance assessments carried out by the retrieval engine. At the same time, one can expect such systems to better address user needs because they are not limited to pre-defined aspects of the user needs (e.g., the support of keyword-based queries).

One representative of the cognitive viewpoint is the *principle of polyrepresentation* (PoP) [Ingwersen 1996; Ingwersen & Järvelin 2005], which forms one half of the theoretic background of this dissertation aiming at the development of a preference-based approach to interactive MMIR. Roughly speaking, the PoP theorizes that representations describing a document are based on various cognitive processes dealing with it, e.g., a title, its color or shape features, its creator, or its date of creation. This multitude of representations can be fused to form a conjunctive *cognitive overlap* (CO) in which highly relevant documents are likely to be contained.

However, the PoP does not answer how a retrieval model has to be implemented in a technical sense which is of interest in the field of computer science. Thus, it is necessary to find a retrieval model that is capable of implementing the PoP in order to be used in a retrieval engine.

One way to implement the PoP, that is also investigated in this thesis, are quantum mechanics-inspired IR models. In his seminal work, van Rijsbergen [2004] used findings from quantum mechanics to build a consistent theory for IR that motivated many publications, e.g., by Melucci [2008]; Schmitt [2008], or Piwowarski et al. [2010]. As a result, the combination of the cognitively motivated PoP with results from quantum theory has formed a new field of research that gained momentum recently [Frommholz & van Rijsbergen 2009; Zellhöfer & Schmitt 2010c; Zellhöfer et al. 2011].

In this dissertation, the *commuting quantum query language* (CQQL) [Schmitt 2008] is used to implement the PoP. As a consequence, the retrieval model behind CQQL makes up the other half of this thesis’ theoretic background. CQQL is particularly interesting

## 1.2 Research Questions and Contributions

because it can also combine data access paradigms used in the fields of DB and IR which is crucial for exploiting synergies between various media [Lew et al. 2006, cf. p. 14], or, in other words, for exploiting the polyrepresentative nature of multimedia documents. Furthermore, CQQL allows the personalization of retrieval results using a preference-based relevance feedback approach called *PrefCQQL* in order to respect the dynamic nature of the search process and information need.

To sum up, this dissertation combines two approaches to MMIR/IR that try to offer a holistic and consistent theory for IR, i.e., the PoP and CQQL. Recently, theory returned again into the focus of IR, which is also put forward by Fuhr [2012], who motivates the need for a solid theory as follows:

1. "Theories give us a deeper *insight* into the foundations of our field, thus satisfying the scientific interest.
2. Theoretic models possess general *validity*, thus forming the basis for broad ranges of applications – in contrast to experiments where we just don't know to what extent their results can be generalized.
3. Only theories allow us to make reliable *predictions* – which is important from the engineer's point of view."

[Fuhr 2012, p. 22]

Besides satisfying the scientific interest of the author, this dissertation investigates the validity of the hypotheses of the PoP in the field of interactive MMIR. As such, it contributes to the formation of the PoP as a verifiable or, respectively, falsifiable theory. This research is motivated by the objective to develop a preference-based approach to interactive polyrepresentative multimedia information retrieval.

Eventually, this thesis examines whether the hypotheses of the PoP can predict how a MMIR system will perform in different retrieval scenarios.

## 1.2 Research Questions and Contributions

The core of this thesis is formed by the utilization of a novel query language, the *commuting quantum query language* (CQQL) [Schmitt 2008], in the field of *multimedia information retrieval* (MMIR). In order to address the inherent dynamics of the search process, a preference-based relevance feedback approach called *PrefCQQL* relying on machine-based learning is discussed. These two components build the foundation for an investigation of the utility of the *principle of polyrepresentation* (PoP) [Ingwersen 1996] implemented by CQQL in interactive MMIR. In order to evidence its utility, various experiments are conducted that answer the following research questions.

### 1.2.1 Can the Hypotheses of the Principle of Polyrepresentation Be Verified in Multimedia Information Retrieval?

This research question is basal as it is necessary to answer whether the hypotheses of the PoP can be interpreted as a theory for MMIR. This question is in particular interesting

## 1 Introduction

in conjunction with studies supporting the validity of the PoP in textual IR [Skov et al. 2004; Larsen et al. 2006, 2009].

Nevertheless, a comprehensive investigation of the utility of the PoP in the domain of MMIR is still missing. There are plenty of studies dealing with feature fusion techniques in MMIR, which combine representations (or features) in a similar fashion to the PoP in order to improve retrieval effectiveness. Although these techniques have been shown to be effective, they often lack a theoretical foundation [Kokar et al. 2004].

Hence, an investigation of the PoP and its implementation in form of a probabilistic logic (i.e., CQQL) based on findings from quantum mechanics is appealing, because it relies on a consistent theoretic framework. Unfortunately, the utilization of a solid theory is a factor being often disregarded in favor of optimizing parameters for (machine-based learning-supported) information fusion MMIR systems [Fuhr 2012].

Because of the complexity of this research question, it is fragmented into three parts described in the following subsections.

### 1.2.1.1 Can CQQL Be Used to Implement a Formal IR Model on the Basis of the Principle of Polyrepresentation?

As said before, quantum mechanics-inspired IR models attracted attention by researchers trying to implement the PoP. CQQL, as a representative of the quantum mechanics-inspired IR models, has some features suggesting it for implementing the PoP, i.e., it supports the incorporation of multiple representations into the search process, it is based on logics as the PoP, it can handle representations based on various data access paradigms, and it supports means for personalization with the help of weighting variables (which will become important later).

In addition, it is reasonable to explore how CQQL, as a quantum mechanics-inspired query language, is related to common formal IR models.

The answer to this question acts as a precondition for the following questions because this thesis can only make statements about the utility of the PoP on the basis of its implementation with CQQL.

### 1.2.1.2 Do the Hypotheses of the Principle of Polyrepresentation Apply in Multimedia Information Retrieval?

As implied before, the fusion of different representations according to the PoP introduces an “intentional redundancy” [Ingwersen 1996] amongst the representations, which serve as the basis of the relevance assessment of a document with regard to a query. In CBIR and MMIR systems it is often attempted to avoid such forms of redundancy because redundant information, e.g., in form of similar representations, is commonly assumed to not increase retrieval effectiveness [Eidenberger 2003; Deselaers et al. 2008].

Various benchmarks, e.g., different tasks of the ImageCLEF benchmark [Popescu et al. 2010; Tsikrika et al. 2011; Thomee & Popescu 2012; Zellhöfer 2012e; Caputo et al. 2013], have shown that *feature fusion*, i.e., the combination of different representations

during retrieval, improves retrieval effectiveness. However, there is no clear theoretically sound advice on how representations have to be fused. This answer is given by the PoP recommending the formation of the CO. In addition, the best performing feature fusion approaches often rely on training data, which might not be available in every MMIR use case, for their machine-based learning algorithms.

To reveal whether the hypotheses of the PoP hold in MMIR, three further subquestions are answered in this thesis with the help of an experimental evaluation:

1. Is the formation of cognitive overlaps superior to other approaches towards feature fusion that do not rely on any kind of training data?
2. Does the usage of “intentional redundancy” limit retrieval performance?
3. Can the observations regarding the principle be generalized or are they limited to certain usage domains?

According to Fuhr [2012], this form of experimentation has to be regarded as *why-experimentation*, which tries to validate a given model’s assumptions relying on a solid theory (the PoP implemented with CQQL). This separates the experiments presented in this thesis from typical IR research that often relies *how-experimentation* aiming at the optimization of retrieval parameters [Fuhr 2012, cf. Sec. 4.2].

### **1.2.1.3 Can Polyrepresentation Compensate the Weak Retrieval Effectiveness of Some Low-Level Features in Multimedia Information Retrieval?**

In MMIR, the retrieval effectiveness of different representations varies to a large degree. For instance, while textual representations usually perform very well, representations relying on some sort of automatic image segmentation still perform weakly in general MMIR, i.e., when they cannot be optimized for a particular usage domain. Hence, it would be desirable if the PoP could compensate such effects to increase the principle’s overall utility in MMIR.

### **1.2.2 Do the Hypotheses of the Principle of Polyrepresentation Hold in a Preference-based Interactive Multimedia Search Process?**

Without doubt, the actual user interaction is in the focus of every user-centered system design. However, this research question primarily aims at examining the retrieval effectiveness of the polyrepresentative PrefCQQL approach in an interactive MMIR scenario, i.e., when relevance feedback is provided. In order to investigate if the hypotheses of the PoP hold in such a scenario, the aforementioned three subquestions are answered with the help of an experiment.

### **1.2.3 Can a Usable Multimedia Information Retrieval System Be Built on the Basis of the Principle of Polyrepresentation and PrefCQQL?**

This research question deals with the user experience of a MMIR system built on the basis of the PoP and PrefCQQL. From this perspective, it takes a similar role as question

## 1 Introduction

1.2.1.1 as it examines the actual implementation of a conceptual user interaction model based on the PoP.

To answer the research question, a prototypical MMIR system is built in order to test it in form of a proof of concept in a real world scenario, i.e., the search in a personal photo collection. The user experience and usability of the prototype is evaluated on the basis of a quantitative and qualitative user study to get comprehensive insights into the test persons' user experience.

Unlike before, the system is no longer limited to PrefCQQL alone. As said before, users in MMIR typically rely on different information seeking strategies to satisfy their current information need. Hence, the prototype offers different strategies such as browsing and directed search between which a seamless transition is possible. In order to assess the impact of different information seeking strategies on the user experience, different user interface variants are examined.

To conclude, further properties of CQQL/PrefCQQL are studied in order to reveal whether they can make an impact on the overall user experience, e.g., by supporting or disturbing users during the retrieval process. Exemplary properties include conflicting preferences or statistics that might indicate the need for the suggestion of a new query that better reflects the user's dynamic information need.

### 1.2.4 Positioning of the Dissertation

To recapitulate, the aforementioned research questions try to answer how a MMIR system, consistent from the retrieval engine to the user interface, can be built. Unsurprisingly, the dissertation is therefore clearly positioned in the computer science-related part of the wide research field of IR, or more precise, of MMIR. Overviews of general IR are available by van Rijsbergen [1979]; Dominich [2008]; Croft et al. [2009], or Baeza-Yates & Ribeiro-Neto [2011]. More specific publications about MMIR are available by Schmitt [2006] or Blanken et al. [2007].

This dissertation also has a strong relation to databases [Silberschatz et al. 1999; Garcia-Molina et al. 2000; Date 2000; Saake et al. 2010; Elmasri & Navathe 2011], and relational databases [Codd 1970; Date 1982] in particular. This relation is established via two links. First, via the usage of a relational complete query language, i.e., CQQL, as the core of the designed MMIR system, and second, via preferences, which are used as a means for "relevance feedback" input in the field of databases.

Preferences also link this thesis to psychology and microeconomics [Lancaster 1991] in which the examination of consumer preferences forms a part of utility theory [Fishburn 1968].

Because of its reliance on the PoP, this dissertation is also related to the research areas within information science [Ingwersen 1992] that deal with IR. However, this research is only streaked because of the thesis' focus on the computer scientific problems in MMIR.

Further research areas that have an impact on the research described in this thesis are interaction design and usability engineering [Preece et al. 2002; Shneiderman & Plaisant 2005; Cooper et al. 2007; Nielsen 2009], and machine-based learning [Russell et al. 2007; Liu 2011].

These examples make clear that this dissertation takes an interdisciplinary view on MMIR. Thus, it cannot discuss all field in the required depth. As a consequence, when in doubt, this thesis will take an IR perspective on the described research. This is particularly true for the information seeking process which assumes layperson users that cannot communicate with the retrieval system using a structured query language as it is the case in the field of databases.

For practical reasons further elaborated in Chapter 6, the rather broad concept of MMIR is limited to visual perceivable media such as images and (textual) metadata in the context of this dissertation. However, many conclusions drawn might also apply to other media.

To conclude, a list of publications related to the research presented in this thesis is available in Appendix F.

### 1.3 Structure of the Dissertation

Neglecting the appendices, the thesis is divided into five main parts. The first two parts establish a theoretical basis for the implementation and evaluation presented in part three and four. The fifth part concludes the thesis by summarizing its contents and providing future prospects of topics that need further research. The detailed structure of the dissertation is as follows:

**Part I. Foundations and Background** Chapter 2 defines the central terms and principles of information retrieval and multimedia retrieval. Furthermore, it gives an overview of traditional IR models that are typically associated with the system-centered approach in IR.

Chapter 3 summarizes the motivation and theories behind interactive information retrieval (IIR) and discriminates this approach from traditional, system-centered IR. In particular, the chapter recapitulates the hypotheses of the PoP. Chapter 3 concludes with a discussion on information seeking strategies that are common in MIR<sup>4</sup>.

**Part II. Learning User-specific Weights in Logic-based Queries** In order to implement an IR model based on the principle of polyrepresentation, Chapter 4 describes the theoretic basis of the quantum logic-based query language *CQQL* and its relation to other formalizations of the PoP and more traditional IR models.

Chapter 5 then introduces the preference-based machine learning approach *PrefCQQL* which itself utilizes *CQQL* at its core. Additionally, this chapter addresses the notion of preference with a focus on preference approaches in databases, i.e., so-called qualitative and quantitative approaches.

As such, these two chapters form the theoretical basis for the conceptual design and the implementation described in the third part of this dissertation.

---

<sup>4</sup>That is, multimodal information retrieval, a specific technique for MMIR (see Definition 2.21).

**Part III. Concept and Implementation** The user-centered design process of the prototypical *Pythia MIR system* is described in Chapter 6. The chapter both addresses the user interaction model and the retrieval engine of the Pythia MIR system. Furthermore, it separates the developed system from other interactive MIR systems.

**Part IV. Evaluation** The fourth part of the thesis is divided into two separate evaluations of the Pythia MIR system. After a brief introduction into the evaluation of IR and IIR systems (see Chapter 7), Chapter 8 gives an evaluation of the retrieval effectiveness of the CQQL/PrefCQQL approach using different matching functions, which determine the relevance of a document with respect to a given query, in a query-by-example MIR scenario.

Chapter 9 evaluates the user experience of the Pythia MIR system using methods borrowed from usability engineering.

**Part V. Conclusions** The last Chapter 10 concludes the thesis by summarizing its contributions and outlining future research options.

To facilitate reading, the dissertation also contains a central list of definitions (see page 507) that are crucial for an understanding of its content. Page 511 contains a list of the included theorems, lemmata, and proofs. The examined matching functions (see Chapter 8) are listed on page 513 and the user stories, which form the basis for the user-centered design process of the Pythia MIR system, are listed on page 515.

To further increase the comprehensibility of this dissertation if it is not read in a linear fashion, the text contains many cross references and repeatedly summarizes findings that are required to understand the different sections. All cross references are hyperlinks in the PDF version of this text (see Appendix E).



## **Part I**

# **Foundations and Background**



## 2 Multimedia Information Retrieval

In principle, information storage and retrieval is simple. Suppose there is a store of documents and a person (user of the store) formulates a question (request or query) to which the answer is a set of documents satisfying the information need expressed by his question. He can obtain the set by reading all the documents in the store, retaining the relevant documents and discarding all the others. In a sense, this constitutes 'perfect' retrieval. This solution is obviously impracticable. A user either does not have the time or does not wish to spend the time reading the entire document collection, apart from the fact that it may be physically impossible for him to do so.

---

*van Rijsbergen [1979]*

Before the field of multimedia information retrieval (MMIR)<sup>5</sup> and its distinction from information retrieval (IR) is discussed, the core terms of media and multimedia deserve closer attention.

### 2.1 Media and Multimedia

The term "medium" is derived from the Latin word for middle, center, or in-between. Typically, it either describes a communication channel for transmitting messages between a sender and a receiver or a means of information storage.

The first interpretation, based on the seminal work by Shannon & Weaver [1949], is the pre-dominant interpretation in the humanities, especially in communication and media theory, and has been extended to investigate the process of communication primarily between humans. As such, it also examines how media affect the communication, e.g., by McLuhan [2010] and others.

In contrast, a more techno-centric definition focusses on the medium as a storage of information (not necessarily physical) and is often used in computer science and engineering disciplines, e.g., by Bruns & Meyer-Wegener [2005] or Blanken et al. [2007]. This dissertation follows this interpretation.

**Definition 2.1 Medium:** *In accordance with Blanken et al., we define "a medium to be a type of information representation" [Blanken et al. 2007, p. 3].* ◇

**Definition 2.2 Multimedia:** *The combination of multiple media, e.g., audio and images, of which at least one medium is non-alphanumeric (i.e., text), is called multimedia [Blanken et al. 2007]. These multimedia are commonly multimodal, i.e., they utilize multiple modalities.* ◇

---

<sup>5</sup>"Multimedia information retrieval" is also often abbreviated as "multimedia retrieval" in this dissertation.

## 2 Multimedia Information Retrieval

**Definition 2.3 Modality:** *In order to process information, humans can rely on different receptors or sensors that process different media types, e.g., auditive or visual data. These communication paths are called modalities, such as the vision modality or the haptic modality.*  $\diamond$

Some authors limit the scope of the definition of multimedia to include only digital [Schmitt 2006] and multimodal media, e.g., Bruns & Meyer-Wegener [2005], or distinguish between static and non-static (or time continuous media such as interactive animations) [Schmitt 2006; Blanken et al. 2007]. This discrimination is not crucial for this thesis. Nevertheless, because of the nature of this dissertation all considered media objects are assumed in digital form as they are processed by a computer.

### 2.2 Information Retrieval

From a historic perspective, information retrieval deals with the retrieval of information<sup>6</sup> that is implicitly present within a set of textual documents [Croft et al. 2009] and is relevant to the search task submitted by a user. As such, the expected results can be inaccurate because the relevance of the documents cannot be determined without fault as outlined in this section.

This discriminates the field from relational databases [Date 1982] or other data retrieval techniques which operate on data with a well-defined structure and semantics [Baeza-Yates & Ribeiro-Neto 2011]. In consequence, databases (DB) can guarantee the retrieval of accurate results with respect to a well-defined query – otherwise they would be considered defective.

In reference to Dominich [2008], we formulate information retrieval (*IR*) as the following mapping:

$$IR : (U, IN, Q, D) \rightarrow \mathcal{R} \quad (2.1)$$

where  $U$  stands for a *user* who has a specific *information need* ( $IN$ ) that should be satisfied by the *information retrieval system* ( $IRS$ ). In order to interact with the  $IRS$ , the user states a *query*  $Q$  that reflects the  $IN$  completely or in parts (although the latter is the usual case). As a response to  $Q$ , which is later evaluated against a *collection of documents*  $D$  forming the  $IRS$ ' database, the user is confronted with a list of *retrieved documents*  $\mathcal{R}$  that reflects the system's assessment of relevant documents with respect to the query provided by the user, i.e., the "best matching" documents to the query.

#### 2.2.1 Principal Concepts in Information Retrieval

The point of view on  $IR$  taken in Equation 2.1 is often called the *system-centric* viewpoint (or computer-centered view, e.g., by Baeza-Yates & Ribeiro-Neto [2011]) because it focuses mainly on the development of  $IR$  algorithms. Often, this viewpoint is also re-

---

<sup>6</sup>If not stated otherwise, we refer to the term "information" as a synonym for (relevant) knowledge or facts in contrast to the mathematical definition used in information theory founded by Shannon [1948].

ferred to as the standard IR model or the basic laboratory model [Ingwersen & Järvelin 2005]. There are two major problems with the system-centric approach.

First, the user is assumed to only interact with the IRS via some kind of query that does not necessarily reflect his IN but that is expected to retrieve only (or at least as many as possible) relevant documents. In other words, the user provides a vague idea of the IN but expects a precise retrieval result. Other forms of interaction or factors affecting the retrieval are not considered but are addressed by user-centered approaches discussed in Section 3.

Second, the imprecise (and possibly misleading) query is matched against the document collection to determine the most relevant documents according to the current IN. In order to retrieve the relevant documents, the query is interpreted by the means of the IR model and compared to document representations. How well these relevance assessments align with the user's notion of relevance is highly dependent on the used IR model. Thus, the system-centric viewpoint is tightly coupled to the evaluation of the algorithmic implementation of IR models.

Before presenting common IR models in Section 2.2.2, central terms of IR need to be defined for the understanding of this thesis. Figure 2.1 illustrates their relation while omitting the evaluation, which is covered separately in Section 7.1.

**Definition 2.4 Information need:** *According to Belkin et al. [1982, p. 2], the “information need arises from a recognized anomaly in the user’s state of knowledge concerning some topic or situation [...]” – a famous hypothesis known as the ASK hypothesis<sup>7</sup>.*

*An IN is inseparably connected to a user and his context (e.g., prior knowledge or expertise in the field of research, work task, understood languages etc.).*

*Thus, it augments a query with additional information that could be exploited by an IRS. This additional information is obvious to the user and therefore not communicated to the IRS making the retrieval of relevant documents more complicated.*

*Eventually, the objective to satisfy this IN motivates the user to interact with an IRS.* ◇

**Definition 2.5 Query:** *A query is often considered to be the primary means of communication with an IRS, which expresses parts of the user’s IN. For example, in order to express the IN in form of a query, keywords or sample documents such as images or texts can be used.* ◇

**Definition 2.6 Query formulation problem:** *Due to the inability of the user to specify the IN precisely [Belkin et al. 1982; van Rijsbergen 1986b], the user is challenged to input a query that 1) reflects a large part of the IN, and 2) is interpreted by the IRS in the way intended.* ◇

**Definition 2.7 Query representation:** *Depending on the IR model, the query is transformed into an internal representation  $q^r$  (where the superscript  $r$  denotes representations throughout this thesis) that is later used to match against the document representations to calculate the relevance of a document.* ◇

**Definition 2.8 Semantic gap:** *The semantic gap describes the difference at a semantic level between two expressions describing the same object, e.g., an IN in natural language and its representation within the IRS. This gap manifests in an ambiguous, inappropriate, or poor ex-*

<sup>7</sup>Anomalous state of knowledge

## 2 Multimedia Information Retrieval

pression of the higher level semantics (e.g., the image of a flower, i.e., its meaning) in its low-level counterpart (e.g., a color histogram, i.e., statistical information) which usually results in a loss of semantic (and contextual) information. See Section 2.3.3 for an illustration of the semantic gap in the scope of this dissertation.  $\diamond$

**Definition 2.9 Document:** For the scope of this thesis, a document  $\mathbf{d}_i$  out of a document collection  $\mathbf{D}$  ( $\mathbf{d}_i \in \mathbf{D}$ ) is a container of textual or multimedia information in its original form that can be cognitively processed by a user, i.e., it can be read or interpreted directly. The traditional media IR is concerned with are text documents.  $\diamond$

**Definition 2.10 Document representation:** A document representation  $\mathbf{d}_i^r \in \mathbf{D}^r$  represents the document  $\mathbf{d}_i$  within the IRS. This technical representation depends on the utilized IR model and is used for the calculation of the relevance of  $\mathbf{d}_i$  with respect to a query  $\mathbf{q} \in \mathbf{Q}$ , whereas  $\mathbf{Q}$  denotes all possible queries.  $\diamond$

**Definition 2.11 Document storage:** Although out of the scope of this work, the document storage holds the document representations and appropriate indices<sup>8</sup> to allow fast access to the documents. The document storage is directly linked to the matching as it can only operate on the data, e.g., full texts, metadata etc., available in the indices and the index vocabulary.  $\diamond$

**Definition 2.12 Index vocabulary:** In traditional IR, the index vocabulary  $\mathbf{K}$  consists of index terms or keywords that can be used during the retrieval [Croft et al. 2009]. Only a keyword  $\mathbf{k}_j$  out of  $\mathbf{K}$  ( $\mathbf{k}_j \in \mathbf{K}$ ) can be used for the representation of a document. In order to indicate whether a specific index term  $\mathbf{k}_j$  is present in a document representation  $\mathbf{d}_i^r$ , an index term weight  $\mathbf{w}_{i,j} \geq 0$  is used. If  $\mathbf{w}_{i,j} = 0$  then  $\mathbf{d}_i$  does not contain  $\mathbf{k}_j$ . Hence, a document representation has the following (vector) form:  $\mathbf{d}_i^r = (\mathbf{w}_{i,1}, \mathbf{w}_{i,2}, \dots, \mathbf{w}_{i,j})$ .  $\diamond$

**Definition 2.13 Relevance:** Traditionally, relevance is assumed binary, e.g., by Robertson [1977]. That is, a document can be relevant or irrelevant with respect to a query. More recent studies indicate that a gradual relevance scale is more appropriate as it better matches the users' notions of relevance [Spink et al. 1998].  $\diamond$

**Definition 2.14 Matching:** The matching between the query representation and the document representations is carried out by an algorithm that mostly outputs a ranking sorted decreasingly by the relevance of the documents. Consequently, the matching process determines a list (or a set) of potentially relevant documents that are presented to the user.  $\diamond$

**Definition 2.15 Matching function:** In order to carry out the matching, the matching algorithm relies on a matching function which calculates the similarity between a query representation and a document representation. Typically, a matching function calculates a numerical score that is interpreted as the relevance of a document during the matching process where this score serves as the sorting criterion for the result list (see above). The matching function is a central point in the system-centric viewpoint.  $\diamond$

**Definition 2.16 Relevance feedback:** The system-centric approach features an optional

---

<sup>8</sup>Although the term "indexes" is pervasive, especially in the field of databases to describe data structures for fast access, this thesis uses the form "indices" as recommend by the Oxford English dictionary for technical usage.

personalization technique called *relevance feedback (RF)*, which includes user feedback to adjust the matching algorithm in order to retrieve documents that satisfy the IN to a higher degree. RF is usually iterated in order to obtain results that better satisfy the IN [van Rijsbergen 1986b]. Section 5.1.1 discusses RF in more detail. Further personalization techniques relevant to this dissertation are addressed in Section 5.1.  $\diamond$

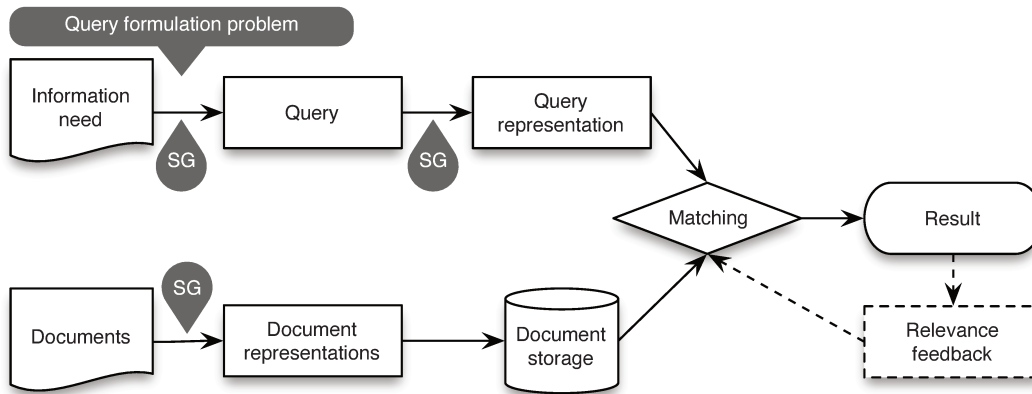


Figure 2.1: Schematic illustration of the system-centric/standard IR model without evaluation; SG denotes impacts by the semantic gap

## 2.2.2 Information Retrieval Models

The decision whether a document is relevant or irrelevant to an IN is the central problem in IR. Information retrieval models furnish an answer for the relevance decision mechanism, although this process is not necessarily transparent to the user. After decades of research, plenty of IR models have been proposed, implemented, and evaluated. This dissertation cannot provide a full discussion on this issue. Instead, we focus on models that are directly related to the IR model discussed in Chapter 4. We subdivide these models in accordance with Dominich [2008] and Baeza-Yates & Ribeiro-Neto [2011] in set theoretic, algebraic, and probabilistic approaches. Considering their importance in this dissertation, quantum theoretic IR models are discussed separately in Section 4.5.

For the sake of brevity, the following techniques dealing with text processing are not discussed: stemming, i.e., to reduce a word variant to its grammatical stem form; stop word elimination, i.e., to remove words with low semantic information such as articles or high frequent words from a text; or indexing techniques as they are widely covered in various text books. Instead, this section emphasizes the mathematical properties of the models.

Further information about common IR models and their implementation can be found in various publications, e.g., by van Rijsbergen [1979]; Dominich [2008]; Croft et al.

[2009], or Baeza-Yates & Ribeiro-Neto [2011]. Finally, Schmitt [2006] relates these models to multimedia retrieval and databases.

### Retrieval Models based on Set Theory

**Boolean Retrieval** In general, the Boolean retrieval model (BRM) is considered very simple. It is based on set theory and Boolean algebra (see Appendix A.1.5). It restricts the index term weight to be false or true, i.e.,  $w_{i,j} \in \{0,1\}$ , denoting the absence or presence of a term in a document. A query  $q$  is composed by linking index terms with the help of logical connectors (conjunction, disjunction, or negation<sup>9</sup>).

As a result, the matching function is exact and can determine whether a document is relevant or not based on the binary relevance decision by comparing the document representations against the logical expression specified by  $q$ . Due to this clear discrimination between relevant and irrelevant documents, the retrieved documents form a set without an order that can also be empty if no documents match  $q$ .

The risk to retrieve too few or too many documents, e.g., by a false usage of the Boolean connectors by users [Soergel 1985], is often criticized in addition to the missing partial or best matching (see below) in the BRM.

The relational model (RM) [Codd 1970; Date 1982] used in relational database systems (RDBMS) shares its Boolean origin with the BRM. The RM is based on the mathematical foundations of set theory and first-order predicate logic. Via this theoretical connection, a relation between IR and DB can be established. In fact, some authors, e.g., van Rijsbergen [1986b] and Nottelmann & Fuhr [2003], regard IR as a generalization of DB. The RM assumes that all data can be expressed as *tuples* with various *attributes* that form a *relation*. As mentioned at the beginning of Section 2.2, the RM requires a well-defined structure of the “documents”, i.e., the tuple, in a tabular form (the relation). As a result, one can assume  $d_i = d_i^r$ . These tuples are then matched against a query that is comparable to the approach taken in the BRM. Typically, a relational database consists of multiple relations that can be used for retrieving data using a first-order predicate logic. In contrast to the BRM, the RM’s logic is commonly 3-valued. That is, it uses the truth value “unknown” in addition to true and false to indicate a missing or unknown information. For a more thorough discussion of the model, see Codd [1970]; Date [1982]; Silberschatz et al. [1999]; Garcia-Molina et al. [2000], or Date [2000].

**Extended Boolean Retrieval** To overcome the strictness of the Boolean model, various extensions have been suggested. These models have in common that they try to provide a ranking (or partial matching) of the retrieved documents, means to express the subjective importance of query terms using weights, and “softened” versions of the Boolean operators. A number of authors have discussed these extended BRMs, e.g., Waller & Kraft [1979]; Salton et al. [1983]; Yager [1988]; Fox et al. [1992], or Lee [1994].

---

<sup>9</sup>An informal introduction of Boolean logic is available in Appendix A.1.4.



Roughly speaking, the core idea of “softened” Boolean operators is to equip these connectors with a parameter that affects their logical characteristic. That is, a parameter steers the “conjunctiveness” or “disjunctiveness” of an operator connecting two terms. The  $p$ -norm [Salton et al. 1983] is a representative of such approaches. Depending on its  $p$  parameter setting, the  $p$ -norm’s matching characteristic shifts from vector space-like evaluation (see below) to fuzzy logical, a logical model that is described in the next section. This effect is due to the arithmetic evaluation of the logical  $p$ -norm connectors, which yield the inner product if  $p$  is set to 1 and min/max if set to  $\infty$ . For a more thorough discussion, see Salton et al. [1983].

Often, extended Boolean approaches also allow a weighting of the query terms in order to introduce a ranking and to overcome the aforementioned drawbacks of the BRM. Comparable weighting approaches can also be found in the field of DB, e.g., by Fagin & Wimmers [2000].

**Fuzzy Logic Retrieval** Fuzzy logic retrieval models (FRM) are based on fuzzy set theory [Zadeh 1965] and logic [Zadeh 1988]. Fuzzy set theory can be regarded as a generalization of set theory and the BRM [Kraft & Buell 1983].

**Definition 2.17 Fuzzy set:** *The core idea is to associate a membership function  $\mu \rightarrow [0, 1]$  with the elements  $x$  of the space of objects  $X$  in order to determine their membership to a certain fuzzy set. That is, every fuzzy set  $A$  is defined as follows:  $A = \{x, \mu_A(x)\}$  where 1 denotes full membership to  $A$ , while 0 indicates no membership. The values in between represent the “grade of membership” [Zadeh 1965, p. 339].*  $\diamond$

In other words, fuzzy set theory allows a gradual membership of elements to one or more sets in contrast to the strictly binary membership in traditional set theory. To cope with the membership values, Zadeh [1965] defines the set operations conjunction, disjunction, and complement (negation) between the fuzzy sets  $A$  and  $B$  as follows.

$$\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x)) \quad (2.2)$$

$$\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x)) \quad (2.3)$$

$$\mu_{\neg A}(x) = 1 - \mu_A(x) \quad (2.4)$$

The logical connectors are defined accordingly by Zadeh [1988]. Because of different advantages and disadvantages of these functions, which are covered in Section 4.5, numerous alternatives have been proposed [Zimmermann 1996].

Based on this theory,  $w_{i,j}$  can be in the interval  $[0, 1]$  to express gradual relevance of a document regarding an index term. Similar to the BRM, a query can be composed using logical connectors that are evaluated as described before. For instance, the actual calculation of  $w_{i,j}$  can be based on a thesaurus that contains related terms to the index terms in order to determine the semantic neighborhood of a term found in a document and an index term  $k_j$  [Baeza-Yates & Ribeiro-Neto 2008].

### An Algebraic Retrieval Model – The Vector Space Model

The vector space model (VSM) is the most prominent algebraic retrieval model. It is called algebraic because both queries and document representations are modeled in an  $t$ -dimensional vector space. The dimensionality  $t$  is determined by the number of index terms, i.e.,  $t = |K|$ . The determination of the similarity between a query vector and a document representation vector is solved by using methods from linear algebra, e.g., the cosine of the angle between both vectors.

In the VSM,  $w_{i,j} \geq 0$  holds, but it is not restricted to the interval  $[0, 1]$ . The same applies to the terms  $w_{q,j}$  of the query representation  $q^r$ . The resulting vectors are as follows:  $d_i^r = (w_{i,1}, w_{i,2}, \dots, w_{i,t})$  and  $q^r = (w_{q,1}, w_{q,2}, \dots, w_{q,t})$ , whereas  $t$  is the total number of index terms. As said before, the cosine of the angle between  $d_i^r$  and  $q^r$  (or variants) is often used to calculate the similarity between  $q^r$  and  $d_i^r$ :

$$\text{sim}(d_i^r, q^r) = \frac{\sum_{j=1}^t w_{i,j} \cdot w_{q,j}}{\sqrt{\sum_{j=1}^t w_{i,j}^2 \cdot \sum_{j=1}^t w_{q,j}^2}} \quad (2.5)$$

Obviously,  $\text{sim}(d_i^r, q^r) \rightarrow [0, 1]$  holds, where 1 denotes a full match or the highest degree of similarity. This value is used to sort the retrieved documents in decreasing order to present a ranking to the user that also includes partial matches, e.g., documents that do not contain all index terms of the query. How the index term weights can be calculated, e.g., by counting the occurrence of terms in a document or by using the well-known  $tf * idf$  formula<sup>10</sup>, falls out of the scope of this thesis and has been discussed by the authors referenced at the beginning of this section.

Because of its neat mathematical foundation, the VSM can be used in every scenario that can be represented with vectors [Schmitt 2006, cf. p. 33f.]. In fact, “most of the new systems adopt at their core some form of vector retrieval” [Baeza-Yates & Ribeiro-Neto 2008, p. 38]. The utilization of the VSM in multimedia retrieval is discussed in Section 2.3.2. Additionally, it is a fertile ground to implement relevance feedback as shown by the original work by Rocchio [1971].

On a more theoretical level, the VSM is criticized for its low degree of guidance regarding index term weighting, the choice of matching algorithms, and their relation to relevance [Croft et al. 2009].

### Probabilistic Retrieval Models

The core idea of probabilistic retrieval models (PRM) is to use the theoretic framework of probability theory from mathematics (see Appendix A.2) to solve the IR problem. Most models are based on the probability ranking principle (PRP) [Robertson 1977] that provided an early theoretic justification for rankings based on the probability of relevance (POR) of a document.

<sup>10</sup>The  $tf*idf$  formula assigns each  $w_{i,j}$  with a product of the term frequency ( $tf$ ) of  $k_j$  in  $d_i^r$  and the inverse of frequency ( $idf$ ) of  $k_j$  in the collection [Salton & Buckley 1988]. The  $idf$  expresses how rare an index term is within the collection.

**Definition 2.18 Probability of relevance (POR):**

“The probability ranking principle (PRP): *If a reference retrieval system’s response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data.*” [Robertson 1977, p. 281]

◇

With this definition, Robertson outlines the two main ideas of the probabilistic approaches. First, retrieved documents are ordered by their decreasing POR<sup>11</sup>. Second, the estimation of the probability of relevance is made on basis of a request (or a query) and the document representations (“data that has been made available to the system”) alone. Additionally, Robertson assumes the relevance of a document regarding a query to be binary (a “dichotomous criterion variable” [Robertson 1977, p. 280]) and independent from other documents in the collection. Although the PRP leaves open how to compute the POR, it has motivated the development of probabilistic retrieval models, becoming the pre-dominant retrieval model of today<sup>12</sup> [Croft et al. 2009]. In order to estimate the POR, formally denoted as  $P(R|q, d)$ , probabilistic approaches often rely on Bayes’ theorem (see Appendix A.2.4) and assume the index terms to be independent from each other. Being one of the most widely used models, we cannot discuss all its variants in depth. Thus, we present one representative and then focus on a sub-area of the probabilistic models: probabilistic logics. A more thorough discussion is available by a number of authors, e.g., by Fuhr [1992]; Pearl [2008]; Aly & Demeester [2011], or by Blanken et al. [2007] who put an explicit focus on multimedia retrieval as an application domain.

One representative of the PRM approaches is the *binary independence retrieval* (BIR) model [Robertson & Spärck Jones 1976]. In the BIR model,  $w_{i,j} \in \{0, 1\}$  and  $w_{q,j} \in \{0, 1\}$  hold, i.e., each document representation’s index term weight indicates whether an index term is present in the document or not. As before, a query is a subset of  $K$ . The POR is calculated as follows, where  $P(R|d_i^r)$  stands for the probability of picking  $d_i$  from the relevant documents  $R$  regarding  $q^r$  based on its representation  $d_i^r$ . Analogously,  $P(\neg R|d_i^r)$  is its complement.

$$\text{sim}(d_i^r, q^r) = \frac{P(R|d_i^r)}{P(\neg R|d_i^r)} \quad (2.6)$$

By applying Bayes’ theorem, the equation can be transformed:

$$\text{sim}(d_i^r, q^r) = \frac{P(d_i^r|R) \cdot P(R)}{P(d_i^r|\neg R) \cdot P(\neg R)} \quad (2.7)$$

<sup>11</sup>In his work, Robertson mostly equates usefulness with relevance (and user satisfaction).

<sup>12</sup>Please note that this does not contradict with the statements about the VSM made above because PRM systems often operate on document representations in vector form.

As  $P(R)$  and  $P(\neg R)$  are the same for all  $d_i^r \in D^r$ , they can be omitted. By assuming that all index terms are independent, the formula can be simplified even further, e.g., as described by Baeza-Yates & Ribeiro-Neto [2008, pp. 33ff.], in order to transform it to a form that is only relying on the probability of the presence of certain index terms  $k$  in the set of relevant documents  $R$ , i.e.,  $P(k_j|R)$  (and its complement). Initially, the probabilities  $P(k_j|R)$  and  $P(k_j|\neg R)$  are unknown and have to be estimated using different methods [Baeza-Yates & Ribeiro-Neto 2008]. Additionally, relevance feedback can be used to re-weight the index terms in query.

**Probabilistic logics** Early on, Nilsson [1986, 1994] coined the term “probabilistic logic” in the field of computer science. In contrast to classic Boolean logic that operates on truth values (true and false), probabilistic logics (PL) are dealing with probabilities. These probabilities can be linked with logical connectors to provide a framework for formalized reasoning on basis of probabilities in order to estimate the POR of a document. Roughly speaking, the truth values are replaced with uncertainties or probabilities that can also be interpreted as the confidence that a particular proposition is true. Although originally dealing with membership values of different sets, fuzzy set theory and logic might be regarded as a generalization of probabilistic logics. This viewpoint is also taken by Zadeh in his late work [Zadeh 2005] that regards fuzzy logic as a generalization of uncertainty<sup>13</sup>.

Because of their strength in formal reasoning and their flexibility, various probabilistic logics have been proposed in the field of IR. The examples range from the retrieval of structured documents<sup>14</sup> [Lalmas 1996, 1997] to models of subjective belief in order to model theoretic cognitive models such as the principle of polyrepresentation [Lioma et al. 2010, 2012] (see Section 3.2) and approaches integrating DB and IR in one query language [Fuhr 2000].

**Uncertain inference** Probabilistic logics are also closely related to the interpretation of IR as *uncertain inference* originally proposed by van Rijsbergen [1986b]:

**Definition 2.19 Uncertain inference:**

$$P(d_i^r \rightarrow q^r) \tag{2.8}$$

*The uncertain inference tries to express the probability that a document representation  $\mathbf{d}_i^r$  implies the query representation  $\mathbf{q}^r$  in contrast to the POR discussed before.*  $\diamond$

This implication must not be mistaken with the frequently used material implication in classical logic, i.e.,  $A \supset B = \neg A \vee B$ <sup>15</sup>. Instead, van Rijsbergen refers to an alternative

<sup>13</sup>Please note that there is much debate about the relation of fuzzy logic to probability theory – especially in the mathematics community. This debate is mostly related to the interpretation of the values out of  $[0, 1]$  both approaches make use of, i.e., (partial) memberships to classes in the case of fuzzy logic or probabilities of events/uncertainty in probability theory. In computer science, this semantic difference is often neglected and both approaches are used interchangeably, e.g., by Hajek et al. [1995].

<sup>14</sup>That is, IR that exploits the structure of a text such as chapters, abstracts, or other means of structuring.

<sup>15</sup>Here, we use van Rijsbergen’s original notation [van Rijsbergen 1986b] of the material implication.

form that interprets the implication as a conditional probability – “the probability of consequent” – leaving the grounds of classical Boolean logic in direction of conditional logic:

$$P(A \rightarrow B) = P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (2.9)$$

With this conditional interpretation of the implication (“if  $A$  is true, then  $B$ ”), the intuitive soundness criterion holds. That is, it is impossible for the premise of the inference to be true while its conclusion becomes improbable. For a more detailed discussion on the two implications, see Appendix A.2.3.

In his work, van Rijsbergen [1986b] argues further for an understanding of document representations and queries as logical sentences or assertions. In consequence, a logical query is inferred from logical sentences present in a document. In other words, a document is considered relevant if it implies a given query. Thus,  $q^r$  can take the following form, where  $\odot$  denotes a logical connector and  $c_i$  a proposition:

$$q^r = c_1 \odot c_2 \odot \dots \odot c_n \quad (2.10)$$

This inference is similar to the one made in databases in the sense that a relevant document (or a tuple) has to satisfy the logical query. The relation of the interpretation of IR as uncertain inference to database retrieval has also been pointed out by van Rijsbergen [1986b] and Nottelmann & Fuhr [2003], who consider uncertain inference as a “probabilistic generalisation of the logical view on databases” [p. 1]. In contrast to IR, the inference in databases is usually not uncertain. That is, only documents (or tuples) with  $P(d_i^r \rightarrow q^r) = 1$  contribute to the result set due to the Boolean basis of databases discussed before.

By using logical formulae, various knowledge or facts derived from metadata, thesauri, databases, or similar sources, can be incorporated into the retrieval model. Thus, this non-classical logical model of IR is not limited to the usage of index terms, making it a powerful yet comprehensible approach towards IR.

As said before, the relation of uncertain inference to logic-based retrieval is one of the strengths of this approach. Furthermore, probabilistic retrieval models following the notion of uncertain inference have been proposed. One example is *Probabilistic Datalog* [Fuhr 2000], which is discussed separately in Section 4.5.3. Another implementation is CQQL, which is presented in depth in Chapter 4.

To conclude, the formal relationship between uncertain inference and probabilistic logics has been shown by Nottelmann & Fuhr [2003] by mapping the POR ( $P(R|q, d)$ ) to a result ranking ordered by decreasing probability of the conditional implication  $P(d_i^r \rightarrow q^r)$ .

### 2.3 Principles of Multimedia Information Retrieval

In Section 2.1, different media types were introduced. In contrast to IR that often denotes the retrieval from textual documents<sup>16</sup>, the perspective of multimedia information retrieval (MMIR) is much broader. Common media types include images, animations, videos, music, sounds, 3D graphics, or texts. As mentioned before, these different media types can also be combined arbitrarily in a static or non-static manner [Blanken et al. 2007].

Because of the constraints on time and resources, this dissertation only deals with static visual media. That is, the primary focus of this thesis is on image retrieval using techniques from content-based image retrieval (CBIR) and multimodal retrieval (MIR).

**Definition 2.20 Content-based image retrieval:** *The domain of CBIR is concerned with the retrieval of images based on features such as colors, edges, or shape information that can be extracted automatically from these documents (see Section 2.3.1). As such “content-based methods are necessary when text annotations are nonexistent or incomplete. Furthermore, content-based methods can potentially improve retrieval accuracy even when text annotations are present by giving additional insight into the media collections” [Lew et al. 2006, p. 1].* ◇

After the development of the first CBIR systems such as QBIC [Flickner et al. 1995], which is based on color and texture properties of images, it quickly became obvious that the retrieval quality of content-based methods is limited. Thus, numerous systems combining textual information and CBIR techniques were proposed [Lew et al. 2006; Datta et al. 2008]. The combination of textual annotations and content-based features is still a major area of research as Datta et al. [2008, p. 5] state: “while the former [textual data] is considered more reliable from a user viewpoint, there is immense potential in combining the two to build robust image search engines[...] This endeavor will hopefully be actualized in the years to come.” A more comprehensive discussion of the historic development of the CBIR domain can be found in Del Bimbo [1999]; Lew et al. [2006]; Datta et al. [2008]; Müller et al. [2010], or Ruthven & Kelly [2011].

These combination techniques are often referred to as *multimodal retrieval* techniques. Unfortunately, the term “multimodal retrieval” is somewhat misleading. It describes the retrieval of multimedia documents using different document representations (see Definition 2.10). Although operating on the same modality (vision), the combination of textual and image information is widely considered multimodal, e.g., by Yang et al. [2001]; Lew et al. [2006]; Datta et al. [2008]; Myoupo et al. [2010]; Slaney [2010], and Arampatzis et al. [2011]. Hence, representations can share the same modality, e.g., the visual modality in terms of a combination of texts and images (see Definition 2.2). Frequently, these representations differ by the communication type they rely on, i.e., verbal (spoken or written language) or non-verbal (e.g., a picture, tones, or spatial information) communication.

---

<sup>16</sup>Albeit it is frequently used in this context, the term “information retrieval” is not limited to textual documents alone. For instance, it may refer to the retrieval from XML document collections or multimedia data.

This imprecise usage of multimodal retrieval became necessary in order to discriminate multimedia retrieval systems that use only textual information (such as annotations or a document's title) and methods from IR [Fuhr 2001].

**Definition 2.21 Multimodal retrieval:** *In consequence, we define multimodal retrieval as the retrieval of multimedia documents using different document representations relying on one or more non-exclusive modalities and contextual factors such as the user group or search history. In other words, MIR is a technique for multimedia information retrieval (MMIR) relying on different data that is associated with a document to “exploit the synergy between the various media, including text and context information” [Lew et al. 2006, p. 14].* ◇

In this work, multimodal retrieval always refers to multimodal multimedia information retrieval. Because this text deals mainly with multimodal retrieval, the notions of MIR and MMIR are synonymic throughout the following chapters.

### 2.3.1 Features and Feature Extraction

The term *feature* is pervasive in MIR. Generally speaking, a feature is a document representation [Del Bimbo 1999] based on a particular quality of a multimedia document, e.g., the tempo of a song, the color histogram of an image, or even metadata. The term is used interchangeably with the expression “(content) descriptor” [Feng et al. 2003] – in particular in the scope of MPEG-7<sup>17</sup> [Manjunath et al. 2002].

Features are often subdivided into *high-level* and *low-level* features [Del Bimbo 1999; Schmitt 2006; Blanken et al. 2007]. The “level” of a feature refers to its semantics, i.e., how well it can be understood by a user, or, in other words: “high-level concepts or terms which would be intuitive to the user” [Lew et al. 2006, p. 2]. In CBIR, a similar discrimination is made between visual content descriptors (low-level) and semantic content descriptors (high-level) [Feng et al. 2003]. High-level features can be understood by users because they are at a comprehensible semantic level, e.g., a verbal description of an image (its meaning), while low-level features are mostly statistical information or extractable patterns of the underlying data. Thus, they are very dependent on the type of multimedia document [Blanken et al. 2007]. The main advantage of low-level features is that they can be extracted fully automatically, i.e., without human intervention.

Obviously, this discrimination into high- and low-level features promotes the impact of the semantic gap in MIR as discussed in Section 2.3.3. To bridge this gap, the central hypothesis in MIR is that one can infer a semantical valuable information from a combination of different low-level features [Del Bimbo 1999; Feng et al. 2003; Lew et al. 2006].

For the sake of brevity, this section does not provide an in-depth discussion of all available features available in CBIR and MIR. Instead, only features used in this dissertation are sketched out here. A detailed comparison of commonly used features is presented by Deselaers et al. [2008]. Further comprehensive descriptions of various

<sup>17</sup>MPEG-7 is a standard (ISO/IEC 15938) that defines a machine-readable description of the content of multimedia documents to allow searching in such data.

## 2 Multimedia Information Retrieval

features can be found, e.g., in Del Bimbo [1999]; Feng et al. [2003]; Schmitt [2006], or Blanken et al. [2007], while Eidenberger [2003] concentrates on features in the scope of MPEG-7. Another concise overview of state-of-the-art techniques, features, and their performance can be found in Müller et al. [2010]. To conclude, Kosch & Maier [2010] review current CBIR and MIR systems on the basis of a survey by Veltkamp & Tanase [2002].

In accordance with the usual classification in the field, we further subdivide low-level features into global and local (low-level) features.

**Definition 2.22 Global low-level feature:** *A global feature is extracted from an image considering the image as a whole, non-separable entity. That is, no regions of interest or special key points (see below) are extracted. Instead, a quality of an image, e.g., its texture, is regarded as dependent on the whole image and not on certain regions.* ◇

Typical examples of global features are color histograms that count the frequencies of certain colored pixels over an image, or the Tamura feature [Tamura et al. 1978] that extracts texture properties of an image. If global features are extracted from patches or regions of images, they are called *pseudo-local* throughout this thesis.

**Definition 2.23 Local low-level feature:** *Local features mirror the idea that an image consists of independent parts, the so-called key points, which can be modeled as such. Roughly speaking, these key points are then used to model the most interesting parts of an image instead of the whole image as in the case with global features. In order to discover key points in an image, methods from object recognition are mostly used [Deselaers et al. 2008].* ◇

Representatives of local features are the scale-invariant feature transform (SIFT) [Lowe 2004], its variant speeded up robust features (SURF) [Bay et al. 2006], or the binary robust independent elementary features (BRIEF) [Calonder et al. 2010]. The main advantage of local over global features is their robustness with respect to invariances such as scalings or translations, which *may*<sup>18</sup> lead to a better retrieval performance in comparison to global features [Deselaers et al. 2008]. In fact, their performance has been shown at a number of evaluation initiatives such as ImageCLEF [Müller et al. 2010] or at typical conferences such as the ICMR [Ip & Rui 2012; Jain & Prabhakaran 2013]. Unfortunately, this potential improvement comes at the cost of a more complex representation of such features and higher computational complexity during the retrieval process. This is the main reason why they are neglected in the experiments described in Section 8.4 and 8.5 of this thesis.

### Features Used in this Thesis

Table 2.1 lists the features used in this dissertation. The table specifies whether they are global or local and which characteristics of an image they represent. To facilitate the understanding of this thesis, a full discussion on each feature is omitted. For details refer to the original publications given in Table 2.1. A deep comprehension of the actual algorithms is not needed to understand this thesis.

<sup>18</sup>This effect is dependent on the usage scenario as reported by Deselaers et al. [2008].



Instead, it is important to point out that some features (e.g., BIC or the color histogram) operate only on the color pixel data, while others extract texture or edge information from a grayscale representation of an image (e.g., edge histogram and Tamura). Additionally, some features aggregate textural and color information into one feature, e.g., CEDD and FCTH.

Local features are not examined in this work due to their high computational cost and the nature of the conducted experiments described in Section 8.4.

In addition to the aforementioned low-level features, four high-level features (Table 2.1; 20-23) are used. These features can be extracted directly from an image's Exif<sup>19</sup> metadata. The high-level features contain the time of creation (20,21), the location (22) in global positioning system (GPS) coordinates of a photograph, and the used camera model (23). This information is usually provided by the used camera without user intervention.

Table 2.1: Available Features and Origin; high-level features are shaded gray

R#	Name	Type	Origin
1	Auto Color Correlogram	color-related, global	Huang et al. [1997]
2	BIC	color-related, global	Stehling et al. [2002]
3	CEDD	texture/color-related, global	Chatzichristofis & Boutalis [2008a]
4	Color Histogram	global	512 bin RGB histogram (own implementation)
5	Color Layout*	color-related, global	Cieplinski et al. [2001]
6	Color Structure*	color-related, global	Cieplinski et al. [2001]
7	Dominant Color*	color-related, global	Cieplinski et al. [2001]
8	Edge Histogram*	edge-related, global	Cieplinski et al. [2001]
9	FCTH	texture/color-related, global	Chatzichristofis & Boutalis [2008b]
10	Scalable Color*	color-related, global	Cieplinski et al. [2001]
11	Tamura	texture-related, global	Tamura et al. [1978]
12	Color Histogram (region based)	color-related, pseudo-local	Balko & Schmitt [2012]
13	Contour-based Shape*	global	Cieplinski et al. [2001]
14	Region-based Shape*	global	Cieplinski et al. [2001]
15	Gabor	texture-related, global	Zhang et al. [2000]
20	Date of creation	temporal	Exif
21	Time of creation	temporal	Exif
22	GPS coordinate	spatial	Exif
23	Camera model	metadata	Exif

\* denotes features in the scope of MPEG-7 [Manjunath et al. 2002]

<sup>19</sup>The exchangeable image file format (Exif) is an industrial standard for storing various metadata in image file formats such as JPEG or TIFF.

### 2.3.2 Distance and Similarity Measures

In order to retrieve similar documents to a given query based on a feature, a large number of distance or similarity functions can be used. A (normalized) distance function maps the maximal dissimilarity between two documents to 1 and perfect similarity to 0. Similarity functions invert this semantics, i.e., perfect similarity is expressed by 1.

The features in multimedia retrieval are usually stored in vector form (or sets of vectors in the case of local features), which qualifies them to be combined with methods described for the vector space model, e.g., the cosine measure (see Equation 2.5). More often, distance functions from the class of the Minkowski distances  $L_p$  are used.

**Definition 2.24 Minkowski distance:** *The Minkowski distance  $L_p$  between two  $n$ -dimensional points  $\mathbf{p}_1$  and  $\mathbf{p}_2$ , i.e.,  $L_p : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}_0^+$ , is defined as follows:*

$$L_p(p_1, p_2) = \left( \sum_{i=1}^n |p_1[i] - p_2[i]|^p \right)^{\frac{1}{p}} \quad (2.11)$$

where the parameter  $\mathbf{p}$  determines the type of distance. Widely used variants are the Manhattan distance ( $\mathbf{p} = 1$ ) or the Euclidean distance ( $\mathbf{p} = 2$ ).  $\diamond$

Besides this generic class of distance functions, there are also feature-specific ones [Del Bimbo 1999] or more sophisticated functions as the earth mover's distance (EMD) [Rubner et al. 1998] in use. For instance, the EMD regards the distance calculation between two vectors as a transport problem. Informally, it interprets the two vectors as ways of piling up dirt. It then tries to calculate the minimal cost of transport to turn the piles into each other. Further information about distance and similarity functions and their mathematical properties can be found in Schmitt [2006, Ch. 5f.].

Although this example cannot replace a thorough discussion of these functions, it points out a general problem in MIR: the distance (or similarity) calculation between features can become very complex in terms of computation costs. This problem becomes even more severe with local features.

One way to circumvent this issue is the usage of indices that minimize the amount of needed distance calculations, which are not covered in this work. Overviews about index technologies can be found in Schmitt [2006, Ch. 7] or Feng et al. [2003, Ch. 8].

### 2.3.3 The Semantic Gap in Multimedia Retrieval

As one of the central problems in CBIR and MIR [Datta et al. 2008], Smeulders et al. describe the *semantic gap* fittingly as:

“[...] the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation.”

[Smeulders et al. 2000, p. 5]

As mentioned before, low-level features provide only a weak means for representing high-level semantics in a multimedia document. Hence, the semantic gap between the

internal representation of a multimedia document and its meaning to a user is wide. See Figure 2.2 for an illustration of the representation of the high-level concept “image of a water lily” and its low-level counterpart as a color layout feature based on the pixel values of the image. The information loss is obvious.

This gap widens the more unspecific the contents of a collection that is used for retrieval gets. While the retrieval in limited (or narrow) domains has been shown to be quite successful, e.g., in the field of face recognition [Zhao et al. 2003], the retrieval performance in general image collections (or broad domains) is usually low [Smeulders et al. 2000; Lew et al. 2006].

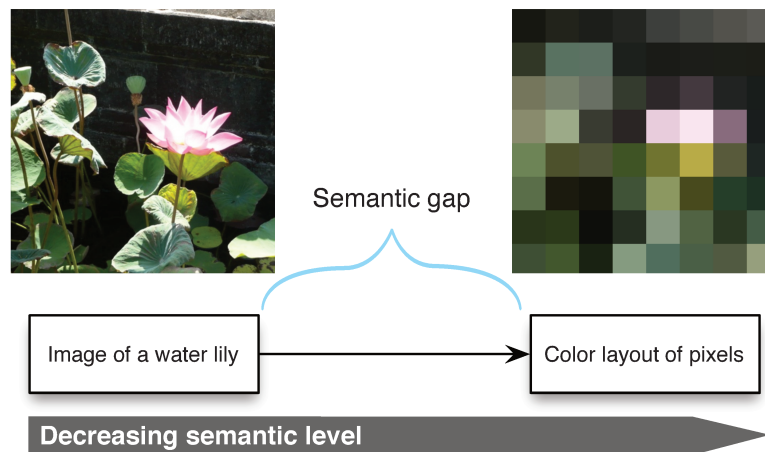


Figure 2.2: The semantic gap in content-based image retrieval

In order to bridge, or, more fittingly, to narrow the semantic gap, *it is assumed that an appropriate combination of multiple low-level features will model higher-level concepts that are understandable to and formulated by the user* [Del Bimbo 1999; Feng et al. 2003; Lew et al. 2006]. Whether the goal of a complete bridging of the gap can be reached is doubtful, as there are reports on glass ceiling effects<sup>20</sup>, e.g., by Aucouturier & Pachet [2004] in the domain of music retrieval. Furthermore, there are no approaches in CBIR, MIR, or MMIR that have succeeded in bridging the gap as yet. Nevertheless, there are numerous contributions that show that narrowing the gap by combining different features is possible even in broad domains [Deselaers et al. 2008; Müller et al. 2010; Tsikrika et al. 2011; Zellhöfer 2012e]. How the actual combination is carried out best remains an open issue in the research community. Though, there is empirical evidence that a large number of multimodal features *might* perform best if one roughly summarizes the results of the last ImageCLEF evaluation benchmarks [Müller et al. 2010; Forner et al. 2012].

<sup>20</sup>“Glass ceiling effect” is a term from economics or sociology originally referring to the effect that a minority is stopped by an invisible and unbreakable barrier from reaching higher career levels, ignoring the individual achievements of the minority. In this work, it refers to the effect that after a number of low-level features are combined, no more significant improvements in the retrieval performance of an IR system can be observed.

### The Query Formulation Problem

As mentioned before, the motivation to submit a query to a MIR system arises from a “recognized anomaly in the user’s state of knowledge” [Belkin et al. 1982, p. 2] regarding an IN. To overcome this anomaly of the state of knowledge, the user provides a query to the system. While formulating this query, the user is confronted with feelings of uncertainty and a lack of understanding regarding the IN [Kuhlthau 1991]. Furthermore, the *mental model* of the user, i.e., a subjective explanation of the system’s function the user interacts with, affects the form of the query input and the provided query details (see Definition 2.4).

Obviously, these effects are also present in an IR scenario. Alas, they get amplified in multimedia retrieval. This is due to the greater width of the semantic gap in this field in comparison to IR and the polysemic nature of images<sup>21</sup>.

If one assumes that a user can express the current IN in keywords, the semantic gap between the high-level IN and the textual representations within the MIR system can be narrowed using techniques from IR. Unfortunately, this approach relies on the presence of annotations for the document corpus, which are likely to be incomplete, subjective, non-reliable, or even missing [Lew et al. 2006; Datta et al. 2008]. In theory, the latter could be solved with the help of automatic annotation that is “widely recognized as an extremely difficult issue” [Datta et al. 2008, p. 40], e.g., because of the needed segmentation of images that is in general still impossible [Deselaers et al. 2008].

Theoretically, users could directly input their query using the MIR’s model of low-level features and distance functions. In practice, users are not able to express their IN at this level [Smeulders et al. 2000]. Interestingly, the same publication claims that users should be able to indicate at least “the class of features relevant for the task, like shape, texture, or both” [p. 17]. The authors continue to postulate that users are also able to select a distance function used during retrieval. We do not agree with their assumptions because the authors provide no justification for their claims. Although Smeulders et al. [2000] are most likely discussing expert user scenarios, their statements remain doubtful – especially, if one considers the tremendous effort a user must put into building a realistic mental model of the low-level features and the accompanying vector operations. This finding is reflected by various interactive approaches that try to assist during query formulation and that are discussed in Section 3.3 in more detail. Frequently, a query-by-example (QBE) approach is used that relieves users from a direct interaction with low-level features. QBE approaches allow a query formulation by providing a sample image and have been used from early on, e.g., in the QBIC system [Flickner et al. 1995].

Regarding the polysemy of images, probably the most used catchphrase in CBIR is “a picture is worth a thousand words”. That is, an image allows a thousand interpretations of it as whole or of parts of it. Regarding the query formulation problem, this means that a provided QBE document can represent innumerable information needs

---

<sup>21</sup>Polysemy is also present in IR and manifests, e.g., in synonyms and homonyms. Nevertheless, this effect can be compensated better with the help of thesauri or ontologies in comparison to MIR as we point out below.

## 2.3 Principles of Multimedia Information Retrieval

that have to be interpreted by the MIR system using various features which are also subject to the semantic gap themselves. For instance, Figure 2.3 might be interpreted as “a Balinese demon statue”, “a statue in the ‘Puri Kerta Gosa’ palace in Klungkung”, “concentration”, a description of a particular event that is important to the user, or a technical property such as being a black and white photograph. As argued before, these interpretations cannot be expressed by the low-level features, and they are also not covered by high-level features such as the date of creation of the photograph.

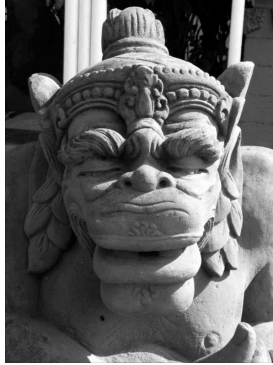


Figure 2.3: A Balinese demon statue

To conclude, CBIR and MIR are still – after decades of research – far from being easy-to-use real-world technologies. This fact is mainly attributed to the semantic gap [Datta et al. 2008] and the related query formulation problem that has been discussed earlier in this section.

## 2 Multimedia Information Retrieval

## 3 Interactive Multimedia Information Retrieval

In the last chapter, IR models following the *standard model* of IR were presented. The standard model (see Figure 2.1) assumes that the interaction between a user and an IR system consists of the specification of a query, an examination of the result documents, and an optional modification of the initial query. This interaction sequence is repeated until the information need is satisfied. In other words, an IR system based on the standard model “assumes almost total control of the interaction, by doing the representation, comparison and modification automatically, and without reference to the user” [Belkin 1993, p. 61].

### 3.1 Interactive Information Retrieval

Interactive information retrieval (IIR) tries to take a broader view onto the interactions during the search process. As a field of research, IIR can be located in the stress field between traditional system-focused IR, e.g., represented by the models described in Chapter 2.2.2, and information, cognitive, and library sciences [Kelly 2009].

While system-centric approaches are primarily focussing on the development or improvement of algorithms and often neglect user needs, the latter fields of research are analyzing cognitive processes and interactions that appear during a user’s attempt to satisfy the current information need (IN). IIR fills the gap between these two worlds by combining their findings. Typical theoretic approaches in IIR are the user-oriented viewpoint (see Section 3.1.1) and the cognitive viewpoint (see Section 3.1.2)

Although out of the main focus of this dissertation, IIR cannot be investigated detached from other fields of research it interacts with. These fields are, in particular, usability and user experience that are addressed by a number of authors, e.g., by Preece et al. [2002]; Shneiderman & Plaisant [2005], and Cooper et al. [2007]; cognitive research focussing on human-computer interaction [Wiedenbeck & Zila 1997]; or visualization [Zhang 2008] to name some.

Interestingly, there are only a few approaches in CBIR and MIR that take the user-centered perspective on retrieval advocated by IIR. While numerous authors, e.g., Belkin [1980]; Bates [1989]; Kuhlthau [1991]; Belkin [1996]; Ingwersen [1996], or Marchionini et al. [2000], contributed to IIR with a (primary) focus on textual retrieval, there are only a few approaches that are based on a theoretical concept of IIR in CBIR, e.g., suggested by Campbell [2000]; Liu et al. [2010], or Urban et al. [2006], whereas the latter is based on Campbell’s model. A more comprehensive review of common IIR approaches can be found in Hearst [2009]; Kelly [2009], or Ingwersen & Järvelin [2005, Ch. 3].

#### 3.1.1 The User-oriented Viewpoint

To put it simply, the user-oriented viewpoint can be regarded as the complement of the system-centric view on IR depicted in Figure 2.1. Whilst system-centric approaches following the standard model assume users and their IN as constant during the retrieval process, the user-oriented models focus on the user's interaction with a mostly black box (or constant) IR system. In consequence, this perspective on IR can be considered broader than the standard model because it regards IR as a goal-oriented interaction aiming at the satisfaction of an IN. Belkin summarizes the main objective of user-oriented approaches fittingly as "to make people's interactions with information the central process of IR, with the other processes and components being seen as providing methods for the appropriate support of such interaction" [Belkin 1996, p. 4]. During this interaction with information, the user undergoes different phases that eventually affect the IN and the interaction with the system. This finding establishes a linkage between the user-oriented and the cognitive viewpoint on IR (see Section 3.1.2). In order to understand the behavioral patterns of users, user-oriented models often rely on socio-psychological methodology such as empirical studies with real users [Ingwersen 1992, cf. pp. 83ff.].

Well-known examples of user-oriented models are Bates' berry picking model [Bates 1989] or Kuhlthau's study on information seeking behavior [Kuhlthau 1991].

The *berry picking model* [Bates 1989] reflects the finding that a user's IN is not static, i.e., it changes over the time during which a user interacts with an IR system. For instance, this change is motivated by learning effects that occur while the user explores a collection. Additionally, the model assumes that the IN is not satisfied by a final set of result documents. Instead, it supposes that parts of the IN are satisfied by documents (or parts of them) that are examined during the search process. Thus, the whole interaction contributes to the solution of the IR task. As Hearst states [Hearst 2009, Ch. 3.3], the berry picking model is supported by several studies, e.g., by Ellis [1989]. Furthermore, its utility for CBIR is shown by an early implementation by Campbell [2000] (see Section 3.3.2).

Kuhlthau's observations leading to the *information search process (ISP)* model [Kuhlthau 1991] are based on a study of a large group of users (385 users at 21 different library sites) and their search behavior. Her study shows that users experience different phases (or stages) during the information search process, e.g., an exploratory phase aiming at learning more about their current task. These phases are linked to different emotions; for example, in the case of exploration to the feeling of insecurity whether they can properly express their IN and confusion [Kuhlthau 1991, cf. p. 366]. Kuhlthau notices also different interaction strategies that are used to overcome the problems of each distinct phase. The change of *information seeking strategies (ISS)* is also described by Belkin [1993] and Ellis & Haugan [1997]. In contrast to Kuhlthau's model, the latter two contributions do not assume a linear sequence of the search stages. Instead, they assume an arbitrary transition between different search strategies. This dynamic nature of search strategy selection is also observed in another study by Reiterer et al. [2000].

The need to reflect these changes in the actual interface is discussed, e.g., by Mar-



chionini et al. [2000] who suggests the usage of with different views supporting interactions based on different search strategies. Furthermore, the findings of Belkin [1993] have led to the BRAQUE (*browsing and query*) system [Belkin et al. 1993]. BRAQUE shares the support for the same information seeking strategies with I<sup>3</sup>R [Croft & Thompson 1987]. Another noteworthy contribution is the THOMAS system [Oddy 1977], which can be considered one of the first systems that allowed users to overcome an interaction scheme based on the specification of a query. Instead, THOMAS allows users to browse document collections in order to satisfy their dynamic IN. Further discussion about overcoming the query-response paradigm is presented by White & Roth [2009].

Another theoretical model that tries to explain this phenomenon and that has been suggested for the usage in CBIR by Liu et al. [2010, 2009] is *information foraging* [Pirolli & Card 1995; Pirolli 2007]. Information foraging links information seeking behaviors with the biological and ecological theory of optimal foraging. That is, information seeking is compared to food gathering strategies of animals analyzing the trade-offs between the value of energy spent on acquiring a food source (an information source such as a document) and the (expected) energy gain (information gain regarding an IN) from it.

Although most of the aforementioned models are based on studies with text documents<sup>22</sup> and the interaction with them, they clearly show that users apply more than one information seeking strategy (ISS) while trying to satisfy their IN. In consequence, a useful retrieval system should support more than one ISS. In accordance with Belkin et al. [1993, p. 328], “we can consider ISSs as types of user interactions within the IR system”. Hence, some possible search strategy implementations in graphical user interfaces (GUI) that are relevant in the field of multimedia retrieval is discussed in Section 3.3.

#### 3.1.2 The Cognitive Viewpoint

Sections 2.2 and 3.1.1 presented two antagonistic perspectives on information retrieval. System-centric approaches focus on the development and tuning of retrieval engines mostly neglecting the user, while the user-oriented viewpoint brings users into the focus of attention at the cost of ignoring the IR system.

A more holistic perspective on IR research is taken by the *cognitive viewpoint*, which is commonly attributed to deMey [1977]. In fact, the cognitive viewpoint “seeks to cover all aspects of IR” [Ingwersen & Järvelin 2005, p. 112]. From a historical perspective, it can be considered as the first coherent alternative to the system-centric IR viewpoint (see Section 2.2) [Kelly 2009, cf. p. 15]. For instance, an early and still central hypothesis arguing on a cognitive level is Belkin’s anomalous state of knowledge (ASK) hypothesis [Belkin 1980] that motivates the user’s need to satisfy an IN (see Definition 2.4).

We omit here the historical development of the cognitive viewpoint in information science because it is covered extensively by a number of authors, e.g., by Larsen [2004]; Ingwersen & Järvelin [2005]; Kelly [2009], and Ingwersen [1992, Ch. 6f.].

<sup>22</sup>For instance, Belkin [1993] and Belkin et al. [1993] explicitly refer to any information-bearing object.

### 3 Interactive Multimedia Information Retrieval

In agreement with the aforementioned authors, Ingwersen & Järvelin [2005] recapitulate the cognitive viewpoint's core ideas as follows:

1. "Information processing takes place in senders *and* recipients of messages;
2. Processing takes place *at different levels*;
3. During communication of information any actor is *influenced* by its past and present experiences (*time*) and its social, organizational and cultural environment;
4. Individual actors *influence* the environment or domain;
5. Information is *situational* and *contextual*."

[Ingwersen & Järvelin 2005, p. 25]<sup>23</sup>

One important assumption of the cognitive viewpoint is the dynamic nature of information seeking as described in Section 3.1.1 in combination with the impact of all actors, e.g., users or IR systems, contributing to the retrieval process. Unlike in the user-oriented viewpoint, actors are not limited to users alone. That is, actors can also be "retrieval engine designers; database producers; algorithm developers; authors; indexers [...]"<sup>24</sup> [Ingwersen & Järvelin 2005, p. 27]. Although these actors might not be present directly during the information seeking activities of a user, their cognitive structures (e.g., experiences, usage context, or emotions) are *represented*. For instance, an algorithm developer's cognitive structure is represented by the actual implementation of a low-level feature affecting the retrieval of multimedia documents.

Additionally, the cognitive viewpoint acknowledges the differences of information processing levels or their "interpretative capability" [Ingwersen & Järvelin 2005, p. 26]. That is, each actor is only able to process information at a distinct level. For instance, a low-level feature describing a document is representing information at a lower semantic level than a user's description of the same document. The resulting gap between these levels has been described as the *semantic gap* in this thesis (see Section 2.3.3).

Finally, the viewpoint recognizes that the involved actors (user and system) interact with each other (and their environment), continuously altering their interpretations of the current IN. This separates the cognitive viewpoint from the user-oriented viewpoint, which assumes the human searcher as the sole recipient of messages. The acknowledgement of the interpretative capability of all actors links the cognitive viewpoint to the concept of *mental models* often referred to in human-computer interaction (HCI). Mental models [Craik 1943] are used by actors to explain the working of a system (or the counter-part they interact with) to themselves. These models do not necessarily mirror the actual operation of this system, i.e., the *conceptual model*. For instance, this effect could be shown in IR by Muramatsu & Pratt [2001]. Given a sufficiently large overlap between the mental and conceptual model, an actor is enabled to interact with the system in a meaningful way. In order to disambiguate messages, the context of the users can be utilized [Ingwersen & Järvelin 2005, cf. p. 26], which often plays no role in basic HCI models.

---

<sup>23</sup>The accentuation has been chosen by the original authors Ingwersen & Järvelin [2005].

<sup>24</sup>This list is not meant to be complete.

To recapitulate, the cognitive viewpoint advocates a “*subjective and profoundly dynamic style of information processing – ideally resulting in continuous changes of models and actual state of knowledge for each device*” [Ingwersen 1992, p. 17]. One model that is based on the cognitive viewpoint is the *principle of polyrepresentation* discussed in the next section.

### 3.2 The Principle of Polyrepresentation

“The principle of polyrepresentation is based on the following *hypothesis*: the more interpretations of different cognitive and functional nature, based on an IS&R [information seeking and retrieval] situation, that point to a set of objects in so-called cognitive overlaps, and the more intensely they do so, the higher the probability that such objects are *relevant* (pertinent, useful) to a perceived work task/interest to be solved, the information (need) situation at hand, the topic required, or/and the influencing context of that situation.”

[Ingwersen & Järvelin 2005, p. 208]<sup>25</sup>

Although the principle of polyrepresentation is based on prior work [Ingwersen 1992], it has been first discussed in full detail in Ingwersen [1996]. Being a cognitive model that has been evolved over time, it is not free from influences of other contributors. The discussion of the principle’s origins is out of the scope of this dissertation. A detailed discussion of the principle and its evolution can be found in Ingwersen & Järvelin [2005], while Larsen [2004] provides a good overview of its influences and its relation to other models of IIR.

As outlined above, the central hypothesis of the principle of polyrepresentation (PoP) of the information space (or documents) is that a document is defined by different representations that can be combined to form a *cognitive overlap* (CO), in which highly relevant documents are most likely to be contained [Larsen et al. 2006]. In contrast to the techno-centric definition given in this thesis (see Definition 2.10), a representation in the sense of polyrepresentation can be the result of a cognitive process of any actor taking part in the information seeking process. Thus, it is not limited to a document representation in vector form when the vector space model is used. By determining the overlap amongst different representations, it is assumed that the inherent uncertainty about relevance assessments in IR, i.e., the relevance judgment an IR system makes about a document, will be compensated – eventually improving the retrieval quality [Ingwersen & Järvelin 2005]. In other words, the principle of polyrepresentation (or “multi-evidence” [Ingwersen & Järvelin 2005, p. 206]) tries to exploit the uncertainties of the representations proactively by examining how many representations “point” to (or provide simultaneous evidence of relevance [Larsen et al. 2006] of) a document. This evidence can be equalized with the probability of relevance of the document (see Section 2.2.2).

In addition, the PoP divides representations in *functionally* and *cognitively* different ones. Functionally different representations are created by a sole actor. This can be

---

<sup>25</sup>The original accentuation has been maintained.

### 3 Interactive Multimedia Information Retrieval

the author of a document who provides an abstract and a title. Cognitively different representations are created by different actors, e.g., a librarian or a retrieval model. As stated above, the principle assumes an IR system as a peer actor during the information seeking process. Consequently, a combination of multiple IR systems as described by Larsen et al. [2006] can also be considered polyrepresentative. Larsen [2004] argues that functional representations also contain a cognitive element, e.g., they are affected by a scientific tradition that defines the form of abstracts or summaries. We fully agree with this conclusion but will use here the original discrimination scheme proposed by Ingwersen [1996] because of its clarity.

Figure 3.1 shows a possible CO in MIR. This figure clearly illustrates the original association of the principle with set theory and Boolean logic. Although the principle is not limited to crisp set theory and Boolean logic alone (see below), the CO is commonly visualized as the *intersection* of different result sets formed by different representations. This kind of visualization is also used in the original publication on the PoP [Ingwersen 1996].

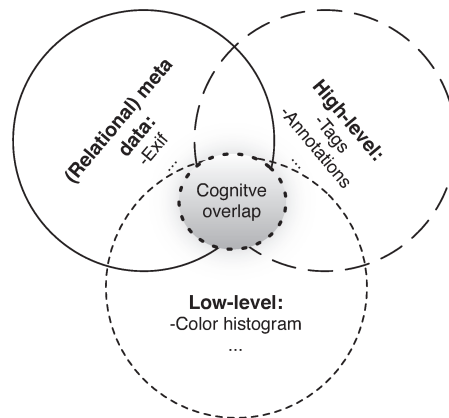


Figure 3.1: Venn diagram of a sample cognitive overlap in MIR

Although the principle of polyrepresentation is not limited to the *information space*, i.e., the system's perspective on IR, this aspect of polyrepresentation attracted much more attention by the research community (see Section 3.2.1 and 4.7) than its view on users and their interaction. We believe that this is due to the simpler implementation potential for the PoP of the information space in comparison to the inclusion of cognitive user processes. Moreover, Larsen [2004] argues similarly by pointing out that further research is needed to understand the user's *cognitive space* in order to develop a functional system.

Nevertheless, Section 6.2.1 discusses the PoP of cognitive space and present a user interaction model based on both aspects of polyrepresentation.

### 3.2.1 The Polyrepresentation Continuum

In order to explore the utility of the principle in best match IR systems that have a far more practical impact on nowadays IR than Boolean IR models [Baeza-Yates & Ribeiro-Neto 2008, cf. p. 38], Larsen [2004] examines the nature of polyrepresentation in more detail. His studies dealing with exact and best match systems lead to the conclusion that there is a “continuum of polyrepresentative solutions” [Larsen & Ingwersen 2005; Larsen et al. 2006, p. 89].

Although the original polyrepresentative reasoning [Ingwersen 1996] is inherently Boolean, Larsen [2004] investigates its application at the two extreme poles of the polyrepresentation continuum (see Figure 3.2). Exact match systems represent the original Boolean proposal by Ingwersen [1996] that rely on calculating the intersection between the result sets based on different cognitive or functional representations. On the opposite side, Larsen places best match systems that are commonly used in multimedia retrieval or IR. These systems cannot use an intersection of result sets because they are based on ranks, e.g., sorted by the probability of relevance of the contained result documents. Instead, they have to *fuse* the distinct ranks of the representations to produce a final rank [Larsen et al. 2006]. Regarding the fusion of ranks in best match systems, Larsen [2004] adverts to a risk regarding the resulting cognitive overlap. For instance, if two representations contribute to the CO and one is only producing a rank with documents of low relevancy (which can be expected in MIR as outlined in Section 2.3.3), the quality of the CO can become very low [Larsen 2004, cf. pp. 36f.]. A comparable problem can occur with the Boolean approach if the intersection of the result set is empty.

The main area for future research regarding the PoP according to Larsen et al. is “to identify flexible and effective matching methods that can generate high quality cognitive overlaps from a variety of the most promising representations” [Larsen et al. 2006, p. 89f.]. This research area is visualized by a cloud in Figure 3.2. Despite this open research question, there is evidence of the utility of the PoP in IR on both poles of the polyrepresentation continuum. Sample studies are available by Skov et al. [2004], who use exact matching, and Larsen et al. [2009], who rely on best matching systems. Both studies show that approaches using the PoP are outperforming the non-polyrepresentative baseline systems, although Larsen et al. [2009] report some instabilities of the retrieval quality. This effect is attributed to the fusion of short ranks, i.e., 15 documents, that have a higher probability that relevant documents “are lurking” [Larsen et al. 2009, p. 653] outside the CO and therefore have no impact on the formation of the CO. Hence, the size of the ranks to be fused has to be sufficiently large.<sup>26</sup>

In conclusion, no matter which kind of matching is performed, “it can be stated that highly structured<sup>27</sup> models of a CO lead to higher precision than little or unstructured

<sup>26</sup>Please note that it might not be sufficient to only increase the number of elements in the rank because the inclusion probability of picking relevant documents follows a hypergeometric distribution that also depends on the number of relevant documents in the collection.

<sup>27</sup>For instance, a query using Boolean connectors to combine representations that aims at incorporating structural properties of the documents (e.g., its title or abstract) [Egnor & Lord 2000].

### 3 Interactive Multimedia Information Retrieval

models such as bag-of-words. These findings resemble the results of Turtle & Croft [1991]; Hull [1997]" [Zellhöfer & Schmitt 2011a, p. 2].

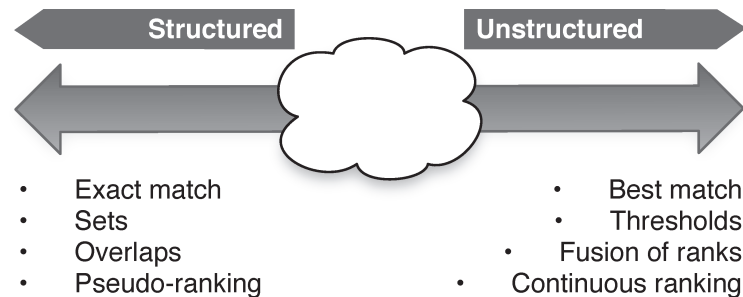


Figure 3.2: The polyrepresentation continuum [Larsen 2004, cf. Fig. 3.3]

## 3.3 Interactive Support of Information Seeking Strategies

Over the decades of IR and IIR research tradition, numerous information seeking strategies (ISS) have been described. To support these ISSs, various user interaction designs have been proposed [Belkin et al. 1993], e.g., directed and exploratory approaches. In this section, ISSs and their supportive user interfaces (UI) that have gained attention in the fields of CBIR and MIR are presented.

### 3.3.1 Directed Search and Query-By Approaches

Probably the oldest information seeking pattern is *directed search*, which is still the predominant information seeking pattern in both information retrieval and databases.

**Definition 3.1 Directed search:** *In directed search, a user provides a query aiming at retrieving specific information relating to the current information need.* ◇

The need for directed search can be considered to be the driving force behind DB and IR. In IR, the query is a central part in the system-centric viewpoint outlined in Section 2.2.

#### Query by Language

The most intuitive way to interact with an IR system is natural language. For instance, Bilal [2000] presents evidence that users that are new to IR systems tend to use natural language questions. Nevertheless, current IR systems mostly use subsets of the natural language, e.g., in the form of keywords. As outlined in Section 2.3.3, such approaches are affected by the semantic gap.

An alternative way to formulate queries is the usage of restricted vocabularies or a language with a high level of formalization. These artificial languages can serve experts

### 3.3 Interactive Support of Information Seeking Strategies

well while they repel layperson users. For instance, SQL<sup>28</sup> is the most often used language in relational databases. It offers powerful means to experts to query tremendous amounts of data stored in a relational DB. In order to restrict the results, it supports various “filters” in form of predicates that are evaluated in a three-valued logic (true, false, unknown) (see Section 2.2.2). For instance, the following query returns all rows from a table `irmodels` that are set as theoretic or probabilistic.

Listing 3.1: Sample SQL query

```
1 SELECT * FROM irmodels
2   WHERE settheoretic=TRUE OR probabilistic=TRUE
```

In order to make keyword-based queries more precise, IR systems are also often supporting Boolean operators, e.g., the InQuery system [Callan et al. 1992]. As said before, the usage of Boolean operators bears the risk of erroneous query input resulting in too many or too few results [Soergel 1985]. Another issue arises if Boolean operators are automatically selected to connect the keywords, which has been shown to irritate non-expert users [Muramatsu & Pratt 2001]. In the same study, comparable effects could be observed with stop word removal and other techniques transforming the user’s query into the representation used within the IR model.

#### Query by Example

Because of the query formulation problem (see Section 2.3.3) and the issues with the exploitation of annotations in CBIR (see Section 2.3), *query by example* (QBE) has attracted attention since the early days of multimedia retrieval research. For instance, Flickner et al. [1995] present a system that uses sample images instead of a language-based query, the *query by image content* (QBIC) system. Because of the ease of use of this approach, the QBE has had and still has a tremendous impact on nowadays research in CBIR and MIR. Prior to its usage in MIR, a similar approach has been suggested in the DB field to support non-expert users during query formulation [Zloof 1975]. In DB, users can formulate predicates, joins etc. using the visualization of a table skeleton. Figure 3.1 illustrates the approach using the SQL query from above, whereas P. stands for the print of all columns and the two used rows denote a disjunction on the both predicates.

Table 3.1: Query by example as a DB query approach

<i>irtable</i>	name	settheoretic	probabilistic	algebraic
P.		TRUE		
P.			TRUE	

Although QBE in CBIR and MIR is often carried out with only one QBE document, it is not limited to it. In principle, multiple QBE documents can serve as positive samples [Assfalg et al. 2000b; Tahaghoghi et al. 2002] or users can provide positive and negative

<sup>28</sup>Structured Query Language, the de facto standard query language in relational DB systems.

### 3 Interactive Multimedia Information Retrieval

samples [Assfalg et al. 2000a]. The aforementioned approaches have in common that they consider the whole provided QBE document as relevant (or irrelevant). In contrast, Awang Iskandar et al. [2007] suggest to use only user-definable subregions of the document as the query.

If no QBE document is available, the approach can be extended to incorporate sketches (query by sketch) [Eitz et al. 2009] or be based on user-definable colors that are expected in the result documents as in subsequent versions of QBIC [Flickner et al. 1995].

Because of its simplicity, the query-by approach has also been transferred to other domains, e.g., to music retrieval in form of query by humming [Ghias et al. 1995].

To conclude, there are also approaches that feature CBIR-based QBE queries in combination with the visual formulation of DB queries, e.g., by Schmitt et al. [2005].

#### 3.3.2 Exploratory Search

The directed search approaches discussed in the last section all require an initial query to start the retrieval process. That is, they require the user to have a concept of their current IN and means to formulate it in a way that can be interpreted by the IR system. The fact that this is not always possible has been discussed in Sections 2.2.1 and 2.3.3 and observed by the studies leading to various IIR models presented in Section 3.1.

For instance, Kuhlthau [1991] discusses that users that are insecure about the formulation of their IN in form of a query tend to explore collections in order to learn about their contents and the current task.

Although there are numerous studies showing the utility of exploratory search, end-user systems most often rely on directed search alone [White & Roth 2009, cf. p. 13].

An alternative way to discriminate exploratory from directed search is to focus on their different goals for precision and recall (see Section 8.1.1). While a directed search typically aims at achieving high precision (i.e., a low number of irrelevant documents is amongst the retrieved documents), exploratory searches usually have higher recall values (i.e., a maximum of relevant documents should be retrieved).

Exploratory search patterns try to tackle these challenges by offering an alternative approach towards information seeking.

**Definition 3.2 Exploratory search:** *Exploratory search patterns are not based on a (user-provided) query but allow the exploration of a document collection, e.g., based on the similarity between the contained documents. Exploratory search is often utilized if users are unsure about their current IN and are therefore unable to specify a query which can then be answered by the retrieval system. While exploring the document collection, users learn about its contents by examining and comparing documents instead of relying on a system-generated result.* ◇

In particular, exploratory approaches are linked to visualization and clustering techniques because they provide the means for navigating through and for understanding complex information spaces. In the following sections, we present two exploratory search patterns – *browsing* and *faceted navigation* – that are often used in MIR.



## Browsing

Roughly speaking, browsing can be defined as a movement through an information space along (not necessarily visible) connections between documents. A well-known example is the World Wide Web, which implements browsing along hyperlinks between HTML<sup>29</sup> documents. In MIR, browsing is often understood as the navigation through a two or three dimensional space that is used to visualize a form of similarity between documents by placing them adjacently. One way to establish a spatial connection between two documents is to calculate their distance regarding a given low-level feature, e.g., a color histogram, in a high-dimensional space (see Section 2.3.2) and map the results to a 2D visualization, e.g., by using a *self-organizing map* (SOM) [Kohonen 1995]. A SOM (or Kohonen map) is an unsupervised neural network that is commonly used to map a high-dimensional input space to a two-dimensional output space (the map). The utility of SOM has been shown in both CBIR [Laaksonen et al. 1999] and IR [Kohonen et al. 2000].

Heesch [2008] provides an extensive survey on browsing techniques used in the field of CBIR. In addition, Zhang [2008] approaches browsing in IR from a visualization perspective taking a mathematical point of view. Finally, Hearst [2009] gives a holistic overview over the field including other ISS.

One browsing model that gained a lot of attention in CBIR is the *ostensive model* (OM) [Campbell 2000]. Basically, it is an implementation of the berry picking model by Bates [1989]. Based on an initial document (the “starting point”, see Figure 3.3), documents which might be relevant to the user’s IN are shown to the user. By clicking one document, the user provides relevance feedback to the system which reacts with a new set of possibly relevant documents (the “candidates”). In consequence, the user moves along a path through the information space.

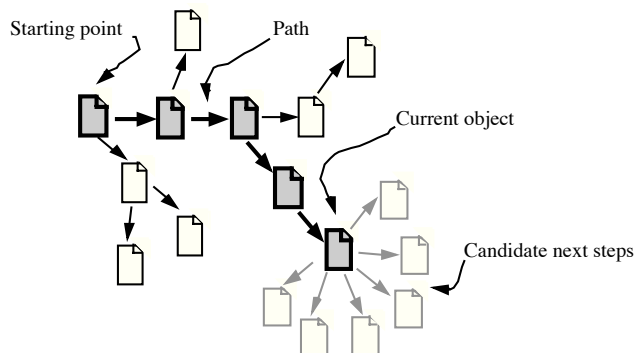


Figure 3.3: Browsing in the ostensive model [Campbell 2000]

Particularly interesting about the OM is that it models a dynamic IN by introducing relevance profiles. These profiles decrease the impact of older documents along the

<sup>29</sup>The hyper-text markup language (HTML) is used to design webpages in the World Wide Web.

### 3 Interactive Multimedia Information Retrieval

path on the system's representation of the current IN. In other words, the IN "ages" if it is not strengthened by additional evidence during the browsing. While the original OM is based on text-based features alone, Urban et al. [2006] extend the approach by additionally utilizing content-based low-level features. A variant of the OM has also been implemented by Google as the "image swirl" (see Figure 3.4) <sup>30</sup>.

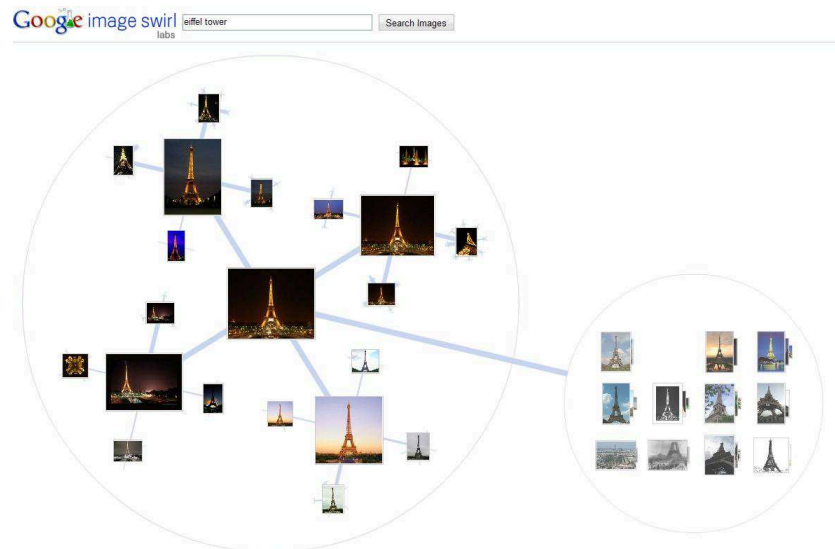


Figure 3.4: Browsing with the starting point "Eiffel tower" in Google's image swirl<sup>31</sup>

#### Faceted Navigation

In contrast to browsing that offers a weakly structured exploration of the information space, faceted navigation (FN) or faceted browsing relies on ordering criteria – the so-called *facets* – to assist users during exploration. Facets are based on attributes or attribute values that are shared by a subset of the retrieved documents or information objects and can therefore be used to filter the contents of a collection with respect to these facets. Normally, facets are orthogonal, i.e., each attribute describes an independent aspect of the document, and often hierarchical, i.e., they describe one aspect at different granularities [Hearst 2009, cf. p. 190ff.]. For instance, the facet "publication year" could be refined from "century" to "decade"<sup>32</sup>. Facets are typically created manually to ensure that they are understandable by users [White & Roth 2009, cf. p. 45]. Nevertheless, an automatic construction with the help of clustering algorithms can

<sup>30</sup><http://image-swirl.googlelabs.com/>; as downloaded at the 18th November 2009. The service has been discontinued in 2011.

<sup>31</sup><http://googleresearch.blogspot.de/2009/11/explore-images-with-google-image-swirl.html>; as downloaded at the 29th January 2013.

<sup>32</sup>In a way, facets resemble dimensions in data warehousing.

### 3.3 Interactive Support of Information Seeking Strategies

also be feasible and can reveal structures within the retrieved documents that were not known beforehand. In any case, the automatic construction of facets comprises the risk that they are not comprehensible to the user<sup>33</sup>. Furthermore, the utility of FN is negatively affected by facets that poorly match the user's mental model, e.g., if they are too specific or use a language that is likely to be misinterpreted by the user. In other words, facets containing documents that are not expected to be part of a given facet by the user lead to confusion. The same holds true for an incorrect assignment of documents to facets, e.g., a facet "red cars" should only contain red cars and no purple ones in order to be usable.

Another factor that might affect the usability of FN are too many facet hierarchies as they might confuse users [Hearst 2009, cf. p. 195].

FN is often combined with keyword search. In this scenario, the keyword search acts as a first filter on the retrieved documents that can then be explored with the help of facets. One of the first uses of FN in MIR is the faceted metadata approach in the Flamenco project [Yee et al. 2003], which also supports keyword search. In this system, FN is used to "navigate along conceptual dimensions" [Yee et al. 2003, p. 401] formed by metadata in a fine arts database. Figure 3.5 illustrates the user interface of the system. Instead of using low-level features directly, the authors rely on metadata-based facets because of the better comprehensibility of textual labels in comparison to low-level feature-based criteria as shown by Rodden et al. [2001].



Figure 3.5: Faceted navigation in the Flamenco fine arts IR system; facets on the left and results on the right [Yee et al. 2003, Fig. 2]

<sup>33</sup>All techniques based on clustering have this risk in common.

### 3 Interactive Multimedia Information Retrieval

As mentioned before, the main advantage of FN over browsing is that users learn about ordering criteria of the information space. This knowledge can then later be used to conduct a directed search (see Section 3.3.1) based on these criteria. With regard to the principle of polyrepresentation (see Section 3.2), one can interpret each representation as a different facet, although not necessarily an orthogonal one. This interpretation is similar to the viewpoint by Larsen [2004, cf. pp. 36f.]. Nevertheless, it is possible to imagine mapping different representations directly on facets during user interaction as described in Section 6.2.

Unlike browsing interfaces, FN has reached the industrial mainstream, e.g., in online shops, as it allows consumers to learn about goods and to decide quickly between different attributes such as the size of a computer screen. Further reviews of systems that support faceted navigation are, e.g., available by White & Roth [2009, Sec. 4.2], Hearst [2009, Sec. 8.6], Morville & Callender [2010, Ch. 4], or Russell-Rose & Tate [2013, Ch. 7]. An in-depth presentation of faceted classification is available in Taylor [2006, Ch. 14].

## **Part II**

# **Learning User-specific Weights in Logic-based Queries**



## 4 A Quantum Logic-based Model for Multimedia Information Retrieval

In his seminal work, van Rijsbergen argues for a unified geometric view on IR using the mathematical formalism of quantum mechanics (QM), i.e., a way “to combine probability, logic and vector spaces into one formalism” [van Rijsbergen 2004, p. 1]. Considering his own contributions to the field and the logical uncertainty principle in particular [van Rijsbergen 1986a, b], this work seems as the logical result of van Rijsbergen’s objective to establish a sound mathematical foundation of IR.

In order to establish a well-grounded theory of IR, van Rijsbergen utilizes the core concepts of QM to formalize the central aspects of IR. By interpreting documents as normalized *state vectors* that are embedded into a *Hilbert space*, i.e., the information space, and mapping queries to *subspaces* of this space, van Rijsbergen reveals a way to interpret the calculation of the probability of relevance as a geometric operation: the *projection* of a state vector to a subspace that represents a query.

Based on this idea, Schmitt [2008] developed a QM-based query language, the *commuting quantum query language* (CQQL), which is used as a query language for MIR in this thesis.

How the core concepts of QM are defined and used by Schmitt [2008] to build a retrieval model leading to CQQL is presented in Section 4.1. The subsequent Section 4.2 introduces CQQL and its evaluation on a formal level. Section 4.3 discusses how queries in CQQL can be personalized using weights. The usage of CQQL in MIR is illustrated with the help of an example in Section 4.4 that points out strengths and weaknesses of the approach.

Section 4.5 relates CQQL to other IR models that have been introduced in Section 2.2.2. The following Section 4.6 reveals how the language can be interpreted as an implementation of the principle of polyrepresentation (see Section 3.2).

The chapter concludes with Section 4.7 that contains a discussion of CQQL’s relation to other polyrepresentative approaches.

### 4.1 Theoretical Concepts behind the Commuting Quantum Query Language (CQQL)

The motivation for van Rijsbergen [2004] to exploit the mathematical formalisms of QM lies in his observed lack of theory in IR. In contrast, Schmitt [2008] approaches QM from the database standpoint in order to find a way to combine traditional Boolean DB queries with concepts of similarity, proximity, or probability.

## 4 A Quantum Logic-based Model for Multimedia Information Retrieval

At first sight, these motivations seem antagonistic because the research fields of DB and IR are traditionally separated. In any case, the disciplines have more in common than their focus on the retrieval of information. For instance, van Rijsbergen [1986b] and Nottelmann & Fuhr [2003] interpret IR as a generalization of DB (see Section 2.2.2).

Because of the DB background of Schmitt's work, he emphasizes other points than van Rijsbergen, although both authors rely on the same mathematical foundation. For instance, Schmitt focuses on a compensation of issues inherited from the usage of fuzzy logic [Zadeh 1965, 1988] in the field of DB. This aspect is covered separately in Section 4.5.3. Furthermore, Schmitt postulates three requirements for a query language linking DB and IR:

1. *database query support*: The language must be relational complete.
2. *information retrieval support*: The language must enable us to formulate and evaluate retrieval and proximity query terms.
3. *unifying theoretical framework*: There must exist just one underlying unifying theoretical framework."

[Schmitt 2008, p. 39]

Due to Schmitt's focus on relational databases, he first requires a relational complete query language [Codd 1970; Garcia-Molina et al. 2000]. That is, a language with the same expressive power as the relational calculus or relational algebra [Codd 1970]<sup>34</sup>. Broadly speaking, this requirement is crucial because it allows the execution of such a language on the most widely used relational DBMS and thus an integration of such systems into MIR.

The second point is surprising from an IR point of view but necessary in terms of a relational DB. In the relational model, each condition (or proposition in a logical sense, see Appendix A.1.4) is evaluated as one out of three truth values, i.e., *true*, *false*, or *unknown* (see Section 2.2.2). In IR, one is confronted, e.g., with probabilities of relevance (POR) of a document with respect to a query. This POR is typically in the interval  $[0, 1]$ . Hence, the query language is required to deal with such values in addition to the aforementioned.

The third requirement is closer to van Rijsbergen's demand for a unifying theoretical framework. Although one can argue that it is possible to consider relational and IR parts of a query separately and fuse their results later, it is necessary to establish a sound unifying theory if one wants to reason consistently with conditions derived from DB and IR at the same time. In the end, this characterizes an elegant approach towards (M)IR.

To satisfy these requirements, Schmitt [2008] builds on the QM-based foundations laid by van Rijsbergen [2004] and extends them to address DB-specific issues by exploiting mechanisms of quantum logic [Birkhoff & von Neumann 1936].

This research resulted in the *commuting quantum query language* (CQQL), which is presented in this chapter.

---

<sup>34</sup>To put it simply, variants of a Boolean first-order logic that have been shown to have the same expressive power [Codd 1972].



## 4.1 Theoretical Concepts behind the Commuting Quantum Query Language

As said before, the works by Schmitt [2008] and van Rijsbergen [2004] rely on the mathematical formalism of QM. This dissertation is not intended to give an introduction or discussion on QM. Thus, it can only provide a general and brief introduction of the core concepts that are needed for an understanding of CQQL. For more information about the definition and theoretical foundation of CQQL, refer to Schmitt [2008] or Schmitt et al. [2008] for its application in the field of MIR. For a more detailed overview of QM, please refer to the appendices of van Rijsbergen [2004], who also provides an excellent, commented bibliography on the topic. A brief introduction also addressing quantum logic (including the four postulates of QM) is available in Schmitt [2008].

In the following, some core concepts of QM are sketched. All definitions as well as the remainder of this thesis use the Dirac (or bra-ket) notation, which is described separately in Appendix A.1.6.

**Definition 4.1 Hilbert space:** *A Hilbert space  $\mathbf{H}$  is a vector space (in the mathematical sense) with dimensions of finite or infinite number. Furthermore, a Hilbert space requires the definition of the inner product of two vectors to measure length and angle between the two. A Hilbert space must be complete. Roughly speaking, this means that the space must have limits in order to allow the usage of calculus techniques (see van Rijsbergen [2004, p. 103] and Appendix A.1.3).*  $\diamond$

**Definition 4.2 State vector:** *The state of a quantum system at a specific point in time is fully characterized by a normalized state vector  $|\varphi\rangle$  that is embedded into a separable, complex Hilbert space  $\mathbf{H}$ .*  $\diamond$

**Definition 4.3 Projector:** *A projector  $\mathbf{p} = \sum_i |i\rangle\langle i|$  is a symmetric, idempotent linear operator that is defined over a set of orthonormal<sup>35</sup> vectors  $|i\rangle$ . Each projector is bijectively associated with a vector subspace of  $\mathbf{H}$ . Hence, the multiplication of  $\mathbf{p}$  with  $|\varphi\rangle$  means to project the vector onto the subspace described by the projector.*  $\diamond$

**Definition 4.4 Quantum measurement:** *Given that each measurable, physical property can be expressed by a projector  $\mathbf{p}$ , the probability that  $\mathbf{p}$  is measured with a system  $|\varphi\rangle$  is:*

$$\langle\varphi|\mathbf{p}|\varphi\rangle = \langle\varphi|\left(\sum_i |i\rangle\langle i|\right)|\varphi\rangle = \sum_i \langle\varphi|i\rangle\langle i|\varphi\rangle \quad (4.1)$$

*More intuitively, the resulting probability value can be seen as “the squared length of the state vector  $|\varphi\rangle$  after its projection onto the subspace spanned by the vectors  $|i\rangle$ . Furthermore, due to normalization (see Definition 4.2), the probability value, furthermore, equals geometrically the squared cosine of the minimal angle between  $|\varphi\rangle$  and the subspace represented by  $\mathbf{p}$ ” [Schmitt et al. 2008, p. 3].*

*This probabilistic interpretation of quantum measurements discriminates QM from traditional mechanics in which each measurement yields a definite outcome.*  $\diamond$

**Definition 4.5 Tensor product:** *In order to combine different 2-dimensional quantum systems into one, the tensor product  $\otimes$  is used. The tensor product of two state vectors is defined*

<sup>35</sup>That is, the vectors are unit vectors and orthogonal to each other.

## 4 A Quantum Logic-based Model for Multimedia Information Retrieval

as follows:

$$|x\rangle \otimes |y\rangle \equiv \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \otimes \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \equiv \begin{pmatrix} x_1y_1 \\ x_1y_2 \\ x_2y_1 \\ x_2y_2 \end{pmatrix} \quad (4.2)$$

◇

In accordance with van Rijsbergen [2004], the main idea to use CQQL in the context of MIR is to interpret documents (or database tuples) as state vectors and queries as projectors embedded into a Hilbert (vector) space<sup>36</sup>. In order to assess the POR of a document with respect to a query, a quantum measurement is conducted. Table 4.1 relates the concepts of DB querying with their QM counterpart.

Table 4.1: Related concepts from database querying and quantum mechanics [Schmitt 2008, Tab. 2]

Database Querying	Quantum Mechanics
Database tuple	State vector
Query	Projector
Query processing	Quantum measurement
Truth values	Probability values
Boolean logic	Quantum logic

As CQQL is meant to be a logical query language, quantum logic (QL) has to be introduced. Originally, QL has been presented by von Neumann [1932] and Birkhoff & von Neumann [1936] and can be seen as a logic for reasoning about the formal structures of the Hilbert space.

To start with, let  $P$  be a set of all projectors in a Hilbert space  $H$  with more than two dimensions. As said in Definition 4.3, each  $p \in P$  is bijectively related to a closed subspace  $vs_p$  via:  $p \leftrightarrow vs_p(H) = \{p|\varphi \mid |\varphi\rangle \in H\}$  [Schmitt 2008, cf. Sec. 3]. The subset relation of the subspaces forms a partially ordered set (poset) (see Appendix A.1.1) over  $P$ , in which  $p_1 \leq p_2 \Leftrightarrow vs_{p_1}(H) \subseteq vs_{p_2}(H)$  holds.

Hence, we obtain a lattice, i.e., a structure in which absorption, associativity, and commutativity is fulfilled, with two binary operations: *meet* ( $\sqcap$ ) and *join* ( $\sqcup$ ) (see Appendix A.1.3).

**Definition 4.6 Meet (lattice operator):**

$$p_1 \sqcap p_2 \rightsquigarrow p \leftrightarrow vs_p(\mathbf{H}) \equiv vs_{p_1}(\mathbf{H}) \cap vs_{p_2}(\mathbf{H}) \quad (4.3)$$

◇

**Definition 4.7 Join (lattice operator):**

$$p_1 \sqcup p_2 \rightsquigarrow p \leftrightarrow vs_p(\mathbf{H}) \equiv \text{closure}(vs_{p_1}(\mathbf{H}) \cup vs_{p_2}(\mathbf{H})) \quad (4.4)$$

where **closure**( $\cdot$ ) yields the set of all possible vector linear combinations.

◇

<sup>36</sup>For simplicity, Schmitt [2008] restricts the space to  $\mathbb{R}^n$ .

## 4.1 Theoretical Concepts behind the Commuting Quantum Query Language

Finally, an unary *orthocomplement* ( $\neg$ ) is defined, which can be interpreted as the negation operator.

**Definition 4.8 Orthocomplement (lattice operator):**

$$\neg p_1 \rightsquigarrow p \leftrightarrow vs_p(\mathbf{H}) \equiv \{|\phi\rangle \in \mathbf{H} \mid \forall |\phi\rangle \in vs_{p_1}(\mathbf{H}) : \langle \phi | \phi \rangle = 0\} \quad (4.5)$$

That is, the orthocomplement in quantum logic is defined as  $\neg \mathbf{p}_1 \equiv I - \mathbf{p}_1$  encompassing all orthogonal projectors to  $\mathbf{p}_1$ .  $\diamond$

Unfortunately, quantum logic violates the law of distributivity. Hence, it does not constitute a Boolean algebra (see Appendix A.1.5) as implied by the requirement for CQQL to be relational complete (see above). For the proof, see Schmitt [2008].

Thus, it is necessary to define a sublattice of quantum logic that is compatible with the laws of Boolean algebra. The identification of such sublattices becomes possible by taking *commuting projectors* into account.

**Definition 4.9 Commuting projectors:** “Two projectors  $\mathbf{p}_1$  and  $\mathbf{p}_2 \in \mathbf{H}$  are called commuting projectors if and only if  $\mathbf{p}_1\mathbf{p}_2 = \mathbf{p}_2\mathbf{p}_1$  holds.” [Schmitt 2008, Def. 2].  $\diamond$

According to the laws of linear algebra, two projectors  $p_1 = \sum_i |i\rangle\langle i|$  and  $p_2 = \sum_j |j\rangle\langle j|$  commute if the vectors  $|i\rangle$  and  $|j\rangle$  are basis vectors of the same orthonormal basis of the underlying vector space. Furthermore, “if two projectors commute then their join corresponds to the union of the respective sets of underlying base vectors and their meet to the intersection” [Schmitt et al. 2009, p. 424]. As shown in Schmitt [2008], we know that the sublattice over every equivalence class comprising commuting projectors is compatible with the laws of Boolean algebra. In other words, all projectors over a given orthonormal basis follow the laws of Boolean algebra because distributivity holds in this particular case. This conclusion is also known as the *Foulis-Holland theorem* [Birkhoff 1993, cf. p. 53].

To conclude, QL can be regarded as a generalization of Boolean algebra violating the law of distributivity. In order to be relational complete, a query language is required to follow the laws of a first-order predicate logic, which is a Boolean algebra. As a consequence, the expressive power of QL has to be restricted to only allow commuting projectors – hence the name for Schmitt’s suggested query language: the commuting quantum query language.

For the sake of brevity, this dissertation does not discuss the full derivation of CQQL from QL nor how database or retrieval conditions are mapped to state vectors etc. in detail. Instead, it concentrates on the relevant evaluation rules of CQQL to show how the query language can be used in MIR. For a complete theoretical examination, refer to Schmitt [2008], which is the central publication on CQQL and its origin in QL. For the remainder of this thesis, it is only important to recognize the mappings recapitulated in Table 4.1.

## 4.2 Construction and Arithmetic Evaluation of CQQL Queries

This section of the dissertation sketches the construction and evaluation of CQQL based on the discussion presented in Schmitt et al. [2008] and Schmitt [2008].

To follow the design of CQQL in detail, a certain understanding of basic database concepts that have been presented by a number of authors mentioned before, e.g., by Abiteboul et al. [1996] or Garcia-Molina et al. [2000], is recommended.

**Definition 4.10 Basic CQQL elements:** Let  $\mathbf{A} = \{\mathbf{a}_j\}$  be a finite set of attributes and  $\mathbf{AC} = \{ac_i(\mathbf{a}_j)\}$  be a finite set of atomic attribute conditions with the following forms:

- ' $\mathbf{a}_j = value$ ' (Boolean or database condition), or
- ' $\mathbf{a}_j \approx value$ ' (retrieval or proximity condition).

The condition set  $\mathbf{AC}$  is called commutative if

$$\forall ac_{i_1}(a_{j_1}), ac_{i_2}(a_{j_2}) \in \mathbf{AC} \mid (type(ac_{i_1}(a_{j_1})) \neq d \wedge type(ac_{i_2}(a_{j_2})) \neq d) \rightarrow j_1 = j_2$$

holds and where **type** returns **d** for a database condition and **p** or **r** for a proximity or retrieval condition, respectively.  $\diamond$

Commutativity on conditions means that no two proximity or retrieval conditions ' $a_j \approx value_1$ ' and ' $a_j \approx value_2$ ' with  $value_1 \neq value_2$  on the same attribute  $a_j$  are allowed.

In our quantum encoding, we assign implicitly to every atomic condition  $ac_i(a_j)$  a set of orthonormal vectors:  $vs(ac_i(a_j)) = \{|ac_i^1\rangle, \dots, |ac_i^k\rangle\}$ , from which a projector  $p_{ac_i(a_j)}$  is constructed.

**Lemma 4.1:** Let  $\mathbf{AC}$  be a commutative set of atomic conditions over  $\mathbf{A} = \{\mathbf{a}_j\}$ . The set  $\mathbf{CVS}(\mathbf{AC}) = \bigcup_{ac_i(\mathbf{a}_j) \in \mathbf{AC}} vs(ac_i(\mathbf{a}_j))$  is a set of mutually orthonormal vectors.

The lemma is a direct consequence of how the mapping function  $vs$  is realized (for more details and the necessary proofs, see Schmitt [2008]). It is essential because it means that the projectors of all conditions over  $\mathbf{AC}$  are mutually commuting and we therefore obtain a projector lattice obeying the rules of Boolean algebra (see Section 4.1, particularly Definition 4.9).

**Definition 4.11 CQQL conditions:** Let  $\mathbf{AC}$  be a commutative set of atomic conditions on  $\mathbf{A} = \{\mathbf{a}_j\}$ . Then a CQQL condition  $\varphi$  is recursively defined by:

$$\varphi \stackrel{def}{=} ac_i(a_j) \in \mathbf{AC}, \tag{4.6}$$

$$\varphi \stackrel{def}{=} (\varphi_1 \wedge \varphi_2), \tag{4.7}$$

$$\varphi \stackrel{def}{=} (\varphi_1 \vee \varphi_2), \text{ and} \tag{4.8}$$

$$\varphi \stackrel{def}{=} (\neg \varphi_1) \tag{4.9}$$

## 4.2 Construction and Arithmetic Evaluation of CQQL Queries

where  $\varphi_1, \varphi_2$  are CQQL conditions with

$$\begin{aligned} vs(\varphi_1 \wedge \varphi_2) &= vs(\varphi_1) \cap vs(\varphi_2) \subseteq CVS(AC) \\ vs(\varphi_1 \vee \varphi_2) &= vs(\varphi_1) \cup vs(\varphi_2) \subseteq CVS(AC) \\ vs(\neg\varphi_1) &= CVS(AC) \setminus vs(\varphi_1) \subseteq CVS(AC) \end{aligned}$$

◇

**Theorem 4.1 CQQL is a Boolean algebra:** *All CQQL conditions over a commutative set of atomic conditions together with conjunction, disjunction, and negation form a Boolean algebra.*

**Proof 4.1 CQQL is a Boolean algebra:** *The function  $vs$  maps every condition bijectively to a subset of  $CVS(AC)$ . Conjunction, disjunction, and negation are mapped to the corresponding set operations. A set combined with these standard set operations constitutes a Boolean algebra. □*

### Arithmetic Evaluation of CQQL

A very interesting feature due to the CQQL's theoretical foundation is also presented by Schmitt [2008]: a CQQL condition (or query) in a specific syntactical form can be evaluated by means of simple, straightforward arithmetics. For the sake of simplicity, the needed syntax transformation algorithm is omitted. A full description of the algorithm is available by Schmitt [2008]. The transformations carried out by the algorithm become possible because CQQL is compatible with the laws of a Boolean algebra. Hence, it is possible to transform every possible CQQL condition into the required syntactical form, on which the following rules can be applied using solely Boolean transformation rules [Schmitt et al. 2008].

**Definition 4.12 CQQL evaluation basics:** *Let  $\varphi_i$  be a CQQL condition and  $\mathbf{d}_j^r \in \mathbf{D}^r$  as defined in Definition 2.10. Given that  $\mathbf{q}$  is a CQQL query (or condition) in the required syntactical form and of the following structure ' $\mathbf{q} = \varphi_1 \odot \varphi_2$ ', where  $\odot$  denotes a logical connector, then  $eval()$  is recursively defined as follows. ◇*

**Definition 4.13 CQQL conjunction (evaluation):**

$$eval(\varphi_1 \wedge \varphi_2, \mathbf{d}_j^r) \rightsquigarrow eval(\varphi_1, \mathbf{d}_j^r) \cdot eval(\varphi_2, \mathbf{d}_j^r) \quad (4.10)$$

◇

**Definition 4.14 CQQL disjunction (evaluation):**

$$eval(\varphi_1 \vee \varphi_2, \mathbf{d}_j^r) \rightsquigarrow eval(\varphi_1, \mathbf{d}_j^r) + eval(\varphi_2, \mathbf{d}_j^r) - eval(\varphi_1, \mathbf{d}_j^r) \cdot eval(\varphi_2, \mathbf{d}_j^r) \quad (4.11a)$$

$$eval(\varphi_1 \vee \varphi_2, \mathbf{d}_j^r) \rightsquigarrow eval(\varphi_1, \mathbf{d}_j^r) + eval(\varphi_2, \mathbf{d}_j^r) \quad (4.11b)$$

*Equation (4.11a) applies when  $\varphi_1$  and  $\varphi_2$  are not exclusive, while Equation (4.11b) applies when  $\varphi_1$  and  $\varphi_2$  are exclusive. A disjunction is called exclusive if it has the following form:*

$$(\varphi \wedge \dots) \vee (\neg\varphi \wedge \dots)$$

◇

**Definition 4.15 CQQL negation (evaluation):** For the unary logical negation operator  $eval()$  follows a similar pattern as shown above.

$$eval(\neg\varphi, d_j^r) \rightsquigarrow 1 - eval(\varphi, d_j^r) \quad (4.12)$$

◇

**Definition 4.16 CQQL atomic condition (evaluation):** The atomic case evaluates to the actual value of the atomic condition, i.e.,

- for a Boolean or database condition, where 0 means false and 1 refers to true:

$$eval(\varphi, d_j^r) \rightarrow \{0, 1\} \quad (4.13)$$

- for a retrieval or proximity condition, where 0 indicates maximum dissimilarity and 1 maximum similarity (or POR respectively):

$$eval(\varphi, d_j^r) \rightarrow [0, 1] \quad (4.14)$$

◇

As a result, the recursive application of  $eval()$  on a CQQL query leads to a score value  $s \in [0, 1]$  for each  $d_j^r$  with respect to the specified query  $q$  [Zellhöfer & Schmitt 2010b]. Similar to other IR models (see Section 2.2.2), this score can be used to establish a total order (see Appendix A.1.2) of the documents in a collection to express a notion of relevance.

The usage of CQQL as a query language in MIR will be illustrated with an example in Section 4.4 after another feature of the language – the integration of weighted logical connectors – has been introduced.

### 4.3 Integration and Evaluation of Weights in CQQL

CQQL offers means to incorporate weights into a logical query as presented in the original work by Schmitt [2007]. Weights are one way to personalize the results of a query. The motivation to personalize query results in order to achieve a high level of user satisfaction is described in more detail and related to other personalization approaches separately in Section 5.1. The weighting mechanism is crucial for an understanding of the later sections of Chapter 5 that describe a relevance feedback approach that is central to this thesis.

The core idea of weighted CQQL queries is surprisingly simple: all logical connectors become equipped with weights in order to steer the influence of their operands on the result of the query evaluation. The next step is to transform this weighted query into a logical formula without weights. To achieve this, Schmitt [2007] suggests to transform the weighting variables  $\theta_i \in [0, 1]$  that are attached to the logical connectors into constants as shown below, while  $\theta_i = 0$  means that the operand has no impact on the result and  $\theta_i = 1$  means that the operand behaves as in the unweighted case.

### 4.3 Integration and Evaluation of Weights in CQQL

Definition 4.17 **CQQL weighted conjunction:**

$$\wedge_{\theta_1, \theta_2} (\varphi_1, \varphi_2) \rightsquigarrow (\varphi_1 \vee \neg \theta_1) \wedge (\varphi_2 \vee \neg \theta_2) \quad (4.15)$$

◇

Definition 4.18 **CQQL weighted disjunction:**

$$\vee_{\theta_1, \theta_2} (\varphi_1, \varphi_2) \rightsquigarrow (\varphi_1 \wedge \theta_1) \vee (\varphi_2 \wedge \theta_2) \quad (4.16)$$

◇

Table 4.2 gives the “truth” table for the weighted conjunction and disjunction when different weight settings are given.

Table 4.2: Impact of Weights in CQQL

$\varphi_1$	$\varphi_2$	$\theta_1$	$\theta_2$	$\varphi_1 \vee \neg \theta_1$	$\varphi_2 \vee \neg \theta_2$	$\varphi_1 \wedge_{\theta_1, \theta_2} \varphi_2$	$\varphi_1 \wedge \theta_1$	$\varphi_2 \wedge \theta_2$	$\varphi_1 \vee_{\theta_1, \theta_2} \varphi_2$
0.6	0.4	0.0	0.0	1.0	1.0	1.00	0.0	0.0	0.00
0.6	0.4	0.0	1.0	1.0	0.4	0.40	0.0	0.4	0.40
0.6	0.4	1.0	0.0	0.6	1.0	0.60	0.6	0.0	0.60
0.6	0.4	1.0	1.0	0.6	0.4	0.24	0.6	0.4	0.76

For an arbitrary  $n$ -ary logical connector, the idea is generalized. That is, the operands become associated with the respective weight constants. Thus, a ternary conjunction  $\wedge_{\theta_1, \theta_2, \theta_3}(a, b, c)$  is transformed into:

$$\wedge_{\theta_1, \theta_2, \theta_3}(a, b, c) \rightsquigarrow (a \vee \neg \theta_1) \wedge (b \vee \neg \theta_2) \wedge (c \vee \neg \theta_3) \quad (4.17a)$$

$$\wedge_{\theta_1, \theta_2, \theta_3}(a, b, c) \rightsquigarrow (a + \neg \theta_1 - a \cdot \neg \theta_1) \cdot (b + \neg \theta_2 - b \cdot \neg \theta_2) \cdot (c + \neg \theta_3 - c \cdot \neg \theta_3) \quad (4.17b)$$

The disjunction is treated analogously:

$$\vee_{\theta_1, \theta_2, \theta_3}(a, b, c) \rightsquigarrow (a \wedge \theta_1) \vee (b \wedge \theta_2) \vee (c \wedge \theta_3) \quad (4.18)$$

Equation (4.18) hints at the potential complexity of the resulting evaluation using the rules presented above. Because of the non-exclusive disjunction in the formula, its evaluation is bloated by the arithmetic sum  $(a + b - a * b)$ , which raises the computational cost of such an evaluation (see Equation (4.11a)).

Fortunately, the extension of CQQL with weights does not violate its Boolean algebra property (see Schmitt [2007] for the proof). Thus, all Boolean transformation rules hold and can be used to simplify complex, logical statements as the following example illustrates:

$$\vee_{\theta_1, \theta_2, \theta_3}(a, b, c) \rightsquigarrow (a \wedge \theta_1) \vee (b \wedge \theta_2) \vee (c \wedge \theta_3) \quad (4.19a)$$

$$\Leftrightarrow \overline{\overline{(a \wedge \theta_1) \vee (b \wedge \theta_2) \vee (c \wedge \theta_3)}} \quad (4.19b)$$

$$\Leftrightarrow \overline{\overline{(a \wedge \theta_1) \wedge (b \wedge \theta_2) \wedge (c \wedge \theta_3)}} \quad (4.19c)$$

$$\Leftrightarrow 1 - ((1 - (a \cdot \theta_1)) \cdot (1 - (b \cdot \theta_2)) \cdot (1 - (c \cdot \theta_3))) \quad (4.19d)$$

## 4 A Quantum Logic-based Model for Multimedia Information Retrieval

The (neutral) double negation added in (4.19b) serves as a preparatory step to apply De Morgan's law in (4.19c). This allows a simplification of the arithmetic evaluation presented in the last step.

Furthermore, a generalization of the principle that simplifies the handling of non-exclusive disjunctions in CQQL is possible:

$$\bigvee_{i=1}^{|\varphi|} \varphi_i \rightsquigarrow 1 - \prod_i^{|\varphi|} (1 - \varphi_i) \quad (4.20)$$

The compatibility with the laws of Boolean algebra of weighted CQQL is a crucial distinguishing feature from other weighting approaches such as the one by Fagin & Wimmers [2000], which violates the law of associativity [Schulz & Schmitt 2003]. Moreover, Schmitt [2007] presents other critical points such as the so-called stability problem of Fagin and Wimmers' approach that is also investigated by Sung & Hu [2009]. For more details about the theoretical properties of the CQQL and other weighting approaches, refer to Schmitt [2007]. For the sake of completeness, Zellhöfer & Schmitt [2010b] examine further properties of the weighting in CQQL that are only of marginal interest for this dissertation and therefore omitted.

To conclude, Schmitt et al. [2008] summarize the main characteristics of the weighting approach of CQQL that are important in the context of this thesis:

1. "A zero-weighted operand [ $\theta_i = 0$ ] has no impact on the result.
2. A one-weighted operand [ $\theta_i = 1$ ] behaves like an unweighted operand.
3. The weighting is realized by conjunction, disjunction, and negation. Thus, all boolean algebra laws remain valid. In contrast, most other weighting approaches are realized by arithmetic operations *outside* the logic violating laws.
4. The effect of weights to the evaluation result is linear<sup>37</sup>."

[Schmitt et al. 2008, p. 5]

### 4.4 CQQL as a Multimodal Logical Query Language

In Section 2.1, multimedia documents have been introduced as composite documents addressing different modalities. From a technical point of view these composite parts can be stored in different types of retrieval systems. For instance, the visual parts of a multimedia document can be stored in a MIR system, whereas its textual parts are held in an IR system. Additionally, metadata about the document, e.g., copyright information or the document's file size, is stored in a RDBMS.

As a consequence, different data access or retrieval paradigms have to be incorporated into the retrieval process of multimedia documents. This problem is not new and is also present in databases dealing with imprecise data [Weikum 2007]. Hence, a number of authors, e.g., Fuhr & Rölleke [1994]; Fuhr [2001]; Chaudhuri et al. [2005],

<sup>37</sup>That is, the "evaluation of the [weighted] CQQL conjunction and disjunction is based on linear formulas." [Schmitt 2007, p. 7]



#### 4.4 CQQL as a Multimodal Logical Query Language

or Schmitt [2008] (see Section 4.1) have addressed this issue by suggesting appropriate query languages combining the worlds of DB and IR.

As outlined before, CQQL provides means to combine different retrieval paradigms and weights to express the (subjective) importance of different conditions. This section describes how CQQL can be used in MIR with the help of an example. Later, the strengths and weaknesses of the query language are discussed.

*Imagine you are searching for photographs to illustrate a brochure about Renaissance paintings similar to a specified photograph of a painting whose intellectual property rights have expired, i.e., images that are in the public domain, or that are free from licensing costs.*

To facilitate the understanding, we assume that the visual similarity of a photograph can be sufficiently expressed by relying mainly on its colors and – to a lower degree – the types of the present edges in it. The visual similarity is calculated by using techniques from MIR (see Section 2.3) based on the automatically extracted low-level features of a provided QBE document (see Section 3.3), whereas the copyright information is stored in a RDBMS.

This sample information need can be expressed as a CQQL query  $q_\theta$  in extended<sup>38</sup> tuple relational calculus [Codd 1970] as follows<sup>39</sup>:

$$\{d^r \mid \text{collection}(d^r) \wedge (\exists q^r)(\text{query}(q^r) \wedge \underline{d^r.\text{color} \approx q^r.\text{color} \wedge_{\theta_1, \theta_2} d^r.\text{edges} \approx q^r.\text{edges} \wedge (d^r.\text{copyright} = \text{'public domain'} \vee d^r.\text{copyright} = \text{'free'})})\},$$

with *collection* being a relation containing all document representations,  $d^r$  a tuple variable (the document representation), and *query* a relation containing the query representation  $q^r$  describing the QBE document. In other words, we seek all  $d^r \in \text{collection}$  that fulfill the query consisting of a QBE and database part, i.e., all relevant documents with respect to the given IN represented by  $q_\theta$ . To keep things simple, we assume the weighting variables  $\theta_i$  as arbitrary constant values.

In order to evaluate this query, the weighted conjunction has to be transformed using the rules presented in Section 4.3:

$$\{d^r \mid \text{collection}(d^r) \wedge (\exists q^r)(\text{query}(q^r) \wedge \underline{((d^r.\text{color} \approx q^r.\text{color} \vee \neg\theta_1) \wedge (d^r.\text{edges} \approx q^r.\text{edges} \vee \neg\theta_2)) \wedge (d^r.\text{copyright} = \text{'public domain'} \vee d^r.\text{copyright} = \text{'free'})})\}$$

Because the required syntactical normalization (see above) is unnecessary in the presented case, the arithmetic evaluation follows. To facilitate reading, the query will be

<sup>38</sup>CQQL extends the tuple relational calculus by a binary similarity operator  $\approx$  on two attributes and weighted logical connectors as the weighted conjunction  $\wedge_{\theta_1, \theta_2}$  seen in the sample query.

<sup>39</sup>For clarity, the CQQL conditions that are subjects to the rules discussed in Section 4.2 and 4.3 are underlined.

#### 4 A Quantum Logic-based Model for Multimedia Information Retrieval

subdivided into parts that are evaluated separately.

$$\begin{aligned}
 & \{d^r \mid \text{collection}(d^r) \wedge (\exists q^r)(\text{query}(q^r) \wedge \\
 & \quad (\underbrace{(d^r.\text{color} \approx q^r.\text{color} \vee \neg\theta_1)}_{\text{col}}) \wedge (\underbrace{(d^r.\text{edges} \approx q^r.\text{edges} \vee \neg\theta_2)}_{\text{edg}})) \wedge \\
 & \quad \underbrace{\hspace{10em}}_{\text{retrieval\_part}} \\
 & \quad (\underbrace{(d^r.\text{copyright} = \text{'public domain'} \vee d^r.\text{copyright} = \text{'free'})}_{\text{pd}}))\} \\
 & \quad \underbrace{\hspace{10em}}_{\text{db\_part}} \\
 \\
 & \text{weighted\_color} \rightsquigarrow \text{col} + \neg\theta_1 - \text{col} \cdot \neg\theta_1 \\
 & \text{weighted\_edges} \rightsquigarrow \text{edg} + \neg\theta_2 - \text{edg} \cdot \neg\theta_2 \\
 & \text{retrieval\_part} \rightsquigarrow \text{weighted\_color} \cdot \text{weighted\_edges} \\
 & \text{db\_part} \rightsquigarrow \text{pd} + \text{free} - \text{pd} \cdot \text{free} \\
 & \Rightarrow \\
 & \text{eval}(q_\theta, d_j^r) = \text{retrieval\_part} \cdot \text{db\_part}
 \end{aligned}$$

This example can be extended arbitrarily, e.g., by using logical implication, more weighted logical connectors etc.<sup>40</sup>.

To calculate the result of  $\text{eval}(q_\theta, d_j^r)$ , all atomic conditions for  $d_j^r$  are evaluated. To give an example,  $\text{col}$  is replaced with the value of the similarity calculation between the QBE image  $q^r$  and  $d_j^r$  (see Section 2.3.2). This evaluation result is then used in the arithmetic evaluation presented above. The other conditions are treated accordingly with the difference that the condition on  $d^r.\text{copyright}$  yields a Boolean truth value.

The result of  $\text{eval}(q_\theta, d_j^r)$  is a score value for each tuple, i.e., each document in the collection, which is used to establish a total order amongst the retrieved documents (a ranking) to express a notion of relevance (see Section 4.2).

It is likely that the similarity calculation returns 0 (maximum dissimilarity) only for a very small amount of document representations because of the underlying structure of the low-level features and similarity/distance calculations (see Section 2.3). As a result,  $\text{eval}(q_\theta, d_j^r)$  only returns 0 for  $d_j^r$  with  $d_j^r.\text{copyright} \notin \{\text{'public domain'}, \text{'free'}\}$ . Consequently, if no Boolean conditions are present in a CQQL query, it is likely that most document representations in the collection obtain a score greater than 0 – even if it is very close to it. These documents are returned as “relevant” to the user, although it can be argued whether this is an appropriate description for such documents. One way to avoid this behavior is to utilize a threshold value  $\tau$  only returning

<sup>40</sup>Additional and more complex examples are available in the source code of the prototypical implementation accompanying this dissertation. They can be found in namespace `dbis.weightlearning.evalfunction` (see Appendix E).

$\{d_i^r \in collection \mid eval(q_\theta, d_i^r) \geq \tau\}$ . Unfortunately,  $\tau$  is dependent on the structure of  $q_\theta$ , e.g., on how the values for  $\theta_i$  are set and which logical connectors are used. Hence, a top- $k$  size limitation of the result ranking is far more viable when using CQQL.

#### 4.4.1 Advantages of Logic-based Multimodal Retrieval

Taking the participants of various ImageCLEF<sup>41</sup> tasks [Müller et al. 2010], important evaluation benchmarks in MIR (see Section 7.1), during the years 2010 and 2013 [Popescu et al. 2010; Tsikrika et al. 2011; Thomee & Popescu 2012; Zellhöfer 2012e; Caputo et al. 2013] as a representative sample of state of the art techniques in MIR, one must infer that the usage of logical query languages such as CQQL is rather uncommon. Typically, techniques from machine-based learning are used that try to find an optimal weighting for the combination of different high- or low-level features (see Section 2.3.1) relying on a linear matching function. Most of these approaches require some form of training data. Similar techniques have been presented at other major conferences of the field, such as the International Conference on Multimedia Retrieval (ICMR) [Ip & Rui 2012; Jain & Prabhakaran 2013].

Most machine-based learning approaches (MBL) – or, particularly in MIR, learning to rank approaches (see Section 5.2) – have three things in common:

1. structured queries are seldom used,
2. the choice of features and matching function is predefined, and
3. some kind of training data is needed.

Without doubt, these approaches have made a significant impact on the improvement of retrieval effectiveness of MIR systems. However, this improvement is achieved at the expense of flexibility. Because of the need for training data, MBL comprise the risk of being over-optimized to a usage domain of MIR – in practice the Cranfield-based test collections (see Section 7.1). Also, training data cannot be assumed to be available in every MIR usage scenario, e.g., during the search in a personal media document collection, which is unknown to the algorithm’s designer by nature.

Additionally, the type of query is frequently predetermined by the system. Hence, users cannot directly modify the query, even though there might be another query that reflects their IN in a better way<sup>42</sup>. Furthermore, although MBL are fully automatable (which is can also be desired by users), they have been shown to not always produce results that are comprehensible to users [Hearst 2009, Sec. 8.8ff.].

In contrast, logic has a long tradition of being used for reasoning going way back to Aristotle (and even further). It is used for reasoning in both mathematics and philosophy because it has been shown to be an effective means to deduce the validity of an argument (see Appendix A.1.4). Normally, the deduced judgment is understandable

<sup>41</sup>The CLEF Cross Language Image Retrieval Track, <http://www.imageclef.org/>

<sup>42</sup>Please note that the query can be learnt by some MBL approaches, but typically will not be subject to direct user modifications.

#### 4 A Quantum Logic-based Model for Multimedia Information Retrieval

for a human<sup>43</sup>. Besides being a sound and aesthetic mathematical and philosophical concept, the usage of logic in MIR has more advantages.

As it is inherent in logic, the usage of a logical query language supports the formulation of structured queries, i.e., logical connectors can be used to form a query for the current IN. As a result, logic in MIR does not only support the inclusion of positive QBE documents as used in the example above. With negation, the provision of negative QBE documents at query formulation time also becomes possible. That is, the user's subjective notion of relevance and irrelevance becomes part of the query. This separates the logical approach from other approaches such as the ones proposed by Kherfi et al. [2003]; Deselaers et al. [2005], or Liu et al. [2009] that can use negative example documents only during relevance feedback (RF) (see Section 5.1.1). In the opinion of the author, positive and negative examples for an IN have to be considered as peers during query formulation time because they help to represent the IN from a broad perspective. By exploiting this information, relevant results can already be retrieved after the first interaction with an MIR system and not after several RF iterations.

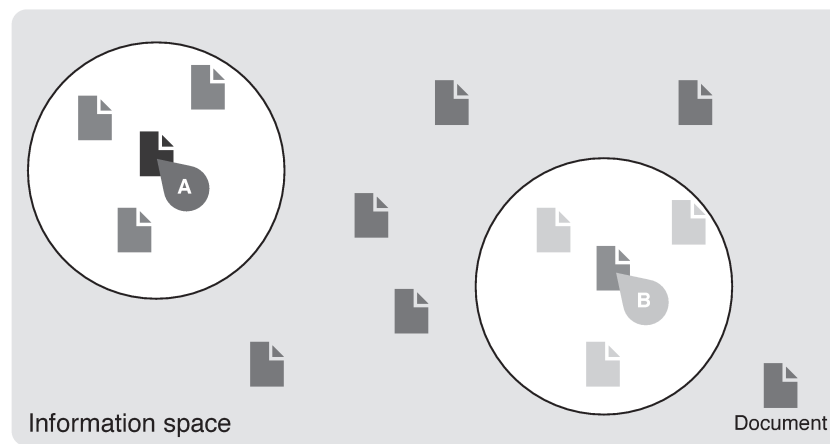


Figure 4.1: The cluster hypothesis versus multi-clustered information needs

Another advantage of the logic-based approach towards MIR is the possible combination of different aspects of the IN that manifest in different subspaces of the information space. Traditional (M)IR techniques assume that the *cluster hypothesis* holds:

**Definition 4.19 Cluster hypothesis:** “This hypothesis may be simply stated as follows: *closely associated documents tend to be relevant to the same requests.*” [van Rijsbergen 1979, Ch. 3]  $\diamond$

In other words, the hypothesis states “that relevant documents tend to be more similar to each other than to non-relevant documents” [Hearst & Pedersen 1996, p. 77].

<sup>43</sup>We assume a perfectly reasonable life-form such as the fictional character Mister Spock of the science fiction series *Star Trek*, who is free from emotions and subjectivity.

## 4.4 CQQL as a Multimodal Logical Query Language

Figure 4.1 illustrates the hypothesis for two disjoint IN  $A$  and  $B$  in a hypothetical information space. In this scenario, the documents in the circle around the given QBE document  $A$  form the cluster of relevant documents for the IN represented by  $A$ . In order to retrieve the documents surrounding  $B$  one has to provide  $B$  as the QBE document. The documents outside the circles are irrelevant to both sample IN.

As described in Section 3.1.1, the user-oriented viewpoint on IR acknowledges that the IN is dynamic. For instance, the *berry picking model* [Bates 1989] explains that users explore the information space to retrieve different documents that eventually satisfy their complex IN. Referring to Figure 4.1, this means that a user moves along a path from  $A$  to  $B$ , which does not need to be straight. With logic-based queries, it becomes possible to use disjunctions to retrieve documents from the cluster  $A$  and  $B$  at the same time<sup>44</sup>. This makes logic a good utility for modeling complex IN that cannot be satisfied by one cluster (or concept) of the information space alone.

Moreover, negation enables users to state queries that retrieve dissimilar documents. Following Belkin's ASK hypothesis (see Definition 2.4), it is not hard to imagine that a user recognizing an anomalous state of knowledge might have access to information resources describing the complement of the information in need. That is, a user can express what is known about the current problem domain but not what is searched. In this example, the user can provide two negated QBE documents  $A$  and  $B$  and retrieve their complement, i.e., the rest of the information space depicted by the documents outside the circles in Figure 4.1 for further inspection.

These examples make clear that logic (and thus CQQL) provides powerful means for users to express their IN. Furthermore, logic can also be used to build even more complex formulas allowing implications, which can become helpful to address different media types in one query. For instance, implications allow the use of distinct high- and low-level features dependent on the media type of the document.

The utilization of logic also brings technical benefits. As shown in Section 4.3, Boolean transformation rules can be applied to logical queries in order to translate them to a form that allows a faster execution by the MIR system comparable to the optimizers found in RDBMS. For instance, queries can be transformed in a way that the information space becomes restricted to contain less documents on which costly distance calculations have to be carried out (see Section 2.3.2).

To conclude, a MIR system supporting user-definable logical queries is (potentially) far more versatile than systems that have a fixed set of matching functions such as the "search by subject" feature available in Google's search or its equivalent in Bing (see Figure 4.2; highlighted regions). Please note that the predefined queries differ in both search engines.

Finally, the deductive nature of a logical query renders the usage of training data obsolete as it does not require training data of any form<sup>45</sup>. In theory, given a sufficiently well-defined query, all relevant documents are expected to be found as described by

---

<sup>44</sup>Please note that the clusters in Figure 4.1 are only circle-shaped to simplify the illustration. In fact, disjunctions in CQQL are associated with subspaces of the information space (see Section 4.2).

<sup>45</sup>That is, if no relevance feedback is used. The relation of MBL and RF is discussed in Section 5.2.

#### 4 A Quantum Logic-based Model for Multimedia Information Retrieval

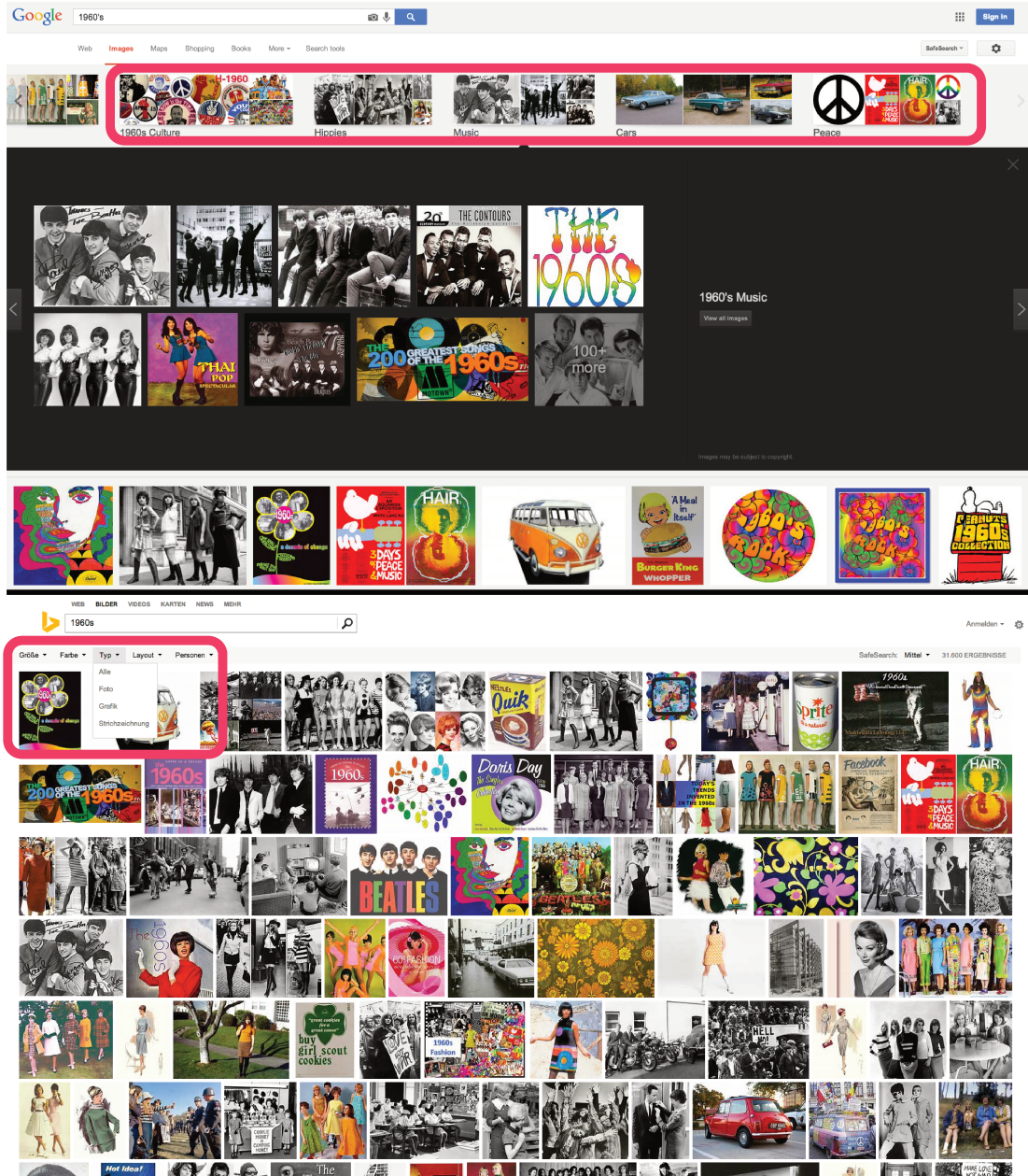


Figure 4.2: Predefined queries in commercial image search engines for the query term "1960s"; top: Google, bottom: Bing (as downloaded on October, 25th 2013)

## 4.4 CQQL as a Multimodal Logical Query Language

the uncertain inference [van Rijsbergen 1986b] (see Section 2.2.2).

Although the example at the beginning of this section does focus on visual MIR alone, the utilization of logical query languages such as CQQL is not limited to this domain. In fact, CQQL has also been used successfully in music retrieval to model musical genres within the “GlobalMusic2one” research project<sup>46</sup>. In this research project, CQQL was used to formulate and evaluate logical descriptions provided by musicologists characterizing musical genres on the basis of automatically extracted features such as a song’s instrumentation, rhythm, approximate length, or the geographic origin of the song etc. [Schiela 2010; Zellhöfer & Schmitt 2011b].

Comparable conclusions can be drawn for other professional user groups such as librarians that use Boolean queries in bibliographic systems or expert search options in Web IR systems such as Google. For instance, the utility of CQQL has been shown in an artwork auction platform<sup>47</sup> used by auctioneers and other professional art researchers [Buckow 2009].

### 4.4.2 Issues of Logic-based Multimodal Retrieval

The expressive power of logic comes at a price. The flexibility of logic is both its biggest advantage over other approaches and its greatest weakness: its flexibility increases the complexity of the query formulation. Although professional users can without doubt be trained to utilize the full expressive power of logic, untrained or layperson users are likely to become overstrained. For instance, there are numerous user studies describing problems of untrained users [Hearst 2009, Sec. 4.4]. Such problems as the retrieval of too many documents because of the wrong usage of logical connectors have already been discussed in Section 2.2.2. Furthermore, the semantics of Boolean logic is perceived counter-intuitive by many users. In natural language, “and” is used to widen a query’s scope, e.g., a query for “databases *and* information retrieval” might be expected to retrieve documents that contain one of each term. Instead, following the Boolean semantics, it will only retrieve documents that address both topics at once. That is, the query’s scope is narrowed. Similar examples can be easily constructed, e.g., “or” in natural language is often used in a mutually exclusive sense and not in the sense of a union of two choices [Hearst 2009, cf. Sec. 4.4]. If parentheses are allowed to group logical expressions, users are not guaranteed to understand the priority of parentheses (or logical connectors) in the correct, mathematical sense.

Another drawback of logic becomes particularly visible with MBL. Roughly speaking, the machine-based learning techniques used in MIR try to find a parameter set for a mostly linear function that describes the training data best. The evaluation of a logical query is not necessarily linear, e.g., its logical structure leads to a non-linear evaluation in case of CQQL (see Section 4.2). This restricts the set of feasible machine-based learning techniques. Furthermore, the logical structure can be interpreted as an additional

---

<sup>46</sup>[http://www.globalmusic2one.net/en\\_summary.html](http://www.globalmusic2one.net/en_summary.html)

<sup>47</sup><http://www.artnet.com/>

constraint for the learning algorithm because it defines the number of features to be used and how they are combined. From this point of view, a linear aggregation function such as the weighted arithmetic average of a number of features leaves much more place to optimize the parameter set. Hence, it is not surprising that logic is seldom used within the field of MBL.

Other common criticism of logic-based models (see Section 2.2.2) such as the missing ranking functionality does not apply to CQQL as shown in Section 4.2.

How the problem of query formulation can be overcome is presented in Section 5.5 of this thesis. In addition, Section 6.2 discusses how users can avoid a direct confrontation with CQQL.

### 4.5 The Relation of the Quantum Logic-based Approach to other IR Models

As mentioned in the beginning of this chapter, van Rijsbergen [2004] places the QM-based approach towards IR in a stress field defined by three corner points: vector spaces, probability theory, and logic. This section follows this structure in order to relate CQQL to these other IR models. The section concludes with a positioning of CQQL in the aforementioned stress field.

#### 4.5.1 Vector Space Model

The similarity to the vector space model (VSM) lies at hand. Section 4.1 described extensively how the theoretical base structures of CQQL are embedded into a vector space, namely a Hilbert space  $H$ . Moreover, the VSM is using vectors to represent document and compares them against vectors (being subspaces as well) representing a query.

To measure the similarity of a document and a query, similar techniques are used. The VSM utilizes the inner product (i.e., the cosine of the angle between query and document vector) whereas a quantum measurement is carried out by calculating the squared cosine of the minimal angle between state vector (document) and subspace (query).

Although both models are relatively close on an algebraic level, the quantum logic-based approach is theoretically more sound because it also addresses logic and probability theory.

#### 4.5.2 Probabilistic Models

A first relation of QM (and thus CQQL) to probability theory has been sketched via the quantum measurement (see Definition 4.4) whose “results can be regarded as probability values” [Schmitt 2008, p. 54]. Van Rijsbergen’s argues comparably by pointing out that “[...] the state-vector is a measure of the space, meaning that each subspace has a probability associated with it induced by the state-vector” [van Rijsbergen 2004, p. 11].



## 4.5 The Relation of the Quantum Logic-based Approach to other IR Models

The relation between QM and probability theory is also established from a more philosophical point of view. The *Copenhagen interpretation* of QM is one of the most widely used interpretations of QM [Burkard et al. 1991]. Roughly speaking, the Copenhagen interpretation argues that QM is not dealing with the measurement of an objective reality. Instead, it replaces exact measurements that are common in “traditional” mechanics with probabilities of measurements or relative frequencies of an observable. The Copenhagen interpretation acknowledges the indeterminism of quantum mechanical phenomena and does not necessarily assume their real existence. In contrast, the interpretation assumes that the objects of the mathematical formalism of QM are used to predict the relative frequency or probability of outcomes of measurements (which are assumed to be real)<sup>48</sup>.

On a more formal level, van Rijsbergen [2004] and Schmitt [2008] utilize *Gleason’s theorem* to link QM and probability theory.

**Definition 4.20 Gleason’s theorem:** *Following van Rijsbergen, Schmitt [2008] presents Gleason’s theorem [Gleason 1957] as follows. Let  $L(\mathbf{H})$  be a set of all subspaces (each corresponding to a projector  $\mathbf{p}$ ) of a Hilbert space  $\mathbf{H}$  with at least 3 dimensions. Then, every countably additive probability measure (see Appendix A.2) on  $L(\mathbf{H})$  has the following form:  $P(\mathbf{p}) = \langle \varphi | \mathbf{p} | \varphi \rangle$  for a state vector  $|\varphi\rangle \in \mathbf{H}$ .  $\diamond$*

In consequence, each query represented by a projector in CQQL yields a probability value (the POR of a document) due to the quantum measurement that is used for query processing (see Table 4.1).

Various evidence for the close relationship between QM and probability theory has been presented so far. From this perspective, it is not surprising that CQQL’s evaluation rules resemble Kolmogorov’s axioms for independent events (see Appendix A.2). For instance, the conjunction of events in probability theory  $P(X \cap Y) = P(X) \cdot P(Y)$  is identical to CQQL’s evaluation rule (see Definition 4.13). The same applies to the other logical operations [Zellhöfer & Schmitt 2011b].

### 4.5.3 Logic-based Models

Section 4.1 has shown that CQQL constitutes a Boolean algebra. Hence, its relation to Boolean logic-based models does not need to be discussed further. Thus, we focus here on CQQL’s relation to other logic-based approaches.

#### Fuzzy Logic

Considering the evaluation results of a CQQL query that are in the interval of 0 and 1, the usage of fuzzy logic seems appropriate because this IR model can deal with such values (see Section 2.2.2). In addition, fuzzy logic has been utilized successfully since the early days of DB and IR [Kerre et al. 1986; Galindo et al. 2005].

---

<sup>48</sup>A conflicting view on QM is the many-worlds interpretation that must not be mistaken with the possible world semantics that is common in probabilistic databases, e.g., see Fuhr & Rölleke [1994].

On closer observation, the values derived from a CQQL evaluation do not constitute membership values of sets (denoted as  $\mu(x)$ , see Definition 2.17) as required by fuzzy set theory [Zadeh 1965] and fuzzy logic [Zadeh 1988]. Instead, Definition 4.4 and Section 4.5.2 have shown that these values have to be interpreted as probability values.

Although Zadeh [2005] tries to establish fuzzy logic as a generalization of uncertainty in his late work, there is still much debate on this issue (see Section 2.2.2).

Besides this theoretical hindrance that sets CQQL apart from fuzzy logic, there are also more practical advantages of CQQL over fuzzy logic in the field of MIR, which are extensively discussed by Schmitt et al. [2008]. The authors' findings can be recapitulated as follows.

*First*, fuzzy set theory is working on membership values alone. That is, the underlying mathematical model does neither include the origin of such values (e.g., the evaluation of a DB or IR condition as addressed in Section 4.1) nor their semantics. Hence, the whole theory is operating on membership values alone, while CQQL addresses conditions in a much broader sense by working with projectors.

*Second*, some fuzzy set operations are subject to the so-called *dominance problem*. In the original publication of fuzzy set theory, Zadeh [1965] suggests the usage of the *min* function for the intersection of two sets, the *max* function for the set union, and  $1 - \mu(x)$  for the complement. We illustrate here the dominance problem with the *min* function and show that it does not meet user expectations [Lee et al. 1994]. The *min* function is returning one out of two values<sup>49</sup>, namely the minimum of both. In other words, one value becomes completely ignored during the function's evaluation, namely the bigger of the two. Additionally, the absolute difference between the two values is not taken into account. For instance,  $\min(0.1, 1.0)$  (case 1) returns 0.1 as well as  $\min(0.1, 0.11)$  (case 2). Intuitively, one would expect that the absolute difference of similarity in cases 1 and 2 would be expressed by the fuzzy intersection operator *min*. Instead, this information, which might be of interest to the user, is completely lost. As a result, one value dominates the other.

*Third*, not all fuzzy set operations follow the laws of Boolean algebra (see Appendix A.1.5). Although *min*/*max* are idempotent, they violate the axiom of complements (see Definition A.12), which can easily be shown:

$$x \wedge \neg x \rightsquigarrow \min(x, 1 - x) = \min(0.5, 0.5) = 0.5 \neq 0 \quad | x = 0.5$$

An alternative to *min*/*max* (amongst others<sup>50</sup>) is to express the intersection by the algebraic product ( $a \cdot b$ ) and the union by the algebraic sum ( $a + b - a \cdot b$ ). This function set overcomes the dominance problem by involving both values in the evaluation. It equals the formula used in CQQL (see Definition 4.13) but is not idempotent in case of fuzzy logic [Schmitt et al. 2008] because the fuzzy model does not take the properties

<sup>49</sup>For instance, similarity scores of two documents with respect to a query in case of MIR.

<sup>50</sup>For an overview of functions commonly used in fuzzy set theory, see Kruse et al. [1993]. For parameterized functions that mainly address the dominance problem by offering a parameter that steers their behavior between conjunction and disjunction, refer to Waller & Kraft [1979]; Yager [1988], or Lee [1994].

## 4.5 The Relation of the Quantum Logic-based Approach to other IR Models

of the projector lattice into account (see Section 4.1):

$$x \wedge x \rightsquigarrow x \cdot x = x^2 \neq x$$

### Probabilistic Relational Algebra and Probabilistic Datalog

The relation of CQQL to other logical IR models such as the *probabilistic relational algebra* (PRA) [Fuhr & Rölleke 1994] and its implementation *probabilistic datalog* (PD) [Fuhr 1995, 2000] can be easily established over their common motivation.

Similarly to the ideas of Schmitt [2008] (see Section 4.1), PRA aims at combining Boolean and vague predicates that incorporate a notion of uncertainty, e.g., the result of “some kind of similarity operation” [Fuhr & Rölleke 1994, p. 53] in MIR.

PRA relies on the principle of uncertain inference [van Rijsbergen 1986b] and the understanding that IR can be regarded as a generalization of DB. Consequently, PRA is developed as a “generalization of standard relational algebra” [Fuhr & Rölleke 1994, p. 32].

The core idea of PRA is to associate every tuple in the DB with a *probabilistic weight* expressing whether it belongs to a result relation (in PRA, each query results in a relation). In order to calculate the probabilistic weight  $pw$  of a tuple, the “probabilities of the underlying basic events, i.e., tuples of the base relations” [Fuhr & Rölleke 1994, p. 36], are used. In case of a Boolean base relation  $pw \in \{0, 1\}$ , while for a probabilistic relation  $pw \in [0, 1]$  holds. To put it simply, an atomic base relation can be understood as a relation with one attribute consisting of tuples, which can be evaluated in a Boolean fashion (an event that can only occur or not) or as a probability (an event with a likelihood of occurrence) with respect to a query. As a consequence, Boolean relations are interpreted by Fuhr & Rölleke [1994] as a special case of probability relations. These basic events are then combined using the operators of the probabilistic event algebra (see Appendix A.2.2).

Understanding tuples of probabilistic relations as (composed) probabilistic events means to take the dependence of the events on each other into account. That is, it becomes important for the calculation of the  $pw$  for a tuple to consider if its base events are disjoint, which simplifies the calculation of probabilities [Fuhr 2001]. In PRA, this information is modeled by an integrity constraint of a relation: the disjointness key. Moreover, Fuhr & Rölleke [1994] formulate a safety restriction in PRA, i.e., “a PRA expression is called safe if, for all possible values of its arguments, no event expression contains more than one event from a set of dependent events” [Fuhr & Rölleke 1994, p. 55]. From a practical point of view, the effects of this restriction are comparable to CQQL’s commutative condition restriction presented in Definition 4.10. In the opinion of the author, CQQL’s restriction affects the syntactical level of the query language and is therefore more user-friendly because it does not need a priori knowledge neither about the dependency of events nor about their disjointness.

Comparably to CQQL, it has been suggested to use PRA in MIR in form of its implementation *probabilistic Datalog* (PD) [Fuhr 2001]. PD [Fuhr 1995, 2000] is based on Data-

## 4 A Quantum Logic-based Model for Multimedia Information Retrieval

$\log^{51}$ , a declarative logic programming language for deductive databases, that extends the domain relational calculus with recursions and probabilities. A brief overview of PD can be found in Fuhr [2001, Sec. 6]. In this paper, Fuhr describes how PD can be used to combine DB and IR functionality in a MIR system. Fuhr's argumentative layout is similar to the one given in Section 4.4.

To recapitulate, Fuhr & Rölleke [1994] rely on classical probability theory, while CQQL uses the mathematical formalisms of quantum mechanics to investigate how conditions can be combined and how they can be evaluated as simple arithmetic expressions. Eventually, this leads to a discrimination of similarity-based IR and Boolean DB conditions on a syntactical level, which is not present in this form in PRA. Instead, to evaluate a query, Fuhr & Rölleke [1994] have to define if conditions are mutually exclusive.

Besides their different mathematical origin, another major discrimination feature of PRA from CQQL is PRA's missing support for weighted logical connectors. To sum up, both approaches are relational complete and extend the relational model by incorporating a concept of uncertainty. Although PRA's introduction is based on the relational algebra and CQQL's presentation on domain relation calculus, the resulting query languages have almost the same expressive power because relational algebra and domain relation calculus have been shown to be equal in terms of expressiveness [Codd 1972].

### 4.5.4 Quantum Mechanics-inspired Models

Section 4.1 gave an overview over the theoretical foundation of CQQL – namely quantum mechanics and logic and CQQL's motivation by van Rijsbergen's seminal work [van Rijsbergen 2004]. It is not surprising that van Rijsbergen's work has motivated other contributions to the stress field of IR and quantum mechanics. Because of the novelty of the QM/QL motivated IR models, a closing discussion of the models is not yet possible. Nevertheless, two exemplary approaches are presented briefly in this section of the dissertation.

#### QIR – Quantum-based Information Retrieval

The *quantum-based information retrieval* (QIR) model has evolved from contributions of van Rijsbergen's work group at the University of Glasgow [van Rijsbergen 2004; Piwowarski & Lalmas 2009; Piwowarski et al. 2010].

The QIR model is presented here in more detail than Melucci's Quantum Retrieval Model (see below) because it has also been used – like CQQL – to formalize the principle of polyrepresentation (see Section 4.7). Both CQQL and QIR are based on the same mathematical formalism derived from quantum mechanics, but are complementary in terms of the the mathematical interpretation of documents and queries (or INs) as Table 4.3 shows. An overview over the quantum mechanics-related terms has been given in Section 4.1.

The core assumption of the QIR model [Piwowarski et al. 2010] is the existence of a Hilbert space  $H$  – the so-called information need space. Additionally, each document

---

<sup>51</sup>For an introduction of Datalog, refer to Abiteboul et al. [1996].

## 4.5 The Relation of the Quantum Logic-based Approach to other IR Models

Table 4.3: Related concepts from QIR, CQQL, and quantum mechanics

QIR	Quantum Mechanics	CQQL
Query/IN	State vector	Document representation
Document representation	Projector/subspace	Query/IN
Matching	Quantum measurement	Matching

$d$  can be represented as a subspace in  $H$ , while the user's IN is represented as a state vector  $\varphi$ .

Moreover, the QIR model assumes that each document can give a relevant answer to various *pure* IN aspects [Piwowarski et al. 2010]. The authors give an example of the pure IN aspect "pop music" that is "represented by the terms 'music', 'chart' and 'hit' of the term space" [Piwowarski et al. 2010, p. 62]. Another important assumption is that a document can be split into semantic fragments, of which each addresses an IN aspect. Hence, a document can be seen as the sum of the pure IN aspects it is relevant to.

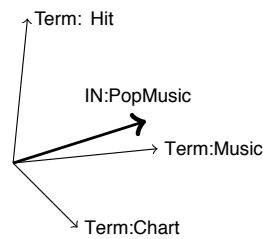


Figure 4.3: A pure IN in a IN/term space [Piwowarski et al. 2010, Fig. 1a]

On a more formal level, each document is modeled by a set of vectors  $\mathcal{V}_d$  representing the various semantic fragments. The determination of the components of each vector in  $\mathcal{V}_d$  is based on the  $tf * idf$  formula (see Section 2.2.2) but is not limited to this approach. Figure 4.3 gives an example for the pure IN aspect "pop music" in relation to its constituting terms. The full document is represented by a subspace of  $H$ ,  $\mathcal{S}_d$ , that is spanned by the vectors  $\mathcal{V}_d$ . One central property of  $\mathcal{S}_d$  is that it constitutes the smallest subspace in  $H$  that is a fully relevant answer to a pure IN aspect  $\varphi$  it contains. In other words, the projection of  $\varphi$  onto the subspace  $\mathcal{S}_d$  yields 1. Obviously, this corresponds to the quantum measurement introduced in Definition 4.4, and the result can therefore be interpreted as a probability value: the probability of relevance of a document to the given IN aspect.

In order to combine multiple IN aspects  $\varphi_i$  that can be present in a user-stated multi-term query, Piwowarski et al. [2010] discuss different approaches but recommend to employ the tensor product (see Definition 4.5), which is also used in CQQL to compose different attributes of a document representation. To recapitulate, a query is represented as the tensor product of the IN aspects it contains. Further examples for other combination techniques are available in a Master's thesis [Döcke 2012] that has been

## 4 A Quantum Logic-based Model for Multimedia Information Retrieval

supervised by the author of this dissertation.

In contrast to CQQL, QIR does not support structured queries based on quantum logic. In QIR, queries are not based on projectors and therefore cannot be composed with the help of different lattice operators. Instead, queries or INs are represented with state vectors that can be combined using the tensor product which does not take a user-specified logical structure into account.

On the other hand, QIR regards the IN as the dynamic part in the search process, projected onto the fixed subspaces spanned by the documents in the IN space  $H$ . This interpretation renders the QIR approach diametrically opposed to CQQL, although both approaches are based on the same origin and make therefore use of the relationship between probability theory and geometry elaborated throughout Section 4.5. Because of their different interpretation of documents and queries, there is no standard way to translate QIR into CQQL or vice versa [Zellhöfer et al. 2011].

From a conceptual point of view, the interpretation of the IN as the dynamic part in the IN space is compelling because it aligns well with the observation of the dynamic nature of the IN, which has also been put forward by Ingwersen & Järvelin [2005] or other researchers in IIR (see Section 3).

A more thorough discussion on formal differences between CQQL and QIR is available in a joint publication by both workgroups [Zellhöfer et al. 2011].

### Melucci's Quantum Retrieval Model

Another quantum mechanics-based IR model has been proposed by Melucci [2008]. For the sake of brevity, we refer to this retrieval model as *Melucci's quantum retrieval model* (MQRM). Melucci's approach is motivated by the need to include the context of a searcher, e.g., the prior knowledge, the motivation, or the search environment, into the vector space model (see Section 2.2.2).

In contrast to QIR, MQRM interchanges the role of the IN and the document. As a consequence, the representation of documents and INs resemble the model proposed by CQQL as shown in Table 4.4.

Table 4.4: Related concepts from MQRM, CQQL, and quantum mechanics

MQRM	Quantum Mechanics	CQQL
Document representation	State vector	Document representation
Query/IN	Projector/subspace	Query/IN
Matching	Quantum measurement	Matching

Roughly speaking, Melucci's core idea is to represent queries and contexts as subspaces in the Hilbert space  $H$ . These subspaces (or their corresponding projectors) can be linearly combined to form arbitrarily complex queries. Documents, which are represented as state vectors, are projected onto a (composed) subspace in order to calculate their probability of relevance to the given contextual query. As a consequence, a document can give different PORs to the same query depending on the current context as the original query subspace is altered with the help of the context subspaces.

## 4.5 The Relation of the Quantum Logic-based Approach to other IR Models

Similarly to CQQL, the POR of a document is expressed as the geometric proximity to the subspace representing the (contextual) query (see Definition 4.4). What is most interesting in MQRM is the fact that both POR and context are modeled in a unified mathematical model – namely quantum mechanics [Melucci 2008, cf. p. 14:3].

Although an in-depth discussion of the formal properties of MORM and its impact on the concept of relevance in IR falls out of the scope of this dissertation, interested readers may refer to Melucci [2008] and Melucci [2011] for further details.

### 4.5.5 Classification of CQQL

The relation of QM to the VSM, logics, and probability theory on a mathematical basis has been shown sufficiently by van Rijsbergen [2004]. To extend the view on CQQL, the last sections compared CQQL against typical IR models. In this section, we investigate where in the stress field of these three points is CQQL located best.

Given the advantages of using logical connectors in a query language for MIR (see Section 4.4), it is not surprising that CQQL has a very strong background in Boolean logic. For instance, this feature is required to be used with the relational model of DB. Unlike the traditionally used relational domain calculus, CQQL supports more than Boolean conditions in order to deal with probabilities or similarities. Hence, it extends (or generalizes) the relational domain calculus.

Consequently, CQQL can be seen as a probabilistic logic [Nilsson 1986, 1994]. That is, a Boolean algebra operating on probability values instead of Boolean truth values. In other words, Boolean logical connectors can be used to connect probabilities. This argumentation does not contradict with CQQL's support of both Boolean or probability conditions. In fact, a Boolean attribute condition is a specialization of a probability condition that can only become 0 or 1 while a probability value is out of the interval  $[0, 1]$ . CQQL shares this characteristic with PRA and PD.

Not surprisingly, CQQL's relation to the uncertain inference principle [van Rijsbergen 1986b] (see Section 2.2.2) as suggested in Section 4.4 can easily be shown by relying on the geometric interpretation of QM and quantum logic.

Table 4.1 on page 50 outlines the central relations between the concepts of DB and QM, pointing out that a CQQL query has to be understood as a projector. From Definition 4.3, we know that every projector  $p$  is bijectively associated with a vector subspace of the Hilbert space  $H$ . Furthermore, a quantum measurement (query processing) resembles the projection of a state vector (a tuple or document representation) onto the subspace associated with  $p$  (see Definition 4.4). To conclude, Section 4.5.2 presented various arguments for the interpretation of a quantum measurement's outcome as probability value.

From this point of view, the probability that one can infer from a document representation to a query representation  $P(d_i^r \rightarrow q^r)$  (the uncertain inference) is represented in QM by the quantum measurement (see Definition 4.4), i.e., the probability that  $p$  (the query) is measured with a system  $|\varphi\rangle$  (the document).

Hence, we know that  $P(d_i^r \rightarrow q^r) = 1$  when the state vector  $|\varphi\rangle$  representing  $d_i^r$  is embedded in the subspace  $vs(q^r)$  associated with the projector  $p_q$  representing  $q^r$ .

#### 4 A Quantum Logic-based Model for Multimedia Information Retrieval

As all state vectors  $|\varphi_i\rangle$  are embedded in  $H$ , we know that we can construct a projector  $p_d$  and thus a subspace in which the projection of one given  $|\varphi\rangle$  yields 1. According to van Rijsbergen [1986b], the uncertain inference is computed as follows:

$$P(d_i^r \rightarrow q^r) = P(q^r | d_i^r) = \frac{P(d_i^r \cap q^r)}{P(d_i^r)}$$

Understanding the document as a “self-reflexive” query expressed by the projector  $p_d$  and the original query as the projector  $p_q$ , the uncertain inference can be interpreted geometrically as follows:

$$P(d_i^r \rightarrow q^r) \rightsquigarrow \frac{\langle \varphi | (p_d \sqcap p_q) | \varphi \rangle}{\langle \varphi | p_q | \varphi \rangle}$$

That is, the probability that the “physical property”  $p_d \wedge p_q$  is measured with  $|\varphi\rangle$  (the document). Given that the document representation equals the query representation,  $p_d \sqsubseteq p_q$  holds. Then, we can simplify the formula and employ the law of idempotency of the meet operator (see Definition 4.6) that holds with CQQL (see Proof 4.1).

$$P(d_i^r \rightarrow q^r) \rightsquigarrow \frac{\langle \varphi | (p_d \sqcap p_q) | \varphi \rangle}{\langle \varphi | p_q | \varphi \rangle} = \frac{\langle \varphi | p_q | \varphi \rangle}{\langle \varphi | p_q | \varphi \rangle}$$

From Definition 4.3, it is known that:

$$p_q \equiv |\varphi\rangle\langle\varphi|.$$

Hence, the following holds because  $\varphi$  is a unit vector and if  $p_d \sqsubseteq p_q$ :

$$P(d_i^r \rightarrow q^r) \rightsquigarrow \frac{\langle \varphi | (p_d \sqcap p_q) | \varphi \rangle}{\langle \varphi | p_q | \varphi \rangle} = \frac{\langle \varphi | p_q | \varphi \rangle}{\langle \varphi | p_q | \varphi \rangle} = \frac{\langle \varphi | \varphi \rangle \langle \varphi | \varphi \rangle}{\langle \varphi | \varphi \rangle \langle \varphi | \varphi \rangle} = \frac{1}{1} = 1$$

This shows that CQQL follows the uncertain inference principle.

Although the argumentation for CQQL being a probabilistic logic following the principle of uncertain inference is straightforward, its positioning as a quantum logic is not that clear. This problematic situation arises from the fact that CQQL is operating on commuting projectors only (see Definition 4.9). This restriction becomes necessary to guarantee that CQQL follows the laws of Boolean algebra. In fact, quantum logic is no Boolean algebra, because it violates the law of distributivity [Schmitt 2008]. Thus, CQQL is only operating with a subset of quantum logic – namely the part in which distributivity holds. As a result, it seems more appropriate to classify CQQL as a *quantum logic-based* query language because Schmitt [2008] relies on the mathematical foundations of QM and quantum logic rather than calling CQQL a quantum logic. The relation of quantum logic (i.e., a non-classical, non-distributive logic) to the uncertain inference principle is extensively discussed in van Rijsbergen [2004, Ch. 5] and therefore will not be covered here.



To conclude, this does not mean that Schmitt’s excursion into QM is useless or purely ornamental. Instead, it provides an interesting and theoretically sound view on the integration of DB and IR while not neglecting practical implementation opportunities. In the end, the restriction of CQQL to follow the laws of Boolean algebra enables the language to be used in common RDBMS, e.g., in the form of custom SQL dialects [Lehrack & Schmitt 2010].

### 4.6 CQQL as an Implementation of the Principle of Polyrepresentation

The discussion in Section 4.4 already showed that a multimedia document is characterized by multiple representations within a MIR system and suggested that a query language in MIR should deal with these different representations, stored and processed by different retrieval systems. An example also showed the application of CQQL as a query language in MIR.

The principle of polyrepresentation (PoP) [Ingwersen 1996] argues comparably to the intuitive understanding that a multimedia document manifests in various representations of different cognitive or functional origin. Furthermore, the PoP assumes that the more representations of different cognitive and functional origins are pointing to a set of documents inside the so-called cognitive overlap (CO), the higher the probability of relevance of these documents is (see Section 3.2).

Being a cognitively motivated principle, the PoP has lacked a formalization in (M)IR for a long time. In fact, there are only a few studies supporting the principle in IR, e.g., by Skov et al. [2004] or Larsen et al. [2009] (see Section 3.2.1 for more details). As reported in Section 4.7, several formalizations of the PoP in IR were suggested around 2010.

Following Larsen’s proposed research direction for implementations of the PoP, i.e., “to identify flexible and effective matching methods that can generate high quality cognitive overlaps from a variety of the most promising representations” [Larsen et al. 2006, p. 89f.], this section outlines how CQQL implements the PoP. Its evaluation in terms of retrieval effectiveness is covered separately in Chapter 8.

The idea to utilize CQQL to implement the PoP in the information space of MIR was first suggested by Zellhöfer & Schmitt [2010c, 2011b]. Subsequently, the approach is extended in later works, shifting its focus towards a *polyrepresentative user interaction model* [Zellhöfer & Schmitt 2011a; Zellhöfer 2012a].

To determine whether the representations’ specificities of a document are relevant for the current CO modeled by a CQQL query, it is important to keep the cognitive processes in mind that lead to the representation.

For instance, reconsider the example from Section 4.4:

$$\{d^r \mid \text{collection}(d^r) \wedge (\exists q^r)(\text{query}(q^r) \wedge \underline{d^r.\text{color} \approx q^r.\text{color} \wedge_{\theta_1, \theta_2} d^r.\text{edges} \approx q^r.\text{edges} \wedge (d^r.\text{copyright} = \text{'public domain'} \vee d^r.\text{copyright} = \text{'free'})})\}$$

#### 4 A Quantum Logic-based Model for Multimedia Information Retrieval

In this CO, we assume that the representation *copyright* is of Boolean nature, i.e., the actor (e.g., the licensing manager of the document collection) that created this representation of a document decided to use a classification scheme. This classification scheme is restricted to attribute values such as “public domain”, “free”, and others that we cannot conclude from the example. Hence, it would not be appropriate to allow some sort of similarity measure to determine how similar a document’s corresponding representation is to the concept “public domain” because this would distort the intention of the representation’s creator as well as the user’s IN expressed by the CO.

Another cognitively different representation created by the MIR system is *color*. Here, we allow a similarity measure to express the IN for similar documents to a given image based on its visual content. According to the PoP, the MIR system that extracts this representation from the images in the collection is also seen as an actor taking part in the retrieval process. It can be argued whether the remaining representation, i.e., *edges*, is cognitively or functional different from *color*. The decision depends mainly on whether one is willing to see the MIR system as a black box providing different representations with different functional objectives or if one wants to see every low-level feature as the result of different cognitive processes of different actors, e.g., the programmers of the feature extractors. However, it is not needed to pursue this argument for the remainder of this thesis as long as it becomes clear that CQQL addresses these different representations with respect to their retrieval paradigm or the cognitive concept behind them.

To recapitulate, *each condition in CQQL corresponds to the probability of relevance of the document’s respective representation with respect to a given query*. Table 4.5 recapitulates the presented findings.

Table 4.5: Related concepts from CQQL and the principle of polyrepresentation

CQQL	PoP
Database attribute	Representation
Query	Cognitive overlap
Query/condition evaluation result	Probability of relevance

Furthermore, Zellhöfer & Schmitt [2011a] postulate two requirements for an IR model implementing the PoP:

1. “Structured queries to model the CO should be possible.
2. The probability of membership to a PCO [see below] of a document has to be computable on basis of its representations.”

[Zellhöfer & Schmitt 2011a, p. 2]

These requirements are directly derived from study results of the PoP presented in Section 3.2.1. The fact that CQQL meets both requirements has been extensively covered in the previous sections of this chapter.

The original concept of the CO is Boolean, i.e., an intersection of different representations is formed [Larsen et al. 2006]. Obviously, this can only be reflected with CQQL restricted to Boolean conditions alone because general CQQL constitutes a probabilistic logic. In other words, the usage of CQQL does not mean an exact matching of a CO

## 4.6 CQQL as an Implementation of the Principle of Polyrepresentation

with a collection. Instead, CQQL follows the best match paradigm whenever more than Boolean conditions are involved. To emphasize this difference, Zellhöfer & Schmitt [2010c] suggest to extend the CO concept to a “*penetrable cognitive overlap*” (PCO). This step is necessary for the development of a polyrepresentative user interaction model (see below). A PCO is characterized by a border that can be penetrated into the direction of each involved representation or even reformed. For instance, the weighting mechanism of CQQL might strengthen the influence of certain representations onto the PCO. Additionally, a PCO consisting originally of four representations may develop into a PCO of three representations during the user interaction (see below). Figure 4.4 illustrates the concept.

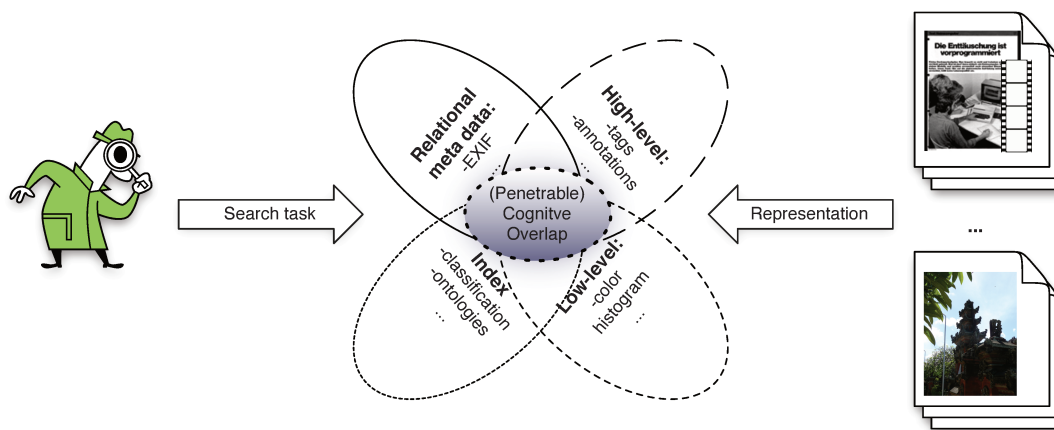


Figure 4.4: Venn-like diagram of different document representations forming a penetrable cognitive overlap [Zellhöfer & Schmitt 2010c, Fig. 1]

One aspect that is often neglected during the modeling of a CO are the subjective user preferences between the different representations. Following Ingwersen [1996], one would assume that all representations contribute equally to the CO. This approach has to be criticized, particularly from a MIR point of view, because users react differently to textures, colors, or other representations of a multimedia document. Although the general IN might be addressed by a CO, we assume that its personalization can only be achieved by steering the influence of the involved representations onto it. In extreme cases, this can also mean that certain representations are neglected in the CO according to the needs of the user; alternatively, the user will leave the original CO focussing on only a subset of the representations – hence the name *penetrable*<sup>52</sup> cognitive overlap. The concept of a PCO and varying important representations is expressed in CQQL with the help of weighting variables. This feature is crucial for the conceptual interaction model used in this dissertation, which is discussed in more detail in Section 6.2.1. The same section also explains how relevance feedback, directed, and explorative information

<sup>52</sup>To a certain extent, “penetrable” in this sense is synonym to “personalizable”.

seeking strategies can be interpreted within the PCO model.

Another side effect of CQQL's probabilistic evaluation is that it can only calculate the POR of a document with respect to a given CO (modeled by a CQQL query). This is not contrary to the PoP. In fact, Ingwersen & Järvelin [2005] refer indirectly to a probabilistic interpretation by emphasizing the insecurity of the relevance judgments carried out by the IR system [Zellhöfer & Schmitt 2011a]. Furthermore, Ingwersen links the principle to the uncertain inference [van Rijsbergen 1986a, b] (see Section 2.2.2) and subsequent work [van Rijsbergen & Lalmas 1996] that he regards as viable model to implement polyrepresentation in the information space [Ingwersen 1996, cf. p. 35ff.].

Finally, CQQL can be positioned in the polyrepresentation continuum suggested by Larsen et al. [2006] as shown in Figure 4.5. The illustration extends Figure 3.2 on page 38 with an axis for the query form. Being a best match fusion system, CQQL has to be placed at the unstructured end of the "retrieval technique" axis suggested by Larsen. This aligns CQQL with virtually all other retrieval models and systems in MIR. As argued before, CQQL's foundation in logic sets it apart from other approaches in MIR that rely mostly on unstructured queries (see Section 3.2.1).

As said before, common MIR techniques also fuse different features and use some form of weighting to combine these features. However, the used techniques are often limited to a certain domain and lack a sound theoretical basis [Kokar et al. 2004; Fuhr 2012]. If one examines the ImageCLEF participants' approaches used in MIR over the last years [Popescu et al. 2010; Tsikrika et al. 2011; Thomee & Popescu 2012; Zellhöfer 2012e; Caputo et al. 2013], a main trend becomes obvious: structure in queries is seldom (almost never) used and the utilization of weighted linear functions fusing features is the current state of the art. In particular, weights are used to increase the impact of textual representations onto the matching process in order to raise its retrieval effectiveness. This finding is not surprising given the system-centric perspective on IR taken by ImageCLEF (see Chapter 7) that mainly encourages the usage of machine-based learning approaches. More user-centered or holistic approaches such as the PoP are usually neglected by such initiatives. As a result, common MIR techniques are clearly positioned in the lower right quarter of the polyrepresentation continuum.

### 4.7 The Relation of the CQQL Approach to Other Formalizations of Polyrepresentation

Around the 2010s, several suggestions were made to formalize the principle of polyrepresentation and to link it with more "traditional" IR models.

An early workshop contribution by Frommholz & van Rijsbergen [2009] uses the geometric QIR framework presented in Section 4.5.4 to implement the principle. The framework is extended in a subsequent publication [Frommholz et al. 2010] focussing mainly on the formal IR model and how it can be used to mirror polyrepresentation with the help of the the described IN aspects and multiple Hilbert spaces for each document representation. In contrast, CQQL uses different state vectors in one Hilbert space to model polyrepresentation. To determine the cognitive overlap of multiple represen-

#### 4.7 The Relation of the CQQL Approach to Other Formalizations of Polyrepresentation

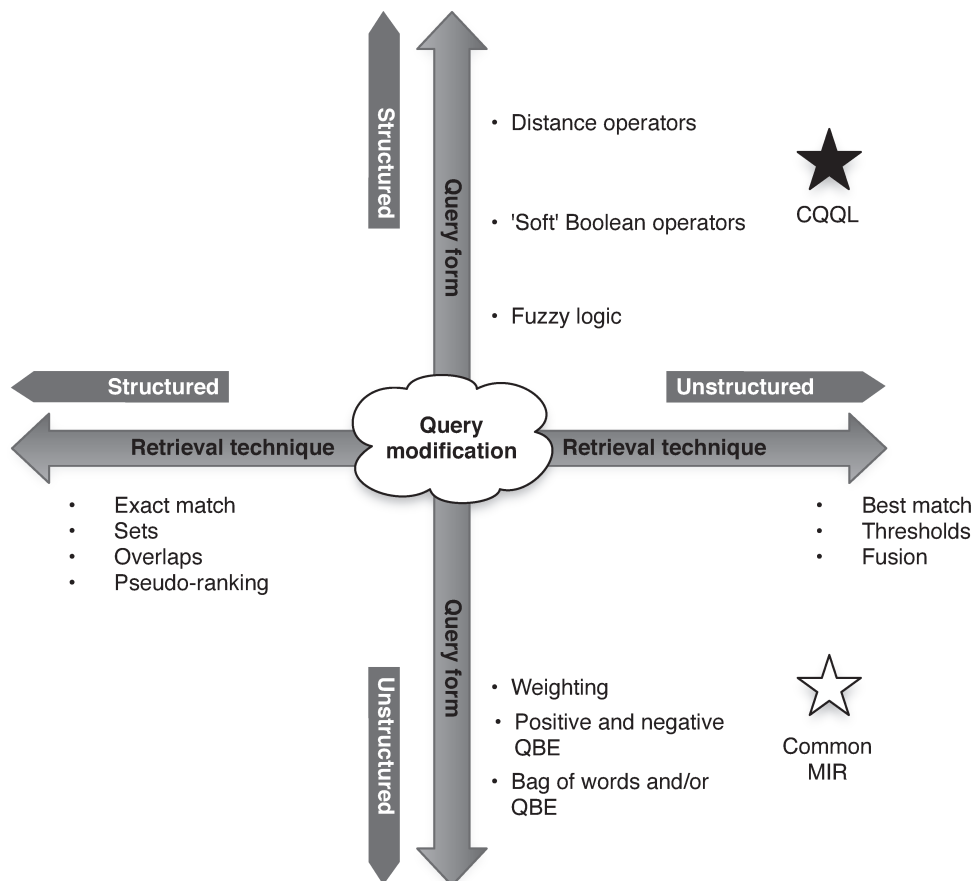


Figure 4.5: The extended polyrepresentation continuum [Larsen et al. 2006, cf. Fig. 6] including samples adapted to MIR; the black star indicates CQQL, the white star indicates common MIR techniques

tations, Frommholz et al. [2010] suggest to combine the single representation spaces using the tensor product in order to eventually establish a polyrepresentation space. At the reported state, the QIR approach can only be applied to textual IR because it heavily relies on properties of index terms and structural characteristics of documents [Zellhöfer 2012a, cf. p. 63] (see Section 4.5.4). An extension of QIR to the domain of MIR has not been published yet and can therefore not be discussed.

An alternative formalization of the principle is suggested by Lioma et al. [2010] and enhanced by an experimental study [Lioma et al. 2012]. In their contributions, the authors create analogies between subjective logic [Jøsang 2001] and the PoP. Roughly speaking, subjective logic is a probabilistic logic dealing with subjective beliefs about the truth value of a proposition. These beliefs are then combined to model the CO in order to assess the relevance of a document.

## 4 A Quantum Logic-based Model for Multimedia Information Retrieval

Taking a more interactive viewpoint, White et al. [2005] present a user interface for an IR system based on the PoP. In addition, White [2006] shows that implicit relevance feedback (see Section 5.1.2) based on polyrepresentation can improve the retrieval performance of an IR system.

Based on the information seeking strategies investigated by Belkin [1993], Beckers [2009] suggests a linkage between polyrepresentation and ISSs. The author adds “aspects” to the PoP that serve to group representations. Unfortunately, it remains unclear whether this approach really differs from the original discrimination into functionally and cognitively different representations [Zellhöfer 2012a]. The question on how a system can be implemented based on the presented concept is left open by Beckers [2009].

Coming from the field of databases, in 2010, Zellhöfer & Schmitt [2011b] suggested independently from van Rijsbergen’s Glasgow group to use CQQL (and hence to apply the work of van Rijsbergen [2004] comparably to Frommholz et al. [2010]) in order to formalize the principle of polyrepresentation in MIR. The subsequent development has been covered extensively in Section 4.6.

Because of their common theoretical origin in van Rijsbergen’s discussion on quantum mechanics in IR, it is not surprising that the PoP formalization in QIR is similar to the one with the help of CQQL. Their major differences at the formal level have been discussed in Section 4.5.4.

The most prevalent differences between QIR and CQQL are the support of MIR in the CQQL-based retrieval model and its holistic view on the interactive retrieval process that is elaborated in Chapters 5 and 6. Furthermore, the QIR-based formalization only focusses on the principle of polyrepresentation in the information space (see Section 3.2), whereas this dissertation also addresses the user’s cognitive space (see Section 6.2.1).

Concerning the support of highly structured queries, the polyrepresentative extension to QIR only supports the conjunction of multiple representations or *total cognitive overlaps*, as they are called by Frommholz et al. [2010, cf. Sec. 5.1]. Admittedly, the QIR model also supports weights to a certain extent. As weights in QIR are interpreted as probabilities that a user is interested in a particular representation, they have to sum up to 1 – a restriction that is not present in CQQL’s logic-embedded weights.

Regarding the emphasis on the logical properties of the retrieval model behind CQQL, it resembles Lioma’s approach to a higher degree than QIR. Albeit this similarity is not present at the level of their mathematical foundations, i.e., quantum logic versus subjective logic, both approaches share the same motivation – to understand IR as a problem that can be best solved using probabilistic logic as suggested by the uncertain inference (see Definition 2.19).

## 5 Machine-based Learning of Personalized CQQL Queries

Seit ich des Suchens müde ward, Erlernte ich das Finden.

---

*Friedrich Nietzsche, Die fröhliche Wissenschaft; 1882*

In this dissertation, *personalization* is generally understood as the tailoring of search results to an individual's information need (IN). The necessity of personalization implies that an initial query result may not adequately satisfy the user's IN.

Personalization has a long tradition particularly in IR. There are many approaches to personalize results ranging from relevance feedback (RF) to the usage of user profiles. Often, these personalization techniques are separated into approaches that rely on the users' explicit provision of information about their current IN or ones that infer the user's current IN implicitly [Hearst 2009, cf. Ch. 9]. The most prominent representative of explicit personalization is RF, while the usage of the user's search history or the data mining of query logs are typical implicit techniques.

Implicit techniques are often used in commercial Web IR engines that incorporate the user's current location and profile, or that re-rank results based on data mining of query logs or click-through protocols. Their explicit counterpart would be recommendations, e.g., often viewed documents or term suggestions that the user can choose from.

Personalization is also closely related to *query reformulation*, i.e., the loop of stating an initial query, the examination of its results, and the adjustment of the query in order to resubmit it. In IR, query reformulation can, for instance, be supported either by term suggestions that expand an initial query or RF.

For the sake of clarity, we subsume personalization and RF under the term *personalization* because RF in CQQL does not change the logical structure of a CQQL query as discussed in Section 5.4. Consequently, the term *query reformulation* is only used if the logical structure of a query is changed.

Similarly to Chapter 4, this chapter examines personalization techniques from different angles. First, Section 5.1 discusses personalization from an MIR point of view. Section 5.2 links MIR to DB personalization research via *learning to rank* approaches. Finally, Section 5.3 presents *preference-based* approaches for personalization, which are used mainly in the field of DB, in more detail because of their strong relation to CQQL's preference approach.

Section 5.4 and the following section discuss an explicit preference-based personalization approach relying on CQQL and machine-based learning that is inspired by the aforementioned fields of research. The subsequent Section 5.5 sketches how the CQQL-based personalization approach can be extended to learn new queries.

Because of the explicit nature of the presented personalization approach, implicit personalization techniques fall out of the focus of this thesis and are therefore only presented briefly.

### 5.1 Personalization in Information Retrieval

As said before, the long tradition of personalization-related research in (M)IR does not allow an in-depth discussion on all typical approaches. Instead, this section focuses on techniques that have a direct impact on the personalization technique presented in this dissertation. Overviews over the field are available in the aforementioned (M)IR textbooks, e.g., by Baeza-Yates & Ribeiro-Neto [2011].

One personalization technique of special interest in the context of this work is relevance feedback that has been briefly introduced in 2.2.1. Another means of personalization is the explicit weighting of certain query parts, e.g., as supported by the extended Boolean retrieval model discussed in Section 2.2.2.

#### 5.1.1 Explicit Relevance Feedback

From a historical perspective, explicit relevance feedback is one of the oldest personalization techniques in IR. The first formalization of explicit, basic RF is commonly attributed to Rocchio [1971] as a personalization technique in the vector space model (VSM). The approach has been continuously extended, e.g., by Ide [1971], in order to include negative RF. For the sake of brevity, the formalization of RF in the VSM is omitted here and its principle is outlined instead. For the mathematical formulas, please refer to the cited original publications or common IR textbooks (see Section 2.2).

Basic RF assumes that the cluster hypothesis (see Definition 4.19) holds. Moreover, RF in general assumes that a user's IN is fixed [Salton & Buckley 1990]. The idea behind RF is straightforward: given a vector  $q^r$  representing the query, the goal is to move  $q^r$  towards the vectors of relevant document representations  $d_i^r$  and away from the irrelevant document representations  $d_i^{-r}$ . Following the cluster hypothesis, this moves the altered  $q_{rf}^r$  towards the cluster of relevant document representations improving its retrieval effectiveness. The selection of relevant documents (and the associated document representations), i.e., the *positive* relevance feedback, is typically carried out by the user. Using the RF approach by Ide [1971], users can also give *negative* relevance feedback by selecting irrelevant documents.

RF has been shown many times to be an effective means of personalization that improves the retrieval effectiveness of an IR system. For instance, Harman [1992] discusses its utility in IR, while Huiskes & Lew [2008a] shows the benefits of RF in CBIR. Comparable results are reported in MIR [Feng et al. 2003]. The principle of RF can also be transferred to the fusion problem of different representations, where it can be used to select query term weights in IR [Ruthven et al. 2002].

Being an explicit personalization technique, the utility of RF heavily depends on the users' willingness to provide relevance (or irrelevance) feedback to the system. Hence,



it is important that the RF mechanism offers a good usability. The importance of the usability of the RF mechanism is also stressed by Ruthven & Lalmas [2003], who also provide a very detailed and comprehensive survey on RF in the field of IR.

In conclusion, the general findings of the utility of RF in IR can be transferred to MIR. In fact, RF is still a state of the art technique in MIR and CBIR. For instance, Assfalg et al. [2000a] uses positive and negative RF in CBIR. Further successful utilizations of RF in CBIR are presented by a number of authors [Kherfi et al. 2003; Deselaers et al. 2005; Liu et al. 2009]. The personalization technique is also used successfully in major MIR benchmarks such as ImageCLEF [Müller et al. 2010].

### 5.1.2 Implicit Relevance Feedback

Although implicit personalization is not examined in detail in this dissertation, implicit RF shall be mentioned for the sake of completeness. The major shortcoming of explicit RF is, if you will, its dependency on user input. In any case, RF can also be carried out by altering the initial query with data inferred from user logs or other sources [Croft et al. 2009, cf. Sec. 7.6.1]. Another source for implicit RF could be the time a user spends examining different documents that are then considered relevant by the IR system. Comprehensive overviews of implicit RF techniques are available by Kelly & Teevan [2003] and White [2004b].

A particularly interesting implicit RF technique is called *pseudo relevance feedback*. The core of the idea of pseudo RF is to interpret the first  $k$  retrieved documents as relevant and to modify  $q^r$  accordingly. This relieves users from explicitly stating RF and can improve the retrieval effectiveness of the initial retrieval step, which can be modified repeatedly during the following explicit RF iterations. Pseudo RF was originally suggested in 1979 by Croft & Harper [1988]. The main problem of pseudo RF is *query drifting*: a decline of retrieval effectiveness because the initial retrieval step includes only few or no relevant documents in the top- $k$  results. As a result,  $q^r$  is moved into the direction of irrelevant document representations.

### 5.1.3 Weighting of Query Parts

Another explicit personalization technique that has also been used from the early days of IR is the *weighting of query parts*. As described in Section 2.2.2, the extended Boolean model allows users to weight their query terms in order to express their importance at query formulation time (QFT). This separates the approach from explicit RF, which is carried out after the formulation of an initial query. Variants of query weighting can be used to express the importance of the co-occurrence of query terms in the text, e.g., to search for compound terms such as “swimming pool”. Similar means for the formulation are offered by most modern IR engines, e.g., by the Indri IR engine [Strohman et al. 2004] of the Lemur project<sup>53</sup>.

The weighting of query parts or representations at QFT is also important in MIR. For instance, the ImageCLEF Wikipedia benchmarks [Popescu et al. 2010; Tsikrika et al.

<sup>53</sup><http://www.lemurproject.org/>

2011] show that the retrieval effectiveness of textual query parts is much higher than the one of CBIR-based parts. Hence, the different weighting of query parts has become the de-facto standard in MIR because it has been shown to be a viable means to improve the retrieval effectiveness of MIR systems [Müller et al. 2010].

In principle, the idea to weight query parts differently is also implicitly present in early CBIR systems such as QBIC [Flickner et al. 1995] (see Section 2.3). Although the most famous demonstration system of QBIC, the CBIR implementation for the Hermitage museum (St. Petersburg, Russia)<sup>54</sup>, does not support the combination and weighting of different representation such as color or texture, the commercial incarnation of QBIC – the so-called Image Extender for IBM’s DB2 RDBMS – allows the combination of different low-level features by utilizing an SQL dialect.

### 5.2 Learning to Rank

*Learning to rank* (L2R) approaches – as typical MIR techniques taken from the field of machine-based learning – have been briefly described in Section 4.4.1. L2R approaches learn (or infer) a *ranking model*, i.e. a characteristic ranking function or parameter set for a pre-defined function that describes a desired rank best, from given labeled training data. In IR, these labels typically contain the relevance (or level of relevance) of documents with respect to a given IN. This ranking function can then be used to place new, unknown objects at the correct position in the desired target rank. In IR, this rank is usually a total or partial order of documents ordered by their relevance.

Most common L2R approaches fall into the large group of supervised learning approaches. L2R is still an active research area in IR. A comprehensive overview is available by Liu [2011]. For further information about machine-based learning techniques, refer to Russell et al. [2007] or comparable text books as this text cannot provide an adequate discussion on this scientific field.

As stated before, L2R relies on training data in order to learn a ranking model. This training data has to be obtained in large amounts from a credible source to infer a widely usable ranking model. Nowadays, this source is formed by human assessors who manually rate document-query pairs in order to provide the needed annotated training data. Thus, it becomes increasingly expensive and difficult to obtain annotations because of the growing amount of data. Moreover, training data is not available for all usage scenarios in IR, e.g., in the retrieval of classified data or personal search scenarios.

Another weakness of training data is that it cannot sufficiently address dynamic information seeking processes described in Chapter 3. Instead, the annotations are fixed to a number of pre-defined document-query relevance pairs.

On the other hand, L2R approaches *per se* do not require user interactions after the training phase to improve the retrieval effectiveness in contrast to explicit RF. Nevertheless, L2R techniques can also be used in scenarios relying on explicit user input of

---

<sup>54</sup><http://www.hermitagemuseum.org/fcgi-bin/db2www/qbicSearch.mac/qbic?selLang=English;> as tested on 27th November 2014.

training data. These approaches link L2R to RF and preference approaches that are discussed in Section 5.3.

At first glance, this emphasizes the similarity between L2R and the aforementioned RF approaches. In any case, the conceptual difference between L2R and RF lies in the modification of the query representation  $q^r$  and the used matching function (see Definition 2.15). Whereas RF actually modifies  $q^r$  and hence the IN stated by the user without altering the matching function of the IR system, L2R typically leaves the stated IN intact and infers a ranking model from it. This ranking model is eventually used as a matching function. For instance, if a user provides two documents describing the initial IN, an explicit RF technique would alter the initial  $q^r$  according to the additionally specified relevant documents throughout the RF process. Using explicit L2R, the same documents would be used to learn a ranking model that describes the current IN best.

From a user's point of view, this difference is hardly discriminable. This is particularly true if implicit L2R has been used to find a weighting scheme for the impact of different document representations, which is used for the initial query, e.g., as suggested by Faria et al. [2010]. Subsequently, this initial query can be modified by means of L2R (e.g., see Hu et al. [2008]) or RF using comparable user interactions.

In practice, implicit L2R techniques are heavily used in MIR (see Section 4.4.1). Although their utilization is often justified with their superior retrieval effectiveness and their abandonment of explicit user interaction, this is only partly true in the opinion of the author of this dissertation. In fact, the burden of user interaction is shifted towards the assessors' side. Although the derived assessments can be re-used through the means of machine-based learning, users are still needed to provide the labeled training data. Furthermore, the Cranfield-based evaluation paradigm (see Section 7.1) leverages L2R approaches because its required ground truth (or subsets of it) can be easily used to train L2R systems. As a consequence, the success of L2R in MIR cannot be separated from the predominance of system-centric Cranfield-based evaluation in the field. The impact of this system-centric viewpoint on IIR evaluation and development is further elaborated in Chapters 7 and 8.

Besides this criticism at the level of methodology, there are also potential problems with L2R approaches during interactive information seeking. Although a direct comparison is complicated because many factors determine the runtime behavior of an implementation (e.g., the efficiency of the implementation, the used computer resources, the underlying IR model etc.), L2R approaches are often more computationally intensive than RF or logic-based approaches. For instance, consider the two winning groups of the ImageCLEF 2013 Personal Photo Retrieval subtask [Zellhöfer 2013, cf. Tab. 4]. The first three places are taken by L2R approaches by Mizuochi et al. [2013], while places 4 to 6 are taken by a CQQL-based approach [Böttcher et al. 2013] that is very similar to the one described in this dissertation.

Table 5.1 compares the best performing experimental runs of both groups, i.e., *ISI\_1* by Mizuochi et al. [2013] and *DBIS\_run3* by Böttcher et al. [2013]. In terms of retrieval effectiveness (for a definition of the used metrics, see Definitions 8.4 and 8.8), the antagonists are relatively close to each other – the fourth placed *DBIS\_run3* reaches roughly 90% of the retrieval performance of the first placed run in terms of the nDCG metrics

## 5 Machine-based Learning of Personalized CQQL Queries

emphasizing the retrieval effectiveness of first placed documents (see Table 5.1, set in italics). As elaborated in Section 8.1.2, the retrieval performance of an IR system at early positions in the retrieved rank is particularly important because it relieves users from inspecting long lists of potentially relevant documents. This improves the user’s efficiency while interacting with the system and thus the user satisfaction (see Section 6.1.4).

Another important factor that affects the user satisfaction is the reactivity of an interactive system. Both discussed groups use prototypical systems without any support of an index. Hence, a general improvement of the runtime per topic, i.e., the time a system takes initially to retrieve all results for one query, can be expected. At any rate, the winning group’s system has much higher hardware demands (e.g., 24 CPU cores vs. 8 cores), uses less features (10 vs. 18), but still takes 10-15 times longer for the initial retrieval of its results<sup>55</sup>. Although one can expect performance improvements for comparable L2R approaches due to the utilization of computational parallelization and the usage GPGPU<sup>56</sup> techniques in the near future, the runtime factor is currently limiting the utility of such approaches in IIR scenarios.

Table 5.1: Comparison of runtime and retrieval effectiveness of the two winning groups at the ImageCLEF 2013 Personal Photo Retrieval subtask

System Characteristic	ISI_1	DBIS_run3
Type of MIR system	L2R, RF-supported	Logic-based, RF-supported
Number of features	10 (4 visual, 6 Exif)	18 (15 visual, 3 Exif)
CPU cores and model	24 cores, 4 x Intel Xeon X5675 3.07 GHz	8 cores, 2 x Intel Xeon E5520 2.26 GHz
CPU launch date	Q1'2011	Q1'2009
Runtime per topic	ca. 10-15 min.	ca. 1 min.
Metric	ISI_1	DBIS_run3
MAP, cut-off @ 100	0.5028	0.3954 (78,64%)
<i>nDCG, cut-off @ 20</i>	<i>0.7425</i>	<i>0.6798 (91,56%)</i>
<i>nDCG, cut-off @ 30</i>	<i>0.7288</i>	<i>0.6546 (89,82%)</i>
nDCG, cut-off @ 100	0.6878	0.6084 (88,46%)

Percentages are given with respect to the first placed ISI\_1 run.

### 5.3 Preference Models for the Personalization of Queries

Preferences are an important concept in psychology and, in particular, microeconomics [Lancaster 1991], amongst other scientific fields. In fact, the examination of consumer preferences forms the *utility theory* [Fishburn 1968], one of the core areas of microeco-

<sup>55</sup>The actual runtime was reported to the author of this text by Mizuochi et al. [2013] in a personal e-mail from October 19th 2013: “[...] for training, we took from 10 to 15 minutes for one topic. We used matlab in 24 core ubuntu 12.04 server. CPU is "Intel(R) Xeon(R) X5675 3.07GHz". The runtime of the second group can be derived from Table 8.8 but is known to the author because of his contributions to Böttcher et al. [2013].

<sup>56</sup>General purpose graphics processing unit, i.e., the usage of GPU to run massively parallel computation tasks.

### 5.3 Preference Models for the Personalization of Queries

nomics. Besides their importance in utility theory, preferences are also central in game and decision theory [Fishburn 1970].

Although the term *preference* is intuitively comprehensible and somewhat ubiquitous in most fields dealing with human decision processes [Becker 2008], this section introduces some important core concepts. A brief but comprehensible discipline-embracing overview of preferences is available by Hansson & Grüne-Yanoff [2012], on which the following terminology and concepts are based.

Put simply, a preference is an actor's choice between two alternative entities  $x_1$  and  $x_2$  from a set  $X$  of alternatives. Formally, an actor's preference between  $x_1$  and  $x_2$  is expressed as  $x_1 \succ x_2$ , i.e.,  $x_1$  is better than  $x_2$ . For now, the reasons why the actor prefers  $A$  over  $B$  are neglected. Instead, we continue with the specification of preferences.

**Definition 5.1 Strict preference:** *Given two entities  $x_1$  and  $x_2$ ,  $x_1 \succ x_2$  means that  $x_1$  is better than  $x_2$  (or  $x_2$  is worse than  $x_1$ ) to an actor, e.g., because  $x_1$  has more value<sup>57</sup> for the actor.* ◇

**Definition 5.2 Indifferent preference:** *The concept of indifference is closely related to preferences. Given two entities  $x_1$  and  $x_2$ ,  $x_1 \sim x_2$  indicates that both entities have the same value to the actor, i.e., the actor has no preference for one of the choices.* ◇

**Definition 5.3 Weak preference:** *A weak preference expresses that entity  $x_1$  is better than or equal in value to  $x_2$ . A weak preference between two entities  $x_1$  and  $x_2$  is denoted by  $x_1 \succeq x_2$ .* ◇

Analogously,  $x_1 \prec x_2$  means that  $x_2$  is better than  $x_1$  etc.. Furthermore, preferences have five central properties:

**Definition 5.4 Anti-symmetry (Preference):**  $x_1 \succ x_2 \wedge x_1 \neq x_2 \rightarrow \neg(x_1 \prec x_2)$  ◇

**Definition 5.5 Symmetry of indifference (Preference):**  $x_1 \sim x_2 \rightarrow x_2 \sim x_1$  ◇

**Definition 5.6 Reflexivity of indifference (Preference):**  $x_1 \sim x_1$  ◇

**Definition 5.7 Incompatibility of preference & indifference:**

$$x_1 \succ x_2 \rightarrow \neg(x_1 \sim x_2)$$

◇

Following these definitions, a *weak preference* can be defined accordingly:

$$x_1 \succeq x_2 \Leftrightarrow x_1 \succ x_2 \vee x_1 \sim x_2$$

**Definition 5.8 Transitivity (Preference):**

$$x_1 \succ x_2 \wedge x_2 \succ x_3 \rightarrow x_1 \succ x_3$$

$$x_1 \succeq x_2 \wedge x_2 \succeq x_3 \rightarrow x_1 \succeq x_3$$

$$x_1 \sim x_2 \wedge x_2 \sim x_3 \rightarrow x_1 \sim x_3$$

<sup>57</sup>Value does not have to be equal to monetary value. It can also be the utility of an entity, e.g., a coat might have a higher utility to a person trapped in a cold cave than an electric heating, although the same person might prefer a heating over a coat while sitting in an apartment.

## 5 Machine-based Learning of Personalized CQQL Queries

Although we assume the transitivity of preferences for the remainder of this text, it is noteworthy that it might not hold in every subjective case. Following transitivity,  $A > B \wedge B > C \rightarrow A > C$  holds. In any case, when a user directly compares  $A$  and  $C$  might  $A \not> C$  hold because in this comparison the user's preference is based on other criteria. This problem cannot be solved at the formal level and is related to the problem of incommensurability addressed in Section 5.4.1.  $\diamond$

From Definition 5.4 follows that a strict preference is also *irreflexive*:  $\neg(x_1 \succ x_1)$ , while a weak preference is *reflexive*:  $x_1 \succeq x_1$ .

Preferences as a means for query formulation in the field of DB appeared for the first time in the late 1980s [Motro 1986; Lacroix & Lavency 1987] and gained more attention in the last decade in general computer science, e.g., in order to determine the run order of processes, to implement economic models, or in DB research [Conitzer 2010]. In the context of this dissertation, we focus on preferences in DB that have been addressed by a number of authors, e.g., by Agrawal & Wimmers [2000]; Börzsönyi et al. [2001]; Govindarajan et al. [2001]; Chomicki [2002]; Kießling [2002]; Kossmann et al. [2002], and many more in the past decade.

Although the focus of this dissertation lies clearly in the research areas of DB and IR, preferences in DB cannot be discussed without acknowledging the impact of artificial intelligence (AI) on the field. Noteworthy overviews of preferences in AI are available by Brafman & Domshlak [2009] or Fürnkranz & Hüllermeier [2010].

Over the years, preference approaches in DB have evolved into two separate research branches: *qualitative* approaches (see Section 5.3.1) and *quantitative* ones (see Section 5.3.2). Because of the huge amount of publications from both schools of thought, this dissertation can only discuss some representative contributions and relate them to their microeconomic origins. Finally, qualitative and quantitative preference approaches are compared to each other in Section 5.3.3.

Roughly speaking, the main motivation to use preferences in DB is the relatively new finding (at least in this scientific field) that DB users might also be subject to query formulation problems, e.g., because of insufficient knowledge of the database and resulting incorrect or inappropriate queries. To be more precise, DB preferences allow the modification of *rigid* DB queries in a sense of weakening “the initially required characteristics if there is no object satisfying them, or to strengthen them if there are too many answers” [Lacroix & Lavency 1987, p. 217]. Similar arguments are provided by authors of both the qualitative and quantitative research field. For instance, Chomicki sees an important usage area of preferences in “information filtering and extraction [in order] to reduce the volume of data presented to the user” [Chomicki 2003, p. 427]. In order to illustrate a sample usage scenario of preferences in DB, consider the following example.

**Sample restaurant preference** For instance, one might query for vegetarian restaurants initially but will also accept Indian restaurants because of their wide range of vegetarian dishes if no completely vegetarian option is available.

## 5.3 Preference Models for the Personalization of Queries

From an IR point of view, the relatively late confrontation with query formulation problems or subjective preferences of users might be surprising as this problem has been in the research focus of IR since its early days. Nevertheless, DB research is driven almost solely by the directed search paradigm (see Section 3.3) and the expectation that a query always returns the correct answer defined by the relational model. Furthermore, DB researchers often assume that only expert users are interacting with their systems, e.g., with the help of SQL (see Section 3.3.1). Hence, the dawn of preferences in DB research has also led to an emphasis of user needs over the past decade in this research area.

Consequently, the users' needs and their interaction with the database management system (DBMS) via preferences have also moved into the focus of DB researchers, e.g., Agrawal & Wimmers [2000] or Chomicki who understands "preference querying as a *dynamic, iterative process*" [Chomicki 2007, p. 81]<sup>58</sup>. As a side effect, the continuous interaction with a DBMS during information seeking (see Chapter 3) requires the confrontation with another aspect of utility theory: the aggregation of preferences of one or more actors.

In economic theory [Becker 2008], multiple agents have different preferences. Given a set of alternative entities, each actor might place them in a different rank according to their subjective preferences. In order to reach a common goal, they decide to cooperate. As some actors have different preferences which cannot simply be combined in the sense of a union, they need a way to aggregate their preferences to reveal the most preferred entities, i.e., a *voting rule* [Conitzer 2010, cf. pp. 85f.]. One way to pick the winning entity would be the use of the *plurality rule*, i.e., to choose the entity that has been placed first by most actors. Alternative voting rules are the *anti-plurality rule*, i.e., the winning alternative is the one that has been chosen last by the least actors. Other voting rules might try to find a balance between the interests – or preferences – of the actors (see [Conitzer 2010] for a presentation of different approaches).

During interactive information seeking, each search step (e.g., a RF iteration) can be interpreted as an actor with different preferences that have to be aggregated. For instance, preferences that might be valid at query formulation time are augmented, but they might also become invalid because the IN has changed.

### 5.3.1 Qualitative Preferences

Fundamentally, qualitative preferences correspond to the intuitive understanding of preferences (see the beginning of Section 5.3). These preferences express that a particular entity is preferred to another one but do not give information about the "degree" of this preference. Review the query for vegetarian restaurants on page 86. The actor in this example prefers vegetarian restaurants over Indian restaurants and Indian restaurants over all other types because of their choice of vegetarian dishes – a strict preference that can be sketched as follows:

$$\text{vegetarian restaurant} \succ \text{Indian restaurant} \succ \text{other restaurant}$$

---

<sup>58</sup>The original accentuation has been maintained.

## 5 Machine-based Learning of Personalized CQQL Queries

This preference expresses a *qualitative* statement about the available restaurant alternative (i.e., “better than”) but does not state that the actor prefers vegetarian restaurants 2 times to Indian ones and Indian restaurants 1.25 times to other types of restaurants. Preferences that are capable of expressing such statements are commonly called *quantitative* (see Section 5.3.2).

Typically, qualitative preferences in the field of relational DB (see Section 2.2.2) are stated at the database attribute level and not at the level of concrete tuples from a database. This allows the formulation of preferences that can be used to deduce matching tuples from a preference-augmented query. Reconsidering the given example, one can imagine a sample relation for all available restaurants: *Restaurant*(*Type*, *Name*). The relation consists of two attributes, *Type* and *Name*, on which a possible preference can be defined. In the present case, the preference can be defined (informally) for the attribute *Type*, i.e., *Restaurant.Type* = “vegetarian”  $\succ$  *Restaurant.Type* = “Indian”.

Obviously, this approach does not require from the user to have an extensive database content knowledge at the time of query formulation. Moreover, the definition of preferences at attribute level is also supported by economic research showing that actors desire entities because of their properties (or attributes) and not because of themselves [Lancaster 1966].

In an early contribution to the utilization of qualitative preferences in DB, Lacroix & Lavency [1987] suggest a preference mechanism extending the domain relational calculus and SQL. Although the authors focus on a certain domain (the retrieval of software components), their general idea is to overcome “the difficulty of expressing in traditional query languages desirable characteristics of what has to be retrieved” [Lacroix & Lavency 1987, p. 217]. In an example [Lacroix & Lavency 1987, cf. Query Q4], the authors illustrate their approach and their SQL dialect by searching for coded versions of the software component *main* that have been developed for 16 bit computers if such components are available. Otherwise, *main* components that have been coded for other computer architectures are accepted (see Listing 5.1).

Listing 5.1: Sample preference query in the Lacroix-Lavency SQL dialect

```
1 SELECT THE versions OF main
2   HAVING status = coded
3   FROM WHICH PREFER THOSE HAVING target=16
```

Regarding the evaluation of such a preference query, Lacroix & Lavency [1987] suggest to first evaluate the query without the preference clause at line 3. Then, the preference clause is applied to the result set of the query in the sense of a *filter*. As a consequence, such a query might return an empty result set or – if there are tuples satisfying the preference – a result set that contains less tuples than the one without the preference clause [Lacroix & Lavency 1987, cf. pp. 219f.]. In addition, the approach supports a basic prioritization of preferences on a syntactical level by allowing the nesting of preference clauses. Hence, preferences can form hierarchies. Unfortunately, the paper lacks a detailed formal definition of the preference approach; therefore, one can only suppose that preferences in the work of Lacroix & Lavency [1987] are *strict*, i.e., they



### 5.3 Preference Models for the Personalization of Queries

form a strict partial order (see Appendix A.1.1). Furthermore, it is reasonable to assume that their preference model does not leave the relational model because the authors present it as “an extension of a language of the Domain Relational Calculus family” [Lacroix & Lavency 1987, p. 217].

To recapitulate, Lacroix and Lavency’s approach features four central characteristics of qualitative preference models in DB:

1. The approach supports only *strict preferences*,
2. Preferences can be *prioritized*, i.e., they can be ordered by their relative importance,
3. Preferences are seen as a *filter* on a result set, and
4. The approach is *compatible with the relational model*, e.g., the set semantics of results is maintained etc..

To continue the discussion on more recent qualitative approaches, it is necessary to introduce some concepts from the field of economics.

**Definition 5.9 Pareto optimality:** *Pareto optimality (or efficiency)<sup>59</sup> is the state of a system in which its resources are allocated most efficiently. Pareto optimality is reached when the resources are distributed in a way that one party’s situation cannot be improved further without deteriorating another party’s situation. The concept of Pareto optimality is central to economic theories that deal with the efficiency of production or the fair distribution of resources.*

*In formal terms, the Pareto optimum is a  $\mathbf{n}$ -tuple  $\mathbf{t}_1 = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$  out of a set  $\mathbf{T}$  of  $\mathbf{n}$ -tuples which is at least as good in its attributes  $\mathbf{n}_i$  as all other elements in  $\mathbf{T}$  and better in one attribute  $\mathbf{n}_p$ .*

*Let  $[\mathbf{i}]$  be an accessor to an attribute of a  $\mathbf{n}$ -tuple and  $\mathbf{i} \in \{1, 2, \dots, \mathbf{n}\}$ . Then, there is no other  $\mathbf{n}$ -tuple  $\mathbf{t}_2 = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) \in \mathbf{T}$  with  $\mathbf{t}_2[\mathbf{i}] \geq \mathbf{t}_1[\mathbf{i}]$  and at least one  $\mathbf{i}$  for which  $\mathbf{t}_2[\mathbf{i}] > \mathbf{t}_1[\mathbf{i}]$  holds.  $\diamond$*

**Definition 5.10 Ceteris paribus:** *In the field of qualitative preferences, the concept of Pareto optimality often coincides with the Latin term ceteris paribus (CP). It translates as “with other things the same” in the sense that all other attributes than the specified one remain equal or constant during a statement, experiment, or observation.  $\diamond$*

Motivated by the formal shortcomings of Lacroix & Lavency [1987] and a qualitative preference implementation in Datalog [Govindarajan et al. 2001], Chomicki [2002] suggests a qualitative preference mechanism that extends the relational algebra and is formally sound. After a critical comparison of qualitative and quantitative preferences (see Section 5.3.3), Chomicki presents a logical framework for the formulation of preferences and their composition. The framework supports only strict preferences, which are formulated in so-called *preference formulas*.

**Definition 5.11 Preference formula:** *A preference formula  $\mathbf{C}(\mathbf{t}_1, \mathbf{t}_2)$  is a first-order logical formula that defines a preference relation  $\succ_{\mathbf{C}}$  between two tuples  $\mathbf{t}_1$  and  $\mathbf{t}_2$ :*

$$t_1 \succ_{\mathbf{C}} t_2 \quad \text{iff} \quad \mathbf{C}(t_1, t_2).$$

<sup>59</sup>Named after the economist Vilfredo Pareto who investigated economic efficiency.

## 5 Machine-based Learning of Personalized CQQL Queries

A preference formula that only uses built-in predicates such as  $=$  and  $\neq$  in the case of a comparison to uninterpreted constants or arithmetic comparison operators such as  $>$ ,  $\leq$ ,  $=$ , etc. when comparing numerical values is called an intrinsic preference formula (IPF). An IPF that only uses arithmetic comparisons is called arithmetical [Chomicki 2002, cf. pp. 36ff.].  $\diamond$

The transformation of a preference into a preference formula can be easily shown with the help of an example. Consider the sample relation instance in Table 5.2 and the given preference “prefer cheaper restaurants of the same type”.

The preference  $(Type, Name, Avg.Price) \succ_C (Type', Name', Avg.Price')$  is equivalent to the preference formula  $Type = Type' \wedge Avg.Price < Avg.Price'$ .

More complex transformation examples can be found in [Chomicki 2002, 2003].

Table 5.2: Sample instance  $r_1$  of relation  $Restaurant(Type, Name, Avg.Price)$

#	Type	Name	Avg. Price
$t_1$	Indian	Shahi	9.20
$t_2$	Indian	Mandala	7.80
$t_3$	Vegetarian	Veggie Delite	12.10
$t_4$	Vegetarian	Happy Cow	11.30
$t_5$	German	Speckpalast	10.88

In order to evaluate a specified preference formula, Chomicki [2002] proposes the *winnow operator* that, to put it simply, returns (or filters) a set of the preferred tuples from a given relation instance. The winnow operator helps to embed preference formulas into the relational algebra (and SQL) by passing them as a parameter to the operator [Chomicki 2003, cf. p. 427].

**Definition 5.12 Winnow operator:** “If  $\mathbf{R}$  is a relation schema and  $\mathbf{C}$  a preference formula defining a preference relation  $\succ_C$  over  $\mathbf{R}$ , then the winnow operator is written as  $\omega_C(\mathbf{R})$ , and for every instance  $\mathbf{r}$  of  $\mathbf{R}$ :  $\omega_C(\mathbf{r}) = \{\mathbf{t} \in \mathbf{r} \mid \neg \exists \mathbf{t}' \in \mathbf{r}. \mathbf{t}' \succ_C \mathbf{t}\}$ .” [Chomicki 2003, Def. 2.3].  $\diamond$

This definition points out the close relationship of the winnow operator to the Pareto optimum. Furthermore, its dependence on a given preference formula that is defined for certain attributes of a relation gives Chomicki’s preference approach *ceteris paribus* semantics. That is, all attributes not associated with a preference formula are considered as of equal importance (or unimportance) during the application of the winnow operator.

Besides his seminal formalization of preferences, Chomicki also addresses the composition of different preferences using logical connectors, by exploiting mathematical properties such as the transitivity of strict preferences, and by prioritizing them similar to the preference hierarchies described by Lacroix & Lavency [1987]. For a full description of the different composition strategies, refer to Chomicki [2002, Sec. 4].

In 2002, roughly at the same time as Chomicki published his preference approach, Kießling [2002] proposed an independently developed preference framework [Chomicki 2003, cf. p. 430] that resembles Chomicki’s formal preference specification for the most part. Hence, only Kießling’s main idea and differences to Chomicki’s work are outlined here.

Kießling’s preference approach supports strict preferences with *ceteris paribus* semantics including preference hierarchies, is compatible with the relational model, uses preferences as a “filter” on a result set, and can combine multiple preferences. In contrast to Chomicki [2002], Kießling [2002] explicitly includes “numerical preferences” (i.e., *quantitative* preferences that resemble distance-based goal queries [Motro 1986] very much) in his preference algebra (see Section 5.3.2). Regarding preference composition, Kießling [2002] additionally supports a method for combining preferences Pareto-efficiently. A comparable composition is proposed by Chomicki as part of a comprehensive examination of preference compositions with the means of mathematical order theory [Chomicki 2003].

The most obvious difference between Chomicki [2002] and Kießling [2002] is that the latter does not rely on preference formulas or an arbitrary logical representation of preferences. Instead, Kießling proposes nine base preferences in form of “functions” for his preference algebra, e.g., *NEG()* to express dislikes or *BETWEEN()* to express a preference for a desired range of an attribute’s value [Kießling 2002, cf. Sec. 3.2]. As a consequence, Kießling [2002] presents his own preference evaluation model, the *BMO* query model (best matches only), which eventually coincides with the semantics of the winnow operator [Kießling 2002, cf. p. 321]. This similarity is also acknowledged by Chomicki [2003], who also points out the equality of the winnow and the *Best* operator [Torlone & Ciaccia 2002]. Considering the technical implication of the preference approaches, Kießling [2002] suggests to transform a BMO query (or Preference SQL) into SQL92 code [Kießling & Köstler 2002], while Chomicki [2002] argues in favor of the inclusion of a separate winnow operator.

#### The Skyline Operator

A special case of qualitative preferences is the so-called *skyline operator* [Börzsönyi et al. 2001; Kossmann et al. 2002]. The skyline operator can be seen as a restricted version of Kießling’s Pareto preference composition that allows only *LOWEST()* or *HIGHEST()* as its base preferences [Kießling 2002, Sec. 6.1.4]. In other words, preferences can only be specified for the maximum or minimum of the existent attribute values and not, e.g., for ranges (see above). Chomicki [2002] takes up a similar position understanding the skyline operator as “a special case of our winnow operator. It is restricted to use an arithmetical ipf [intrinsic preference formula] which is a conjunction of pairwise comparisons of corresponding tuple components” [Chomicki 2002, p. 49]. Following Definition 5.11, an intrinsic preference is a preference that relies only on the values occurring in the involved tuples taking part in the preference evaluation and the built-in predicates. In contrast, extrinsic preferences may also incorporate information from other relations or properties from the relations the tuples were selected from, e.g., aggregate values or properties “like membership of tuples in database relations” [Chomicki 2003, p. 429]. For more details, see [Chomicki 2002, Sec. 5.2].

The skyline operator has *ceteris paribus* semantics and relies heavily on the notion of Pareto optimality (see Definition 5.9), or as stated by Chomicki: “Only tuples with identical values of all DIFF attributes are comparable; among those, MAX attribute

values are maximized and MIN values are minimized.” [Chomicki 2007, p. 80].

To give an example, review the sample preference “prefer cheaper restaurants of the same type” from the beginning of this section (see Table 5.2). Following Chomicki’s terminology, DIFF is *Type* and MIN is *Avg.Price* as inexpensive restaurants should be preferred. Hence, tuples with the same *Type* are directly comparable (no matter how the other attributes look like because of the ceteris paribus semantics) and picked depending on their *Avg.Price*, i.e., the tuples that *dominate* others in the MIN attribute (see Table 5.3). The concept of dominance makes the skyline operator an intuitively comprehensible filter for large data sets because the filter criterion is very simple. The filter only removes tuples that are better or worse than other directly comparable ones in a set of given attributes. Using the example of the aforementioned preference, the filter semantics becomes very clear: if one can choose between two restaurants of the same type, why should one choose the more expensive one?<sup>60</sup> Figure 5.1 shows the dominant tuples from Table 5.3 as black filled circles. The connecting line forms a “skyline” of the resulting data set – hence the name of the operator.

The necessary search for dominant tuples links the skyline operator to the well-known maximum vector problem<sup>61</sup> [Godfrey et al. 2007].

Table 5.3: Dominant tuples of the sample instance  $r_1$  of relation *Restaurant*(*Type*, *Name*, *Avg.Price*)

#	Type	Name	Avg. Price	Dominant
$t_1$	Indian	Shahi	9.20	✗
$t_2$	Indian	Mandala	7.80	✓
$t_3$	Vegetarian	Veggie Delite	12.10	✗
$t_4$	Vegetarian	Happy Cow	11.30	✓
$t_5$	German	Speckpalast	10.88	✓

Because of the ceteris paribus semantics of the skyline operator, users are only required to state strict preferences on attributes they are interested in. In conjunction with its simple filter semantics, the skyline operator is widely considered very user-friendly.

### 5.3.2 Quantitative Preferences

Goal queries [Motro 1986] are an early contribution to the field of quantitative preferences in DB. Motro’s contribution takes a comparable place for quantitative approaches as the Lacroix & Lavency [1987] approach for qualitative preferences<sup>62</sup>. The core idea of goal queries is to relax the rigid exact matching paradigm of relational DB in favor of best matching. To discover the best matching tuples for a given query, Motro [1986] suggests to calculate the distance between attribute values in order to assess the compliance with a specified goal.

<sup>60</sup>This implies that all other attributes are equal because of the ceteris paribus semantics, e.g., the quality of the food.

<sup>61</sup>“The maximal vector problem is to find the subset of the vectors such that each is not dominated by any of the vectors from the set.” [Godfrey et al. 2007, p. 5]

<sup>62</sup>In fact, goal queries motivated the preference approach by Lacroix & Lavency [1987].

### 5.3 Preference Models for the Personalization of Queries

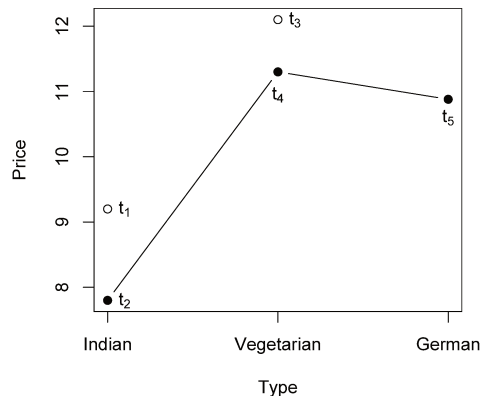


Figure 5.1: Skyline of  $r_1$  with  $\text{DIFF} = \text{Type}$  and  $\text{MIN} = \text{Avg.Price}$

In his paper, Motro motivates the need for goal queries with the help of an example and points out a potential problem of his approach:

“For example, which restaurant is closer to being an inexpensive French restaurant in the downtown area: a moderately priced Continental restaurant a few blocks south, or an inexpensive Chinese restaurant located in the heart of downtown? The need for distances becomes clearer if we examine numeric data. Because such data already embeds a notion of distance, goals that involve numbers are easier to handle.”

[Motro 1986, p. 130]

While the interpretation of distances between numeric attribute values needs no further discussion, the calculation of distances for attributes containing text (e.g., names or addresses) or more complex attributes is not straightforward because such attributes do not embed a natural notion of distance. To tackle this challenge, Motro proposes to augment the database schema with metadata about the attributes of each relation:

1. A *measure*  $\mathbf{M}$  that can be a typical distance function or a relation expressing the distance between the attribute's values,
2. A *scaling factor*  $\mathbf{s}$  that corrects the results of the distance calculation carried out by the measure in order to better express significant differences between attribute values,
3. A *relative weight*  $\mathbf{w}$  that expresses the importance of the attribute in comparison to all other attributes of the relation, and
4. A *neighborhood radius*  $\mathbf{r}$  indicating until which value of the scaled distance an attribute can be interpreted as neighboring.

## 5 Machine-based Learning of Personalized CQQL Queries

In order to state a goal (or best matching predicate), Motro introduces the binary *similar-to* comparator  $\approx$ . Let  $x$  and  $y$  be two attribute values, then the similar-to comparator is defined as follows:

$$x \approx y \quad \text{if} \quad \text{dist}_M(x, y) \cdot \frac{1}{s} \leq r,$$

where  $\text{dist}_M()$  is a distance function relying on the measure  $M$  that has to be defined for the attribute. As a consequence, only attribute values that yield a distance not greater than the neighborhood radius  $r$  are considered similar. If different goals are defined in one query by applying multiple similar-to operators, each distance is calculated separately and then aggregated with the help of the weighted sum using the relative weight for each involved attribute [Motro 1986, cf. p. 139].

The central presumptions of the goal query approach are the existence of a distance measure for each attribute and a valid aggregation mechanism, which allows the calculation of a tuple's proximity to a query containing multiple goals.

This links the approach (and all other quantitative approaches) to a central concept of microeconomics: the usage of *utility functions* to express preferences.

**Definition 5.13 Utility function:** *Let  $X$  be the set of possible alternatives an actor can choose from and  $\mathbf{x}_1, \mathbf{x}_2 \in X$ . The weak preference that  $\mathbf{x}_1$  is better than or equal in value to  $\mathbf{x}_2$  is expressed as follows:  $\mathbf{P} = \mathbf{x}_1 \succeq \mathbf{x}_2$ . Then, a utility function is defined as follows:*

$$u : X \rightarrow \mathbb{R}.$$

A utility function  $\mathbf{u}$  is representing a given preference  $\mathbf{P}$  iff :

$$\forall x_1, x_2 \in X \mid u(x_1) \geq u(x_2) \rightarrow x_1 \succeq x_2$$

◇

Ultimately, the application of a utility function on a tuple yields a (utility) score that can be used to produce a partial order (see Appendix A.1.1) of the examined tuples. Thus, in the field of DB, utility functions are often called *scoring functions*. The application of some sort of utility or scoring function is common to all quantitative preference approaches.

The actual calculation of a tuple's utility can be best described with the following example. Reconsider the vegetarian restaurant preference from page 86. With regard to the additional preference for low prices, we can develop exemplary utility functions for the attributes of the sample relation  $Restaurant(\text{Type}, \text{Name}, \text{Avg.Price})$  with the following properties:

$$u(\text{Type}) = \begin{cases} 2 & \text{if Type} = \text{"vegetarian"} \\ 1 & \text{if Type} = \text{"Indian"} \\ 0 & \text{otherwise} \end{cases}$$

That is, vegetarian restaurants are liked twice as much as Indian restaurants, while other restaurants are disliked.

$$u(\text{Avg.Price}) = \text{Avg.Price}/10 \cdot -1$$

### 5.3 Preference Models for the Personalization of Queries

In order to calculate the utility of a tuple, the sum of  $u(\text{Type})$  and  $u(\text{Avg.Price})$  is computed. Table 5.4 contains the results for the sample instance of the relation *Restaurant* ordered by the tuple's utility. In order to personalize the query, one could also use a

Table 5.4: Sample instance  $r_1$  of relation *Restaurant*(*Type, Name, Avg.Price*) ordered by their utility

#	Type	u(Type)	Name	Avg. Price	u(Avg. Price)	Utility
$t_4$	Vegetarian	2	Happy Cow	11.30	-1.13	0.87
$t_3$	Vegetarian	2	Veggie Delite	12.10	-1.21	0.79
$t_2$	Indian	1	Mandala	7.80	-0.78	0.22
$t_1$	Indian	1	Shahi	9.20	-0.92	0.08
$t_5$	German	0	Speckpalast	10.88	-1.09	-1.09

weighted sum to steer the influence of each attribute's utility on the tuple's utility.

At any rate, it does not matter for the calculation of the utility function whether the weights are part of an attribute's metadata as in Motro's goal queries or part of the user-defined query – although the latter is more intuitive. In fact, every retrieval model that supports the weighting of query terms, starting from the extended Boolean model [Waller & Kraft 1979] (see Sections 2.2.2 and 5.1.3), can be regarded as a quantitative preference model. This interpretation is also supported by Agrawal & Wimmers [2000]. Thus, this section omits a recapitulation of these retrieval models.

As said before, all quantitative preference approaches have in common that they rely on a utility function. Nevertheless, the actual design of this utility function differs. For instance, Fagin & Wimmers [1997] approach quantitative preferences in MIR by applying fuzzy logic [Zadeh 1988]. The authors acknowledge that typical attribute values in multimedia databases are typically in the interval  $[0; 1]$  expressing a similarity to, e.g., the color of an image. This similarity or relevance value is called a *score* by Fagin & Wimmers [1997]. To process complex queries, these scores must be combined using a *scoring function*, e.g., by using logical connectors based on fuzzy logic (see Section 2.2.2) to build complex queries. In any case, the authors criticize that such kind of score aggregation does not address individual user preferences between different sub-formulas, e.g., a user might be more sensitive to the color of an image than to its edges. Hence, Fagin & Wimmers [1997] propose a method that allows the weighting of these sub-formulas.

Extending their first contribution, Fagin & Wimmers [2000] investigate the nature of weighted scoring functions in more detail. They present an explicit formula for the incorporation of weights that is based on the unweighted original scoring function. Their approach is elegant because it uses the underlying scoring function directly when all weights are set equal and ignores sub-formulas if weighted with zero. Moreover, the resulting score for a tuple is a continuous function of the defined weights [Fagin & Wimmers 2000]. In particular, Fagin and Wimmers' scoring function has become important because it marks a reference point for other quantitative approaches such as the ones by Schulz & Schmitt [2003]; Schmitt [2007]; Sung & Hu [2009], or Zellhöfer & Schmitt [2010b]. Weaknesses of the approach have been addressed in Section 4.3.

Another proposal for incorporating quantitative preferences into databases is made

by Agrawal & Wimmers [2000]. The authors suggest to use the so-called *preference functions* that allow a sophisticated definition of *weak* preferences on attributes with the help of a weighting scheme in order to express the importance of a preference, a *veto* functionality (i.e., a dislike), or an indifference judgment. Furthermore, a fair preference composition method that at least preserves all stated vetos is discussed. The resulting preference formula is then used to compute the utility of each tuple. In contrast to Fagin & Wimmers [2000], Agrawal & Wimmers [2000] also address the implementation of their preference approach in relational databases.

To conclude, weighted CQQL (see Section 4.3) also constitutes a quantitative preference approach.

### 5.3.3 Comparison of Qualitative and Quantitative Approaches

Frequently, qualitative preferences are regarded as a generalization of quantitative approaches [Kießling 2002] providing a higher level of expressive power. As reported by Chomicki [2003], the difference in expressive power is also well-known in utility theory [Fishburn 1970].

To put it simply, qualitative approaches can model everything that can be imagined with preferences in the relational model. On the other hand, the result set semantics of qualitative preferences is less differentiating<sup>63</sup> than quantitative approaches that rely on the induction of a total order on a result set.

The difference between both approaches can be explained easily with the following example. In his paper on qualitative preferences, Chomicki [2002] defines a relation *Book*(*ISBN*, *Vendor*, *Price*) and a sample instance of *Book* shown in Table 5.5. In addition, he defines an illustrative sample preference:

“if two tuples have the same ISBN and different Price, prefer the one with the lower Price.” [Chomicki 2002, p. 34]

Table 5.5: Sample instance  $r_1$  of relation *Book* [Chomicki 2002, cf. Ex. 1]

#	ISBN	Vendor	Price
$t_1$	0679726691	BooksForLess	\$ 14.75
$t_2$	0679726691	LowestPrices	\$ 13.50
$t_3$	0679726691	QualityBooks	\$ 18.80
$t_4$	0062059041	BooksForLess	\$ 7.30
$t_5$	0374164770	LowestPrices	\$ 21.88

The evaluation of the given preference on the instance of *Book* yields the following poset (see Appendix A.1.1), which is also illustrated by the Hasse diagram (see Definition A.7) in Figure 5.2:

$$\{t_2 \succ t_1 \succ t_3, t_4, t_5\}$$

For clarity, the elements which are not part of the partial order because they are incomparable are depicted as well. The resulting skyline is emphasized. In order to find a

<sup>63</sup>In the sense that a discrimination between a “little better” and a “significantly better” element is not possible for the elements of a result set.



### 5.3 Preference Models for the Personalization of Queries

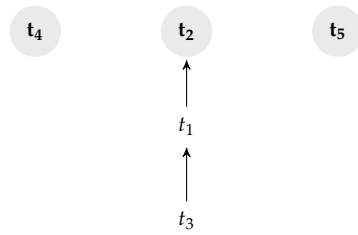


Figure 5.2: Extended Hasse diagram of a Chomicki’s preference example I [Chomicki 2002, cf. Ex. 1]; skyline elements are bold

quantitative preference with the same expressive power, one needs to develop a utility function that returns the same poset [Chomicki 2002, cf. Ex. 2]. Furthermore, Chomicki argues that because there is no preference between  $t_4$ ,  $t_5$ , and the first three tuples, one must assume that these tuples have the same utility as  $t_1 - t_3$ <sup>64</sup>. In a reverse conclusion, this means that the utility of the first three tuples must be equal, which is obviously impossible since  $t_2 \succ t_1 \succ t_3$  holds. This leads to the conclusion that “utility functions cannot represent all strict partial orders. For example, utility functions cannot capture skylines[...]. Also, ordered relations go beyond the classical relational model of data.” [Chomicki 2007, p. 80].

Although the latter argument is unquestionably true because utility functions induce a total order on sets and hence violate the set semantics of the relational model, the first conclusion needs further attention. From an order-theoretical point of view one might argue whether quantitative approaches are *only* a specialization of qualitative approaches because every total order also constitutes a partial order. Though, Zellhöfer & Schmitt [2010b] point out that “every poset with the Dushnik-Miller dimension  $d$  can be expressed by the intersection of  $d$  totally ordered sets” [Zellhöfer & Schmitt 2010b, p. 32]. For a detailed discussion, see Dushnik & Miller [1941] or Trotter [1992].

Although the aforementioned examples are primarily meant to support Chomicki’s pro-qualitative argument, they also show a major weak point of such approaches and skylines in particular. Notwithstanding the fact that skyline operators are relatively user-friendly during formulation (see Section 5.3.1), they can also cause side-effects that can lower their utility.

As illustrated in Figure 5.2 and caused by their *ceteris paribus* semantics, skylines typically contain tuples that are not directly comparable to each other. For instance, the ISBN of every tuple in the skyline differs. Hence, the tuples contained in the result set are considered as having the same utility to the user without acknowledging further attributes that might affect their utility. As a result, this can lead to very large result sets in case of many tuples with different values in at least one attribute (as ISBN in the example above) because each of these tuples are contained in the skyline. In fact, it has been shown that the size of the skyline result set can almost grow exponentially with

<sup>64</sup>Please note that it is arguable whether the absence of an explicit preference implies indifference between the alternatives. If one does not accept this premise, the following conclusions become invalid.

the number of attributes contained in the preference depending on the relative ordering and distribution of the examined data [Bentley et al. 1978].

One way to circumvent this effect is the induction of some sort of ranking on the skyline. Typically, the ranking is based on additional preferences between attributes. For instance, the Telescope algorithm [Lee et al. 2007] allows the formulation of strict preferences between attributes which are then used to rank tuples in the skyline in order to return the top- $k$  results of the skyline. Comparable methods have been suggested under the name of *preference trade-offs*, e.g., by Lofi et al. [2008]. These techniques have in common that they leave the relational model by inducing an order on a set – a common point of criticism on quantitative preferences. In fact, these approaches have to be considered a hybrid preference model because they first rely on a qualitative preference formulation and evaluation, which is later extended by a quantitative preference in order to limit the number of results. Taken to extremes, one could compare the relation of trade-off skyline operators to skylines operators with the relation of the Boolean model to the extended Boolean models (see Section 2.2.2).

Not surprisingly, quantitative preferences do not suffer from an extensive number of results because their induced total order allows the establishment of a result limitation at an arbitrary top- $k$  position. Furthermore, the induced order can be directly interpreted as the level of relevance (see Chapter 2). This ordering by relevance with respect to a given preference (or query) allows a much finer differentiation of the result in comparison to the qualitative approach that regards all elements of the result set as equally important.

The result differentiation is gained at the cost of the need for an utility function (see Section 5.3.2). One major criticism on quantitative approaches is that the formulation of quantitative preference in form of a utility function, e.g., by using weighting, is a complex task for users – especially if multiple attributes are involved. Generally speaking, the usage of weights demands clarity from users about their preferences between the available attributes in addition to the ability to quantify these preferences. In contrast, qualitative approaches rely on intuitively comprehensible preference statements on interesting attributes only and infer the remaining information with the help of their *ceteris paribus* semantics. In his early work, Motro [1986] puts it aptly: “However, selecting the weights that combine different attributes of a description into a measure, is much more difficult. Weights determine what is important in a description; and this may not have a unique answer” [Motro 1986, p. 146]. Consequently, Motro underlines the importance of a supportive user interface whenever quantitative preferences have to be elicited.

Another potential issue with quantitative preference approaches and their utilized scoring function has been presented in Section 4.3: not all scoring function are embedded into a logic. That is, they can violate certain logical laws. An example is the scoring function by Fagin & Wimmers [2000] that violates the law of associativity. In contrast, qualitative preference approaches are fully compatible with the relative model and its underlying logic [Chomicki 2002; Kießling 2002]. On the other hand, it has been shown that CQQL’s quantitative weighting does not violate the laws of Boolean algebra [Schmitt 2007].

An additional aspect of preference approaches that becomes particularly important in interactive scenarios is the computational complexity of the preference evaluation and thus its resulting processing time. It is widely known that the computation of a skyline is very CPU-intensive [Börzsönyi et al. 2001; Chaudhuri et al. 2006] and depends heavily on the structural and statistical properties of the data for which the skyline is computed for [Chomicki et al. 2003]. For instance, Godfrey et al. [2007] show that the average-case complexity of the skyline operator is  $\mathcal{O}(kn)$ , with  $k$  being the number of attributes in the preference and  $n$  the number of tuples in the relation instance. The reported worst-case complexity is  $\mathcal{O}(n^2)$  and normally occurs when the attribute dimensions are anti-correlated [Chomicki et al. 2003; Shang & Kitsuregawa 2013], i.e., one value in attribute dimension  $a_1$  decreases while another in dimension  $a_2$  increases. Similar results are presented for the *Best* operator (see Section 5.3.1) with a worst-case complexity of  $\mathcal{O}(n^2)$  [Torlone & Ciaccia 2002]. This finding can be transferred directly to the winnow operator because of the equality of both operators (see Section 5.3.1).

Unfortunately, anti-correlated attribute dimensions are fairly common in real-world scenarios. Imagine restaurants close to a touristic location. It is very likely that the average price for a meal in a restaurant increases as its distance to the sightseeing location decreases. In contrast, the calculation of the utility of  $n$  tuples has an average- and worst-case complexity of  $\mathcal{O}(n)$  [Chomicki et al. 2003]. This complexity is not dependent on the underlying data because the utility has to be calculated for each tuple. The complexity to store the tuples in a list ordered by their utility is  $\mathcal{O}(\log n)$  (assuming binary search) [Sedgewick & Wayne 2012].

The comparison of qualitative and quantitative preference approaches can be best concluded with a bon mot: *there ain't no such thing as a free lunch*. Obviously, the utility of qualitative and quantitative preferences depends on the usage scenario. The next section reviews the advantages and disadvantages of the particular approaches from a MIR point of view.

### 5.4 PrefCQQL – Preferences within the CQQL-based Retrieval Model

Although early contributions to the idea to combining preferences with CQQL were motivated by potential usage scenarios in the field of MIR [Zellhöfer & Lehrack 2008; Zellhöfer & Schmitt 2010a], the *PrefCQQL* approach presented in this part of the thesis is not limited to this domain. Instead, the central article on the combination of CQQL with preferences [Zellhöfer & Schmitt 2010b] discusses how a widely applicable user-friendly preference elicitation and evaluation method can be realized with the help of CQQL and machine-based learning.

Because of the scope of this dissertation, we only investigate aspects of this approach specific to the usage area of MIR. For a more database-specific point of view, see Schmitt & Zellhöfer [2009] or Zellhöfer & Schmitt [2010b].

Section 4.4 provides an in-depth discussion on how CQQL can serve as a logical

## 5 Machine-based Learning of Personalized CQQL Queries

query language in MIR. Furthermore, Section 5.3.2 has shown that weighted CQQL constitutes a quantitative preference approach.

For the remainder of this text, it is necessary to discriminate between the preference model used in the user's *cognitive space*, i.e., the preference model the user is directly confronted with, and the preference model in the MIR system's *information space*<sup>65</sup>.

To justify this discrimination, it is essential to recapitulate the central challenges in MIR that have been highlighted throughout this thesis with a special focus on users. Consecutively, the applicability of qualitative and quantitative preference models in MIR are assessed on the basis of the following MIR specifics:

1. MIR is multimodal (see Section 2.3), i.e., high- and low-level features of various types are used to represent documents.
2. MIR is subject to the semantic gap and the related query formulation problem (see Section 2.3.3). In particular, low-level features often cannot be used directly during query formulation as they rely on technical terms and concepts that are hardly comprehensible to typical users. To address the query formulation problem, MIR systems often support QBE to state queries (see Section 3.3.1).
3. In order to assess the relevance of a document with respect to a query, MIR engines typically use distance functions to calculate the distance between low- or high-level features (see Section 2.3.2)<sup>66</sup>. The results of these distance calculations or an aggregation of them are interpreted as the relevance of a document which yields a total order of the documents in a collection (see Sections 2.2.2 and 2.3.2).

### Qualitative preference models in the context of MIR

The general user friendliness of the qualitative preference model has been discussed in Sections 5.3.1 and 5.3.3. Its user friendliness is primarily based on its intuitive comprehensibility of the preference formulation at the attribute level.

Without a doubt, the statement of a preference of one attribute value over another implies that the attribute and its possible values are known to the user. For instance, it is easy to express a preference for a restaurant depending on its type attribute as shown with the *Restaurant* relation in Table 5.2.

In MIR, the attributes representing a document are not directly comprehensible; neither are their possible values known or make sense to the typical user. This is in particular true for low-level features. To give an example, imagine a low-level feature counting the number of different edges present in an image (see Table 2.1; edge histogram). It is hard to believe that a user can express a preference for edges running in a 45° angle over the amount of edges running horizontally. Moreover, there is no guarantee that the user's mental model matches the system's interpretation of an edge. That is, a user might interpret an edge as a contour line outlining distinct image regions in an

---

<sup>65</sup>For the sake of consistency, we follow Ingwersen's terminology for polyrepresentation [Ingwersen 1996] introduced in Section 3.2.

<sup>66</sup>The same statement holds for similarity functions.

image, while the system is considering all gradients on the luminance channel of the same image as edges. Figure 5.3 illustrates this effect by juxtaposing a possible user’s interpretation of edges (center) with the system’s interpretation (right) of the original image (left).

This claim is supported by the user study presented in Section 9.4, which clearly shows that layperson users have problems to decipher the meaning behind the low-level features’ names even if they are presented to them with the help of a GUI.



Figure 5.3: Different understanding of edges in an image; from left to right: original image, interesting regions, edges based on gradient detection

In addition, qualitative approaches support only strict preferences. However, it cannot be guaranteed that preferences can always be considered strict. For instance, Bates’ berry picking model and other user-oriented studies argue that a user’s IN might not necessarily be satisfied by a document as a whole (see Section 3.1.1). Furthermore, the polysemic nature of multimedia documents (see Section 2.3.3) suggests that more than one document or one document’s features might be equally relevant to a given IN. To give an example, imagine a document  $d_1$  that addresses one aspect of the IN, while  $d_2$  addresses another one. When compared directly, both documents are equally important to the user because of their different content. Hence, the formulation of a strict preference  $d_1 > d_2$  is infeasible for the user.

The possibility to be embedded into the relational model is one feature of qualitative preferences that is very important in the field of DB. However, this point is of less importance for MIR because modern IR models do not use this theoretical foundation. In fact, the resulting set semantics of the relational model causes effects such as the incomparability of the results of qualitative approaches (see Section 5.3.3) and potentially large result sets. Furthermore, these approaches violate the probability ranking principle (see Definition 2.18), which is central to the dominant (M)IR models [Fuhr 2008].

To conclude, qualitative preference models comprise the risk of a high computational complexity that depends heavily on the distribution of the values of the document representations.

### Quantitative preference models in the context of MIR

Unlike qualitative approaches, quantitative preference models yield a total order of result documents due to the used utility function (see Section 5.3.2). Thus, they are

fully compliant with the probability ranking principle.

Regarding their complexity during preference formulation, they have to be considered more complicated than qualitative preferences. This is due to two main reasons. First, they suffer from the same problems as their qualitative counter-parts in terms of the comprehensibility of the low-level features in MIR. Second, quantitative preference models require the user to quantify their preference between attributes, which is not as intuitive as simply stating that one attribute is more important than another. Hence, a supportive GUI is needed to assist users during the formulation of a preference [Motro 1986].

In contrast to qualitative approaches, quantitative preference models often support the formulation of weak preferences (see Definition 5.3). As a consequence, users are in principle enabled to express *explicit* indifferences between attributes (see Section 5.4.3). Qualitative preferences support only implicit expressions of indifference via their *ceteris paribus* semantics of the attributes that are not participating in any preference relation (see Definition 5.10).

Furthermore, the reliance of quantitative preference models on the calculation of the utility for each document in the collection with the help of a utility function renders their computational complexity independent from the actual data distribution (see Section 5.3.3).

### Summary

To sum up, qualitative preferences are typically easier to elicit than quantitative ones although their practical feasibility depends *in every case* on the comprehensibility of the used attributes or document representations. This distinguishes the preference elicitation process in MIR from databases where each attribute of a relation bears a meaning to the user.

Additionally, the set semantics put forward by qualitative approaches is not natural to common retrieval models. In contrast, modern MIR relies on total orders of documents following the probability ranking principle. As pointed out before, this includes the quantum logic-based retrieval model behind CQQL (see Chapter 4).

In other words, the natural preference model (i.e., qualitative preferences) in the user's cognitive space is different from the quantitative one used in the system's information space. Consequently, a usable MIR system must strive for bridging the gap between the user's mental model and the system's conceptual model. With reference to the cognitive viewpoint in IR, this means that both the user and the system have to step into a dialog (see Section 3.1.2).

### 5.4.1 Preference Reasoning: Deduction and Induction

In order to establish a dialog between the cognitive and information space, it is necessary to examine the reasoning about preferences in addition to their type. In principle, there are two ways of reasoning: deductive and inductive reasoning.

*Deductive reasoning* is the process of inferring the truth of a statement from a general rule. Both preference formulation approaches presented in Section 5.3 are deductive. That is, the preferences define a general, valid rule against which all objects are evaluated. As a result, only objects satisfying this rule are considered “true” or relevant. In other words, the reasoning is *top-down*: individual objects are chosen following a rule which is based on an axiomatic system such as classic logic (see Appendix A.1.4). To give an example, the deductive preference *vegetarian restaurant*  $\succ$  *Indian restaurant*  $\succ$  *other restaurant* can be evaluated against each imaginable instance of the relation *Restaurant*(*Type, Name*) returning only the tuples that satisfy it, e.g., by using the winnow operator (see Definition 5.12).

*Inductive reasoning* means to conclude a general rule from special cases or observed instances. In contrast to deductive reasoning that always yields a certain conclusion, the conclusion of inductive reasoning is only probable and can even be wrong. For instance, one observes that *X* is a robin and *X* is a bird. Using inductive reasoning, one concludes that all birds are robins. Obviously, the conclusion is wrong as there are many more bird species. This effect is known as the “problem of induction” and has been extensively studied in philosophy, e.g., by David Hume or Karl Popper. Nevertheless, the usage of inductive reasoning, i.e., to make a series of observations and to conclude a (potentially) valid general rule, is inherently human [Hume 2009].

**Induction in the context of interactive MIR** In order to overcome the preference formulation problem due to incomprehensible document representations and an un-intuitive preference mechanism in MIR, Zellhöfer [2010a, b] suggests to use so-called *inductive preferences* during user interaction, i.e., preferences that are based on inductive reasoning. In order to elicit an inductive preference, users can define qualitative preferences between documents, which are then used to conclude a general, valid preference by applying machine-based learning. This links the inductive preference approach to learning to rank techniques (see Section 5.2) and other fields of AI such as inductive learning or learning by example. Following this argumentation, traditional explicit RF (see Section 5.1.1) can also be seen as an inductive preference approach. In fact, QBE approaches in MIR are also following the inductive paradigm because they try to conclude the user’s current IN from given examples.

On a more formal level, inductive reasoning is also related to probability theory, statistic inference, and decision theory [Holland et al. 1996] because of its inherent uncertainty. Furthermore, Holland et al. [1996] argue that “induction is (a) directed by problem-solving activity and (b) based on feedback regarding the success or failure of predictions generated by the system” [Holland et al. 1996, p. 9]. This links inductive reasoning to the understanding of the search process in IIR (see Chapter 3) and the need for a constant dialog between the user and the system.

Besides their inherent uncertainty, inductive preferences are typically not *complete*.

**Definition 5.14 Incompletene preferences:** *Preferences are called incomplete, if they are not defined between all available alternatives (or documents in the scope of this dissertation).*  $\diamond$

## 5 Machine-based Learning of Personalized CQQL Queries

Incomplete preferences are typical in the field of MIR because the formulation of complete preferences would require users to inspect all documents in the collection. Incompleteness can be caused by a lack of knowledge, reflection, or disinterest to elicit further preferences. Hansson & Grüne-Yanoff [2012] subdivide incomplete preferences by the way they can be resolved, i.e., how the preference relation can be extended to include all alternatives, into three groups:

First, *uniquely resolvable* incomplete preferences can be resolved in one way. That is, although the user states incomplete preferences, complete preferences can be constructed, e.g., by using observation or logical inference. For instance, quantitative preferences are complete because they quantify the importance of each involved attribute with the help of a utility function (see Section 5.3.2). At first sight, qualitative preferences allow the formulation of incomplete preferences because users are only required to state preferences between attributes they are interested in. For all other attributes, preferences are resolved with the help of the *ceteris paribus* semantics of qualitative preference approaches (see Definition 5.10) which completes the preferences. Hence, all deductive preference approaches are complete.

Second, incomplete preferences can be *multiply resolvable*, i.e., there are multiple ways to complete the preferences. For instance, a machine-based learning algorithm may infer different complete preferences that are compatible with the specified incomplete preferences (see Section 5.4.4).

Lastly, incomplete preferences can be *irresolvable*. Incomplete preferences are irresolvable if two alternatives are incommensurable, i.e., whenever it is not possible to compare both alternatives with the same measurement (or unit, class of advantages and disadvantages etc.). To illustrate the concept, Hansson & Grüne-Yanoff [2012] provide an example of a typical moral dilemma: a person might be unable to express her preference between the death of two acquaintances and the death of a close friend. For instance, the preference for the friend's survival might be based on the shared memories and close relationship, while the survival of the acquaintances might be preferred because both have to support their families. Hence, the "value" of the death of one of the two parties is based on different measurements or assessment criteria.

This example can be extended to different chains of preferences which are *per se* valid but incommensurable to each other. For instance, two valid sets of preferences  $P_1 = \{a \succ b \succ c\}$  and  $P_2 = \{c \succ a \succ b\}$  cannot be combined, because the premise for the stated preferences differs, i.e., the preferences are conflicting.

Another reason for such *preference conflicts* can be little reflection on the stated preferences or erroneous user input. Because of their importance to the preference model presented in this thesis, preference conflicts are covered separately in Section 5.4.7.

Although inductive reasoning with preferences has downsides regarding its general reliability, the next section presents a user-centered preference model which uses inductive reasoning to conclude quantitative preferences that can be used deductively. Besides their disadvantages, inductive preferences offer an intuitive form of preference formulation. To sum up, the reliance on inductive reasoning could not be proven to be problematic in MIR as the success of QBE and RF approaches demonstrates.



### 5.4.2 A User-centered Preference Model for MIR

The last section has shown that typical preference models are based on deductive reasoning. Likewise, the reasoning of most modern IR models is also deductive as shown by van Rijsbergen's interpretation of IR as uncertain inference [van Rijsbergen 1986b] (see Section 2.2.2)<sup>67</sup>. Moreover, Nottelmann & Fuhr [2003] argue that the deduction carried out by IR systems is in fact a generalization of the deduction used in databases that differs in one aspect: the outcome of the deductive conclusion in IR, i.e., the relevance decision for a document, is uncertain.

Unlike the field of DB, the query in IR, i.e., the general rule that will be used in the deductive relevance decision process, can be based on inductive reasoning. In other words, the general IN in form of a query representation is inferred from a sample, e.g., a relevant QBE document in MIR or relevant keywords in the case of textual IR. In contrast, the IN in DB is typically specified in a deductive form, i.e., a logical sentence (see Appendix A.1.4) that is used to infer whether a given tuple (or document) complies with the assertions made in the sentence.

Furthermore, the last sections made clear that common MIR models can be regarded as quantitative preference models as they support the weighting of query parts.

#### Linking Cognitive and Information Space

The recently presented examples point out a general problem of preferences in IR that has already been mentioned briefly at the beginning of Section 5.4: the system's reasoning is deductive, while most users' intuitive way of reasoning is inductive. In other words, the preference model in the user's *cognitive space* differs from the model used in the *information space*: users are familiar with stating qualitative preferences based on examples, typical MIR models are quantitative as outlined before.

In order to address the challenge of linking these two spaces, it is essential to differentiate *for now* between preference elicitation and reasoning in the cognitive and information space. This dichotomy is illustrated in Figure 5.4. The illustration depicts the two actors participating in the IIR process: the user and the system; each having their own representation of the current IN and the corresponding preferences. Eventually, the shown information flow between the different preference representations forms a *relevance feedback cycle*. In other words, a continuous dialog between the cognitive and information space as postulated by Holland et al. [1996] and extensively discussed in Chapter 3 is established.

Before proceeding with the discussion on this cyclic and thus interactive information seeking process, it is necessary to define the central personalization concepts and terminology of the CQQL-based retrieval model – *PrefCQQL* – discussed in the remainder of this thesis. Each of the concepts are explained in more detail in the following sections.

---

<sup>67</sup>In fact, this is true for most computer applications because they rely on axiomatic systems such as logics – in particular for logic-based retrieval as elaborated in Section 4.4.1.

## 5 Machine-based Learning of Personalized CQQL Queries

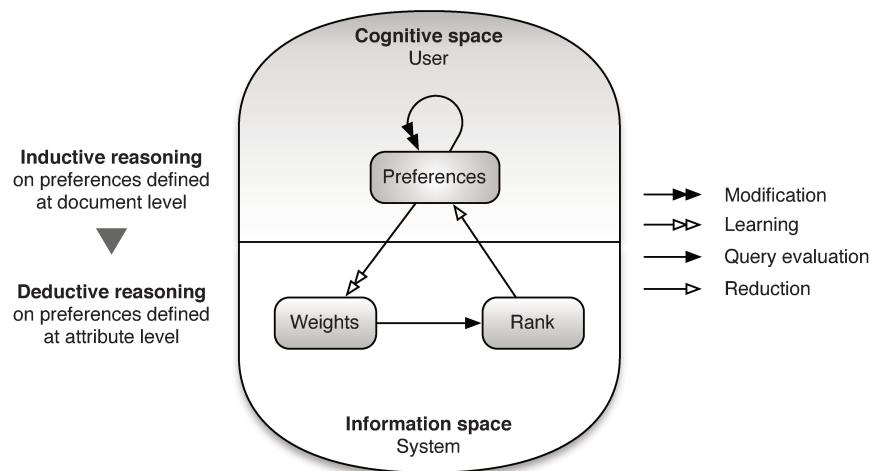


Figure 5.4: Interactive preference information flow within the CQQL-based retrieval model PrefCQQL

### Definition 5.15 Core concepts in the CQQL-based retrieval model PrefCQQL:

*(Inductive) Preferences:* For the sake of clarity, the term preference will only be used for qualitative preferences in the cognitive space, i.e., an inductive preference [Zellhöfer 2010a, b], from now on. Preferences are stated at the document level, e.g.,  $d_1 \succeq d_2$ .

*Modification:* Referring to the dynamic nature of the IN, preferences undergo constant change and are modified by the user. For instance, new preferences can be stated or existing ones may be removed.

*Learning:* Machine-based learning is used to infer a general, valid rule from the user-defined qualitative, inductive preference.

*Weights:* The result of the learning is a set of weights for a given, weighted CQQL query, i.e., a quantitative preference (or utility function, see Definition 5.13), which expresses the user-stated preference  $d_1 \succeq d_2$ .

*(Deductive) Query Evaluation:* The deductive query evaluation is the actual matching step carried out by the MIR engine, which takes the learnt weights to produce a relevance ranking of the evaluated documents.

*Rank:* The results of the query evaluation is a rank, i.e., a list of documents ordered by their probability of relevance.

*Reduction:* The reduction yields a set of characteristic preferences that are necessary to recreate the current rank.

◇

## 5.4 PrefCQQL – Preferences within the CQQL-based Retrieval Model

As said before, inductive preferences are intuitively comprehensible for users because they are used to make preference judgments at the document level. Hence, “it is not necessary that the user has a deeper understanding of the semantics of all features [or representations] that are used for document description nor of the available features or the underlying theoretical IR model. The user directly interacts with the document collection in order to compare pairs of documents” [Zellhöfer 2010b, p. 93]. Without doubt, users are familiar with stating preferences this way because decisions that neglect underlying or hidden features of the judged objects are made frequently in daily life. For instance, if someone has to decide between two apples, one might prefer the visually more pleasing apple, neglecting factors such as the amount of contained vitamins or its potential exposition to toxic pollutants.

How such preferences can be elicited from and modified by the user during the information seeking process is separately discussed in Section 5.4.3. Section 6.2.2 deepens this aspect further and shows how preferences integrate into a usable RF mechanism for MIR.

The elicited preferences serve as the input for the machine-based learning algorithm presented in Section 5.4.4. The learning step is needed to link the qualitative preferences from the cognitive space to the IR system’s quantitative weighting model – *weighted CQQL* – that is used to assess the documents’ relevance with respect to a given query (see Chapter 4).

Section 5.4.8 brings full circle by showing how the learnt weights can be communicated to the user in form of preferences (see Figure 5.4).

### 5.4.3 Interpreting Preferences as Partially Ordered Sets

The idea to interpret preferences as posets becomes obvious if one compares the mathematical properties of preferences as presented in Section 5.3 with those of posets (see Appendix A.1.1). This idea is also advocated by other researchers, e.g., by Chomicki [2002]. Table 5.6 juxtaposes the formal similarities.

Table 5.6: Comparison of the mathematical properties of preferences and posets

Property	Weak Preferences	Weak Posets	Strict Preferences	Strict Posets
Anti-symmetry	✓	✓	✓	✓
Transitivity	✓	✓	✓	✓
Reflexivity	✓	✓	✗	✗
Irreflexivity	✗	✗	✓	✓

Besides their apparent similarities at the mathematical level, one can exploit the fact that posets directly correspond to graphs [Hollas 2007, cf. Sec. 1.4.7]. That is, a set of preferences/a poset<sup>68</sup> corresponds to a directed acyclic graph (DAG), i.e., a Hasse diagram. Moreover, preferences can also be visualized in an intuitively comprehensible manner by using graphs which renders this interpretation compelling from a user’s perspective.

<sup>68</sup>For the sake of brevity, from now on a set of preferences is considered to be a poset.

## 5 Machine-based Learning of Personalized CQQL Queries

Section 5.3 showed earlier that qualitative preference approaches typically rely on strict preferences and thus form strict posets, whereas quantitative approaches support weak preferences.

### Basic preference modification and elicitation

Preferences can be expressed as a directed graph  $G = (V, E)$ , with  $V$  being the set of nodes (or vertices), i.e., the documents  $d_i$  in collection  $D$ , and  $E$  being the set of edges that state the preference relation between two nodes.

It is important to differentiate between actual documents represented by a node and their *utility* to the user represented by an edge, e.g., their contribution to the satisfaction of the current IN. Each preference is expressed as a directed edge between two nodes pointing to the preferred node.

Extending [Zellhöfer & Schmitt 2010b] by including indifference, five user interactions during the modification process become possible with a weak preference graph. The modifications assume that preferences are united into one set of preferences, i.e., no voting rules or the like are utilized during the preference creation or modification phase (see Section 5.3).

For the sake of simplicity, the interactions are illustrated with two nodes depicting the documents and an edge showing the preference. Each figure shows a sample preference's status before and after the respective user modification. Furthermore, we assume the presence of a ranking of the documents  $d_i \in D$  that has been produced by a matching function before the user interaction happens. The extension to more complex samples is left open to the reader.

**Definition 5.16 Creation:** *A preference can be created if it expresses the intended IN better than the currently retrieved rank. That is, the original rank that places  $d_1$  before  $d_2$  (see Figure 5.5, left) is contradicting the user's notion of relevance, in which  $d_2$  is preferred over  $d_1$  (see Figure 5.5, right).*  $\diamond$

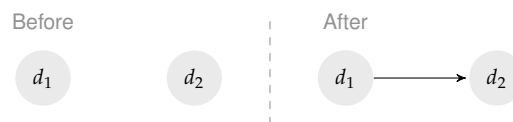


Figure 5.5: Creation of a preference; arrows point to the greater element

**Definition 5.17 Confirmation:** *If a preference is consistent with the user's IN, it can be confirmed<sup>69</sup> (see Figure 5.6).*  $\diamond$

**Definition 5.18 Inversion:** *An existing preference can be inverted if it contradicts the current IN (see Figure 5.7).*  $\diamond$

<sup>69</sup>Although a confirmation does not require active input by the user, it is included in this list for the sake of completeness.

#### 5.4 PrefCQQL – Preferences within the CQQL-based Retrieval Model

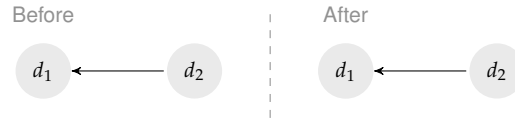


Figure 5.6: Confirmation of a preference; arrows point to the greater element

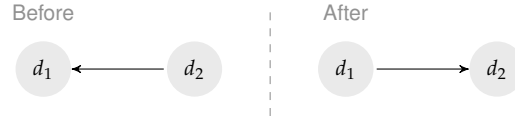


Figure 5.7: Inversion of a preference; arrows point to the greater element

**Definition 5.19 Removal:** *A preference can be removed if it no longer models the user’s IN (see Figure 5.8).*  $\diamond$

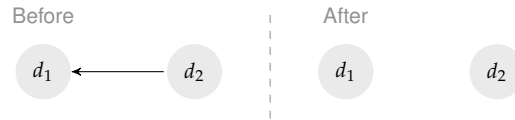


Figure 5.8: Removal of a preference; arrows point to the greater element

**Definition 5.20 Indifference:** *If two documents are of equal value to the user, symmetric indifference (see Definition 5.5) can be expressed (see Figure 5.9).*  $\diamond$

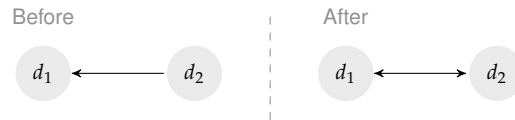


Figure 5.9: Indifferent preference; arrows point to the greater element, bi-directed edges indicate indifference

Indifference takes on a special position because it is not a typical preference judgment *per se* and has therefore been neglected by Zellhöfer & Schmitt [2010b]. In fact, the statement of indifference judgements at UI level can be in many cases considered *syntactical sugar* that are internally transformed into  $\succeq$  (see Section 5.4.7).

Unfortunately, indifference statements increase the chance to formulate cycles within the preference graph because of their symmetry. Nevertheless, indifference statements complete the view on preferences as described in Section 5.3. As such, we assume

that the explicit statement of indifference is more intuitive for the user than working with  $\succeq$  alone. Moreover, Section 5.4 already pointed out that the expression of a strict preference between two documents might not always be possible.

To recapitulate, the aforementioned graph modifications at the user interface level can be used to build a complex preference graph which does not necessarily need to be acyclic. Figure 5.10 shows a Hasse diagram (i.e., two DAG, see Definition A.7 in Appendix A.1.3) of exemplary acyclic preferences. After the user has finished the modification of the preference graph, the altered graph serves as input for the learning algorithm presented in Section 5.4.4.

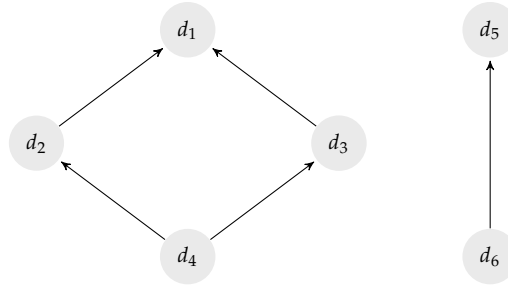


Figure 5.10: Hasse diagrams visualizing  $\{d_1 \succ d_2, d_1 \succ d_3, d_2 \succ d_4, d_3 \succ d_4, d_5 \succ d_6\}$

### Preferences as a Generalization of Traditional Relevance Feedback

Furthermore, the interpretation of preferences as posets emphasizes that preferences are a formal generalization of traditional relevance feedback described in Section 5.1.1. This includes approaches that allow only positive feedback such as the ones by Rocchio [1971] or Campbell [2000] and such that incorporate negative feedback, e.g., by Ide [1971] or Assfalg et al. [2000a].

Regarding positive relevance feedback, the relevant documents  $d_i^+ \in D$  form the *maximal* elements of the poset<sup>70</sup>. Analogously, the *minimal* elements of the poset can be defined by using the irrelevant documents  $d_j^- \in D$ . Figure 5.11 illustrates the preference graph for a collection  $D$  with  $m$  documents  $d_x \in D' = D \setminus (\{d_i^+\} \cup \{d_j^-\})$  (the documents on which no relevance feedback has been given) and the positive and negative relevance feedback documents. Obviously, there is no need for users to construct these preference graphs manually because the needed preferences can be inferred from the definition of the maximal/minimal elements of a poset. Hence, the preference approach also supports *binary*<sup>71</sup> relevance feedback.

To conclude, the utilization of preferences allows a combination of binary and gradual relevance feedback. From a user's point of view, this is desirable because it allows the exploration of a document collection by stating simple RF in binary form while

<sup>70</sup>There is not necessarily a supremum because the poset is not total.

<sup>71</sup>That is, *only* positive and negative RF can be given.

maintaining the opportunity to give more complex, gradual RF in form of finely graduated preferences whenever the user obtains more clarity about the current IN. For example, to introduce the concept of “highly” and “slightly” relevant into the graph shown in Figure 5.11 one would simply insert a new layer of “slightly” relevant documents between the maximal elements and  $D'$ .

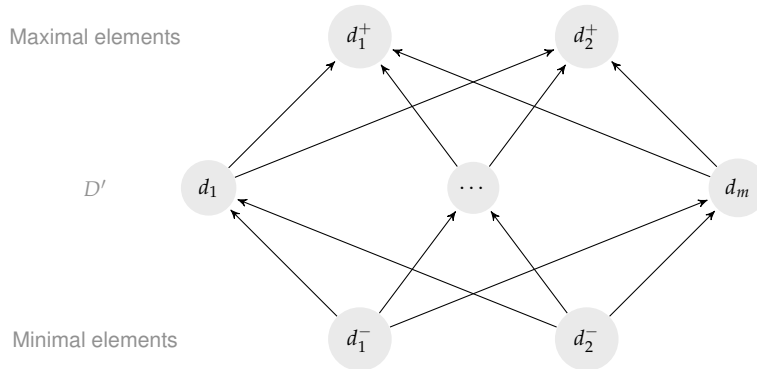


Figure 5.11: Preferences as a generalization of traditional relevance feedback

#### 5.4.4 The Learning of Weights as a Non-Linear Optimization Problem

To keep things simple, it is assumed that the preferences used as input to the learning algorithm are conflict-free and compatible with a given weighted CQQL condition. Preference conflicts and query incompatibilities are discussed in Section 5.4.7.

Following Section 5.4, the learning algorithm’s objective is to *complete* the incomplete inductive preferences that have been defined by the user. To achieve this, the learning algorithm infers weights (or quantitative preferences) for a weighted CQQL query (or condition) from a set of inductive preferences. In other words, the algorithm searches for a utility function (see Definition 5.13) that satisfies all preferences  $p \in P$ , with  $P$  being the set of all formulated preferences. This utility function is defined by a weighted CQQL condition  $q_\theta$  and a weighting scheme  $\omega$ .

**Definition 5.21 Weighting scheme  $\omega$ :** A weighting scheme  $\omega$  is a function that assigns a weight value  $w_{\theta_i} \in [0, 1]$  to each weighting variable  $\theta_i$  that is used in  $q_\theta$ .  $\diamond$

Hence, the utility of a document can be calculated by applying a utility function  $eval()$ .

**Definition 5.22 Evaluation function  $eval()$ :** The utility function  $eval()$  evaluates a given, weighted CQQL condition<sup>72</sup>  $q_\theta$  using the weighting scheme  $\omega$  on a document  $d \in D$ . Following Definitions 4.12ff., it is defined as follows:

$$eval(q_\theta, \omega, d) \rightarrow [0, 1]$$

<sup>72</sup>This condition can be user-defined or pre-defined by the system.

◇

**Definition 5.23 Preference fulfillment:** *In accordance with Definition 5.13, a weighting scheme  $\omega$  fulfills a preference  $\mathbf{p} = \mathbf{d}_1 \succeq \mathbf{d}_2$  if and only if:*

$$eval(q_\theta, \omega, d_1) - eval(q_\theta, \omega, d_2) \geq 0 \quad (5.2)$$

*holds.*

◇

Furthermore, the utility of  $p$  with respect to a given CQQL condition and a weighting scheme is defined as follows.

**Definition 5.24 Preference utility:**

$$util_\omega^p = eval(q_\theta, \omega, d_1) - eval(q_\theta, \omega, d_2)$$

◇

Figure 5.12 visualizes the results of the evaluation of the given CQQL condition  $q_\theta = r_1 \wedge_{\theta_1, \theta_2} r_2$  with respect to different weight values for the weighting variables  $\theta_i$  and with  $r_j$  being the two attributes (and values respectively) of the relation  $Document(r_1, r_2)$  representing the documents as  $d_1(0.7, 0.3)$  and  $d_2(0.6, 0.4)$ .

For the sake of clarity, arrows emphasize some values of  $util_\omega^p$  with  $p = d_1 \succeq d_2$  for various  $\omega$ . Figure 5.12 clearly illustrates that the preference  $p$  cannot be fulfilled with any combination of  $\theta_1$  and  $\theta_2$ . Consequently, a preference can be seen as a restriction on the weight values.

To deepen the understanding of a preference's utility, it is helpful to imagine  $util_\omega^p$  as a hyperplane that is defined on a hyper-unit cube spanned by the weighting variables  $\theta_i$  being used in  $q_\theta$ . Figure 5.13 depicts the three possible preference utility categories using two weighting variables. As before,  $util_\omega^p$  depends on the actual weight values and the result values of the evaluation function  $eval()$  [Zellhöfer & Schmitt 2010b, cf. pp. 39f.].

**Definition 5.25 Preference utility categories:**

***Inconsistent:** A preference  $\mathbf{p}$  is inconsistent if  $\max_\omega(\mathbf{util}_\omega^p) < \mathbf{0}$  holds (see Figure 5.13; left).*

*That is, the preference cannot be fulfilled by any weight values with a given  $\mathbf{q}_\theta$ . Obviously, this is the case if the maximum function value of  $\mathbf{util}_\omega^p$  is less than zero.*

***Useless:** A preference  $\mathbf{p}$  is useless for weight learning if  $\min_\omega(\mathbf{util}_\omega^p) \geq \mathbf{0}$  applies (see Figure 5.13; center). In this case, the preference does not restrict the weighting variables  $\theta_i$  used in  $\mathbf{q}_\theta$ .*

***Useful:** Otherwise, the preference is called useful (see Figure 5.13; right) because it intersects the zero level. The intersection of the zero level indicates a preference that constitutes a meaningful restriction of the weight values.*

◇



## 5.4 PrefCQQL – Preferences within the CQQL-based Retrieval Model

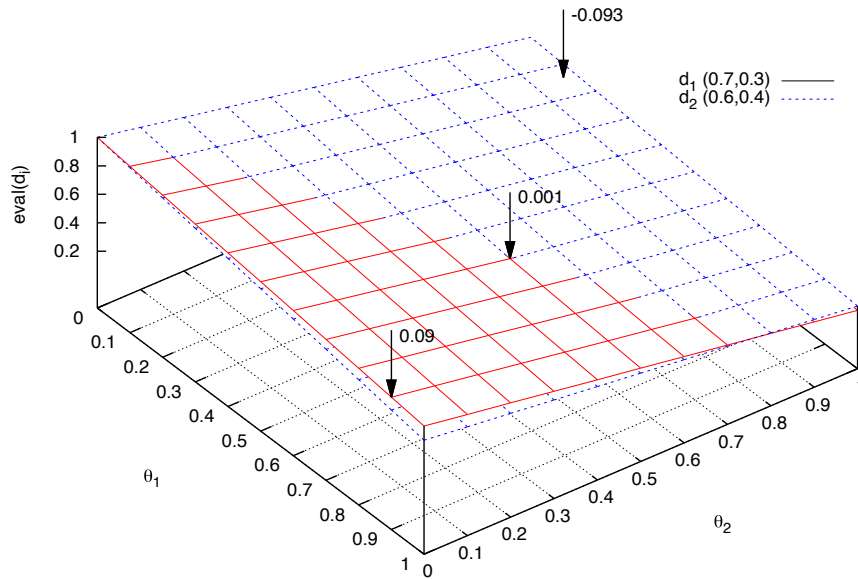


Figure 5.12: Preference utilities for  $q_\theta = r_1 \wedge_{\theta_1, \theta_2} r_2$  and  $p = d_1 \succeq d_2$ ; sampling step size for  $\theta_i = 0.1$

As seen before, Figure 5.12 displays a *useful* preference  $p = d_1 \succeq d_2$  that forms a meaningful restriction on the weight values for the given weighting variables. In this case, we are interested in the red area of the plane with solid lines, i.e., where  $util_\omega^p \geq 0$  holds.

An example for a *useless* preference  $p = d_1 \succeq d_2$  can be easily shown by setting  $d_1(0.7, 0.7)$ . In this case, the evaluation of  $q_\theta = r_1 \wedge_{\theta_1, \theta_2} r_2$  will always yield a bigger score value for  $d_1$  than for  $d_2$  if  $\theta_i > 0$  (see Definition 4.17) because  $\theta_i = 0$  means that the affected attribute will be effectively “disabled” (see Section 4.3).

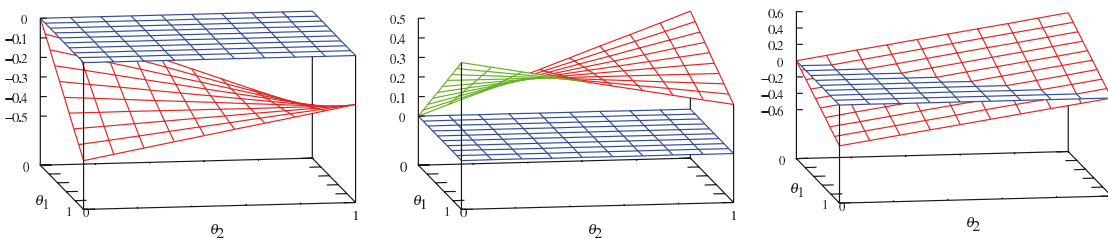


Figure 5.13: Cases of utility of the preference  $d_1 \succ d_2$  with respect to the zero level [Zellhöfer & Schmitt 2010b, Fig. 5]

Analogously, an *inconsistent* preference  $p = d_1 \succeq d_2$  can be created with  $d_1(0.7, 0.3)$  and  $d_2(0.9, 0.9)$ . This preference cannot be expressed with the CQQL condition  $q_\theta =$

## 5 Machine-based Learning of Personalized CQQL Queries

$r_1 \wedge_{\theta_1, \theta_2} r_2$  because  $d_2$  will always get a higher score value than  $d_1$ . Such *query incompatibilities* are addressed in Section 5.4.7 and become important in Section 5.5.

How the category of a preference can be discovered in general is elaborated in Zellhöfer & Schmitt [2010b, Sec. 4.1], which also presents a categorization algorithm. Therefore, a separate discussion is omitted in this thesis. For the remainder of this section, we assume that we deal only with useful preferences.

As said before, every preference in  $P$  constitutes a restriction of weight values. In other words, a preference defines an area in the hyper-unit cube defined by the weighting variables in which weight values are “valid”, i.e., where Equation 5.2 holds.

Typically, the user will provide multiple preferences that have to be fulfilled at the same time. Given that the preferences are not in conflict, two preferences can basically overlap or imply the one or the other. Figure 5.14 illustrates these relations using two preferences and two weighting variables. The horizontally and vertically striped regions depict the restrictions on the weight values defined by the two preferences.

On the left,  $p_1$  *implies*  $p_2$ , i.e., the restriction area of  $p_2$  (vertically striped) is fully contained in the weight value restriction defined by  $p_1$ . Thus, the removal of  $p_2$  from  $P$  will not change the restriction of the weight values that is described by all other preferences in  $P$ . As a consequence,  $p_2$  does not provide additional information to the weight learning algorithm and can therefore be neglected. In case two preferences overlap (see Figure 5.14; right), the only valid weight values lie in the intersection area. Hence, the removal of any preference is not feasible.

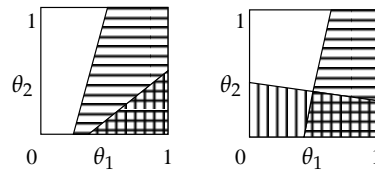


Figure 5.14: Implication (left) and overlap (right) of preferences [Zellhöfer & Schmitt 2010b, Fig. 10]

To conclude, the main objective is to determine the smallest, non-empty intersection region in the weight hyper-unit cube using the minimal number of useful preferences. This is due to two main reasons. First, the size of  $P$  has an impact on the execution time of the learning algorithm as Section 5.4.6 presents. Second, a minimal number of descriptive preferences means potentially less demanding interactions from the user.

### Learning as an Optimization Problem

Acknowledging that every preference ( $p = d_i \succeq d_j$ )  $\in P$  corresponds to a restriction on the weight values that remain assignable<sup>73</sup> by  $\omega$  to  $q_\theta$  and that all  $p \in P$  have to be satisfied, a feasible aggregation into one restriction is needed.

<sup>73</sup>That is, the weight values for  $\theta_i$  that fulfill the preference (see Definition 5.23).

As suggested by Zellhöfer & Schmitt [2010b], the *min* function over all preferences with the additional constraint to be greater or equal than 0 can be used because it guarantees that all formulated preferences are fulfilled, i.e.:

$$\min_{(d_i \succ d_j) \in P} (eval(q_\theta, \omega, d_i) - eval(q_\theta, \omega, d_j)) \geq 0. \quad (5.3)$$

As a result, the finding (or learning) of a weighting scheme  $\omega$  for  $q_\theta$  can be regarded as an optimization problem in which a certain weighting scheme  $\omega^{opt}$  that maximizes the *min* function has to be found. The maximization of *min* improves the significance of the rank produced by the combination of  $q_\theta$  and  $\omega^{opt}$ , i.e., the maximal outcome of the difference between all documents referred to in  $P$ . In other words, the optimization searches for a weighting scheme that clearly discriminates between the score values of all matched documents. It is believed that this equals the intended semantics of the preferences specified by the user.

On a closer examination of Section 4.3, it becomes clear that the optimization problem is non-linear. Roughly speaking, this is caused by the fact that weighting variables can occur within a product during the evaluation of a CQQL condition. Unfortunately, the non-linearity yields a computational hard and complex problem [Bradley et al. 1992].

A commonly used algorithm for solving such problems is the heuristic *downhill simplex* algorithm by Nelder & Mead [1965] that approximates the minimum of a target function. The algorithm and its utilization in PrefCQQL is described in Section 5.4.5. Being heuristic, the downhill simplex cannot guarantee that it will find the exact minimum of the target function. However, Zellhöfer & Schmitt [2010b] argue that it is not necessary to find the exact maximum of the *min* function as long as all user-defined preferences are respected.

To conclude, it is helpful to categorize the learning approach used in PrefCQQL. As hinted in Section 5.4.1, the PrefCQQL learning algorithm is a representative of the learning to rank methods (see Section 5.2) because it tries to find a weighting scheme that produces a rank in conjunction with a CQQL condition which satisfies the user-defined preferences. In contrast to typical L2R approaches, it learns only a special case, namely the user's current IN. It is not meant to draw generalizable conclusions for a large number of documents or IN. Hence, it is appropriate to assume that the algorithm is not dependent on a huge amount of training data. Instead, one can assume that it will work sufficiently with a very small amount of training data, i.e., a supposedly small number of user-specified preferences. The validity of this assumption is investigated in Section 8.5.

#### 5.4.5 Adoption of the Nelder-Mead Downhill Simplex Algorithm

The downhill simplex algorithm by Nelder & Mead [1965] is known to be a simple and robust but not necessarily efficient method for finding the minimum of a non-linear function. Because the discussion on alternative or more efficient algorithms falls out of the scope of this dissertation, refer to Bradley et al. [1992] or Press et al. [2005] that provide an exhaustive overview of the subject of linear and non-linear optimization.

Although the original publication of the downhill simplex method already includes a detailed description of the algorithm, there are many variants with minor modifications. The implication used in this dissertation is derived from a variant implemented in C++ that has been presented in Press et al. [2005, Sec. 10.4]. For the sake of brevity, this section only outlines the rough idea behind the downhill simplex algorithm. For the actual implementation, see the source code<sup>74</sup> attached to this text as a supplement (see Appendix E). An overview of the basic system architecture is available in Section 6.3.1.

### The Nelder-Mead Downhill Simplex Algorithm

The core idea of Nelder and Mead's downhill simplex algorithm is based on Spendley et al. [1962], who suggest to create a simplex in an  $n$ -dimensional parameter space whose defining points are used to evaluate a linear function's value. A simplex is the geometrically simplest volume in a  $n$ -dimensional space with  $n + 1$  points, e.g., a triangle in two dimensions. This simplex is continuously reformed by reflecting one point of the simplex. Consecutively, the target function is evaluated again at each point and the simplex is redefined until the optimum of the linear function is found. In contrast to the contribution by Spendley et al. [1962], Nelder & Mead [1965] do not rely on various parameters that have to be known *a priori* in order to control the simplex' movement. Instead, their algorithm incorporates an automatic adaption of the simplex' movement depending on the evaluation of the non-linear function's value at each vertex of the simplex.

The eponymous simplex of the Nelder-Mead algorithm is defined by  $n + 1$  points called  $P_i$ . Let  $y_i$  be the function value at  $P_i$  and  $h$  (highest) the suffix of  $y_i$ , where  $y_h = \max(y_i)$  and  $l$  (lowest) such that  $y_l = \min(y_i)$ .  $\bar{P}$  is the centroid of the points with  $i \neq h$ . Initially, the points  $P_i$  are determined by taking one randomized  $P_0 \in P$  and deriving the other points as follows:  $P_i = P_0 + \lambda_i$ , with  $\lambda_i$  being specific offsets to define the simplex' other vertices.

In order to find the minimum of the target function, the downhill simplex algorithm relies on three strategies illustrated in Figure 5.15 to move the simplex through the parameter space. First, the algorithm *reflects*  $P_h$  through the opposite face of the simplex (see Figure 5.15; left). If it detects a lower function value at the reflected point  $P^*$ , the algorithm further *expands* the simplex into this direction (see Figure 5.15; center). If the function's value increases into this direction, the simplex *contracts* itself in the transverse direction (see Figure 5.15; right) [Press et al. 2005, cf. p. 414]. The flow diagram of the Nelder-Mead algorithm is available in Figure 5.16.

As shown in the flow diagram, the movement (or redefinition) of the simplex is repeated until the minimum of the target function is found. Typically, the discovery of the minimum is considered to be achieved when the decrease of the target function's value underruns a given tolerance value [Nelder & Mead 1965, cf. p. 309]. Another widely used termination criterion is the number of maximal evaluations of the target

<sup>74</sup>See namespace `dbis::weightlearning`, and `dbis::weightlearning::CQQLWeightLearning` in particular.

## 5.4 PrefCQQL – Preferences within the CQQL-based Retrieval Model

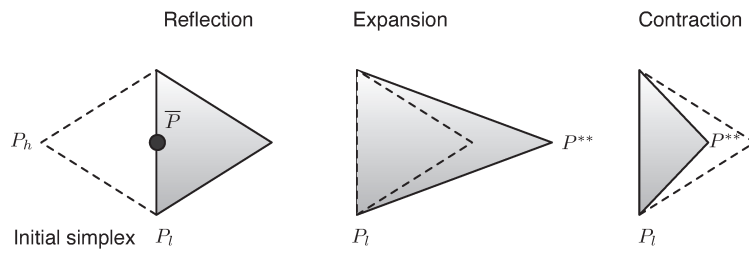


Figure 5.15: Simplex redefinition strategies

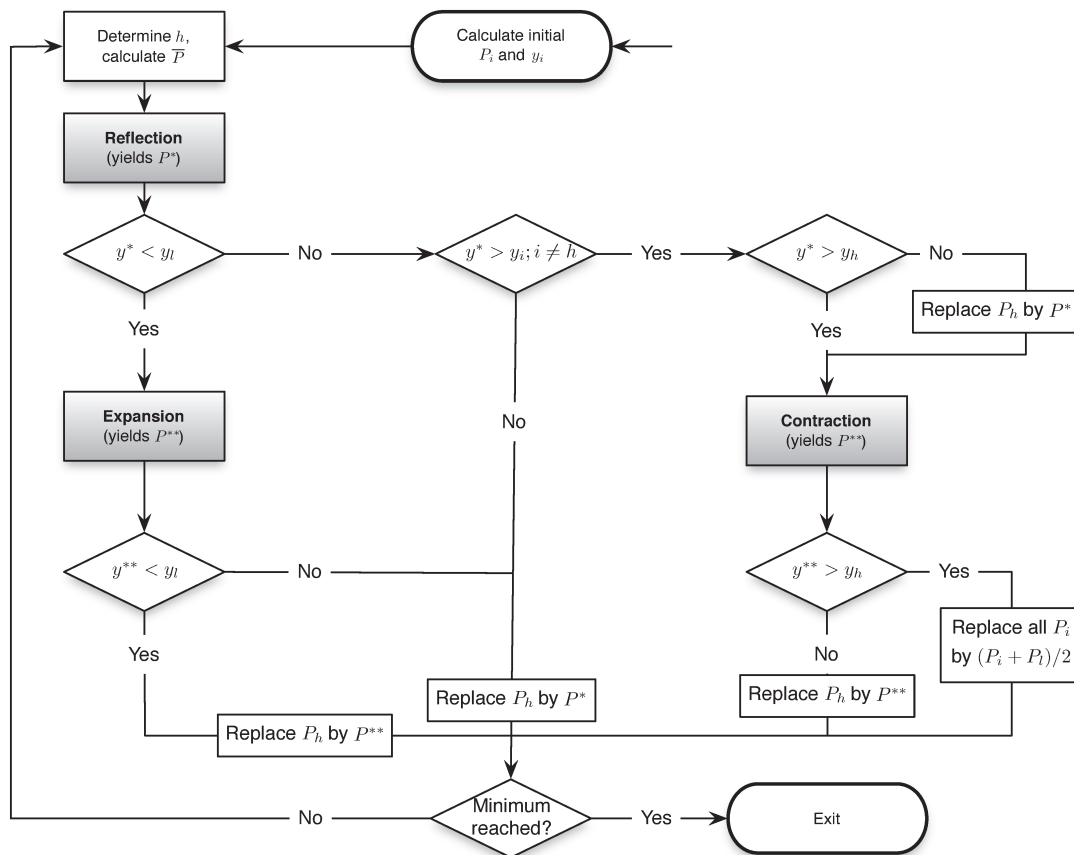


Figure 5.16: Simplified flow diagram of the downhill simplex algorithm [Nelder & Mead 1965, cf. Fig. 1]

function  $kNMax$ .

A general risk of the downhill simplex method is that it might converge to a local minimum instead of the global minimum [Nelder & Mead 1965, cf. p. 311]. In order to

address this issue, the simplex can be started sequentially multiple times with different randomly chosen starting points. This method to increase the approximation quality of the downhill simplex is also used in PrefCQQL [Zellhöfer & Schmitt 2010b]. Another alternative is to restart the algorithm for a number of times whenever it claims to have found a minimum [Press et al. 2005].

### Modifications to the Nelder-Mead Downhill Simplex Algorithm

The original downhill simplex method assumes a target function without constraints [Nelder & Mead 1965, cf. p. 308]. As said before, the preferences in PrefCQQL constitute a constraint on the potential target function (see Equation 5.3), namely a restriction on the valid weight values. Moreover, the downhill simplex algorithm is designed to find the minimum of a function, while we are interested in a maximization of Equation 5.3.

In the case of PrefCQQL, the  $n$ -dimensional parameter space is defined by the number of weighting variables in  $q_\theta$ . The simplex is defined by its points  $P_i$ , each corresponding to a weighting scheme  $\omega_i$ . In order to determine the function value  $y_i$  at  $P_i$ , we have to find a way of including the preference constraints in the function's evaluation. One way is to calculate the sum of all preferences' utilities and to use this value for  $y_i$ . The corresponding function is shown in Listing 5.2 in pseudo-code<sup>75</sup>.

Listing 5.2: Preference utility calculation using the SumMin strategy

```

1 function applySumMin( $\omega_i$ )
2 {
3   min =  $\infty$ ;
4   sum = 0;
5   // Calculate sum of the preference utilities
6   for (i = 0; i < numberOfPreferences; i++) {
7     // each preference has the form:
8     // greaterElement  $\succeq$  smallerElement
9     greater = eval( $q_\theta, \omega_i, \text{preferences}[i].\text{greaterElement}()$ );
10    smaller = eval( $q_\theta, \omega_i, \text{preferences}[i].\text{smallerElement}()$ );
11    preferenceUtility = greater - smaller;
12    if (preferenceUtility < min) {
13      min = preferenceUtility;
14    }
15    sum += preferenceUtility;
16  }
17  if (min <= 0) {
18    return min;
19  }
20  return sum;
21 }
```

<sup>75</sup>For the full implementation in C++, refer to `dbis::weightlearning::CQQLWeightLearning::applySumMin()`.

Because the Nelder-Mead algorithm tries to find a function's minimum, the negated return value of this function is used to calculate  $y_i$ <sup>76</sup>. Further constraints such as Equation 5.3 being greater than or equal than zero are tested subsequently<sup>77</sup>. The same holds true for the maximization.

An understanding of the code at lines 17 - 19 is not crucial for the comprehension of the basic algorithm. The code ensures that the function also returns a value in case preferences cannot be fulfilled, which has been proven practical in the pre-studies to the development of the learning algorithm [Zellhöfer & Schmitt 2010b]. Roughly speaking, it enables the learning algorithm to pick the weighting scheme which still performs best even if not all preferences could be satisfied. In other words, if one preference cannot be fulfilled, the simplex will withdraw from the respective point, i.e., movements into the direction of points that fulfill all preferences are reinforced.

### 5.4.6 Runtime of the Learning Algorithm

Section 5.3.3 already brought up the computational complexity of qualitative and quantitative preference approaches. Not surprisingly, the actual complexity of the PrefCQQL approach depends mainly on the downhill simplex algorithm. Unfortunately, the complexity of the Nelder-Mead algorithm is typically not given in the literature because it is highly instance-dependent. Being a heuristic search method, it depends on many parameters such as the function to be optimized, the number of variables etc..

Because of the importance of the runtime of the algorithm during interactive IR, Zellhöfer & Schmitt [2010b] examine the typical execution time of the learning algorithm used by PrefCQQL with respect to the main parameters affecting the runtime: the number of preferences and the number of weights present in an imaginary function to be optimized.

The experiments have been conducted on an Apple MacBook Pro notebook (version 4,1) from 2008 with one Dual-Core Intel Core 2 Duo with 2.5 GHz and 4 GB RAM running Mac OS X 10.5.5 and Java 5 [Zellhöfer & Schmitt 2010b]. Figure 5.17 shows the results of the runtime analysis.

Regarding the applicability of the Nelder-Mead algorithm in an interactive retrieval scenario, it is noteworthy that the runtime for the algorithm for ca. 10 weights and ca. 40 preferences stays below one second<sup>78</sup>. A further increase of both parameters yields a reasonable runtime extension. In any case, it can be doubted whether users are willing to provide a high amount of preferences because users tend to avoid massive interactions [Shneiderman & Plaisant 2005].

Another factor that has a major impact on the runtime of the downhill simplex algorithm is its implementation. The algorithm is well-suited for parallelization which has not been used in the aforementioned experiment. The main application area of parallelization is the simultaneous launch of the simplex with different starting points. Using parallelization, the runtime of the learning algorithm becomes more and more

<sup>76</sup>See `dbis::weightlearning::DownhillSimplex::maximize()`.

<sup>77</sup>See `dbis::weightlearning::CQQLWeightLearning::findBestWeights()`.

<sup>78</sup>300 different simplex starts and  $kNMax = 1,000,000$

## 5 Machine-based Learning of Personalized CQQL Queries

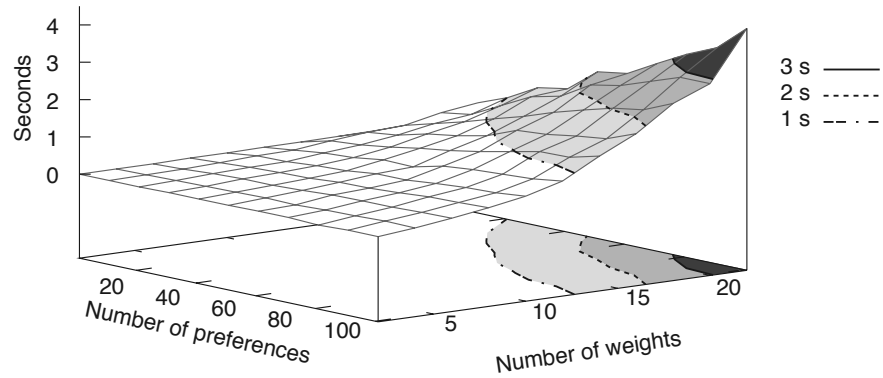


Figure 5.17: Runtime of the learning algorithm in seconds depending on the amount of weights and preference pairs [Zellhöfer & Schmitt 2010b, cf. Fig. 15]

negligible because many more factors, including clustering and rendering, affect the responsiveness of an interactive MIR system. To better illustrate the responsiveness during learning, a real-time sample user interaction video of the prototypical Pythia MIR system (see Chapter 6) is available in the supplement to this thesis (see Appendix E).

### 5.4.7 Preference Conflicts and Query Incompatibilities

The earlier sections of this text assumed the set of preferences  $P$  as a conflict-free and consistent, i.e., compatible with a given query. Furthermore, it has been shown that valid preferences in PrefCQQL form weak posets that correspond to directed acyclic graphs (DAG).

These assumptions are required by the weight learning algorithm presented earlier. In order to provide only the minimal set of preferences as learning input, the preference graph can be reduced by reflexivity and transitivity. A set of preferences  $P$  for  $q_\theta$  is *conflict-free* “if its transitive and reflexive closure yields a partially ordered set (poset). That is, cycles such as  $o_1 \geq o_2$  and  $o_2 \geq o_1$  with  $o_1 \neq o_2$  are not allowed”<sup>79</sup> [Zellhöfer & Schmitt 2010b, p. 38].

#### Preference Conflicts

Unfortunately, users can violate these constraints while interacting with the preference graph, e.g., by adding new preferences. A simple conflict has already been outlined in Section 5.4.1. Given a set of preferences  $P = \{a \succ b \succ c; d \succ c\}$ , the user adds a new preference  $p_4 = c \succ a$  because of either a change in the current state of knowledge or an erroneous input. By adding  $p_4$  to  $P$ , the user creates a cycle in the preference graph that prevents its usage in the weight learning algorithm because it constitutes

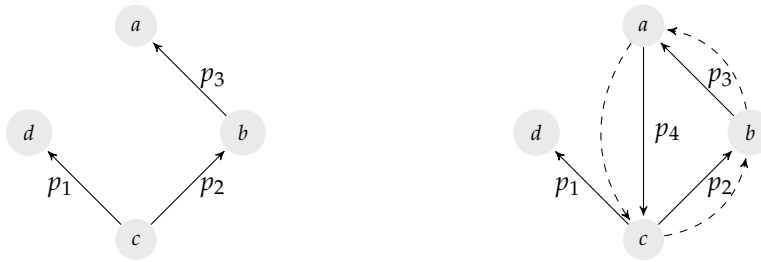
<sup>79</sup>The notation of the original article has been maintained so that  $o_i$  can be considered synonymic with  $d_i$ .



a contradicting preference. Figure 5.18 illustrates the modification of the preference graph and the resulting “contradictory” cycle.

In order to resolve the conflict, the user has to be asked which of the preferences  $p_2$ ,  $p_3$ , or  $p_4$  is invalid and to remove this preference from the graph. If the user cannot decide for the removal of a preference, it is most likely that alternatives are incommensurable as discussed in Section 5.4.1. As stated before, such conflicts are irresolvable as per definition.

This example points out the necessity to find cycles in a graph to reveal conflicts. To achieve this, a property of the topological sorting [Knuth 2013] algorithm can be exploited in a brute force (or very naive way): the algorithm aborts when it detects the first cycle. In order to provide direct user feedback, this algorithm has to check the preference graph after each user interaction to inform the user directly about input conflicts. For more complex graphs or if the topological sorting cannot be carried out in parallel to every user interaction, a depth-first search algorithm can be used (which is by the way far more elegant as it will detect all cycles in a graph and allows deeper insights into a graph’s topology). The depth-first search algorithm’s worst-case complexity is  $\mathcal{O}(n)$ , with  $n$  being the number of edges of the graph.



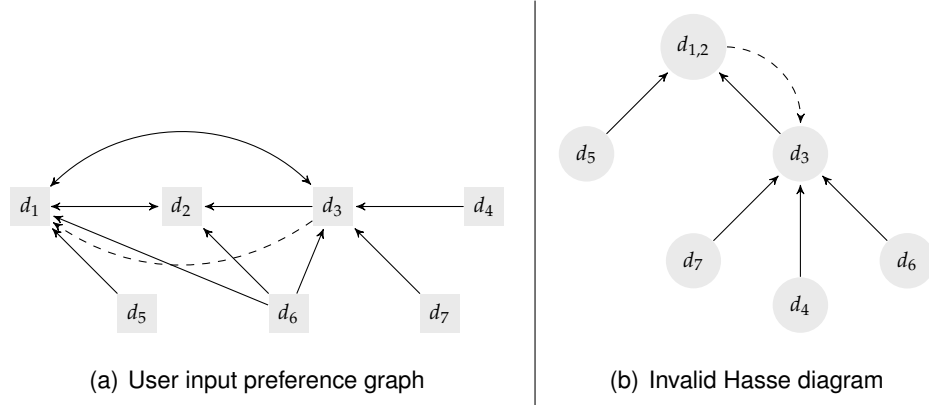
Arrows point to the greater element, dashed lines indicate a cycle.

Figure 5.18: Preference conflict caused by the addition of the preference  $c \succ a$

When dealing with strict preferences alone, each cycle in the user-specified preference graph indicates an automatically irresolvable conflict and each bi-directional edge is forbidden because of the asymmetry requirement. In the case of a user interface that discriminates between strict preferences and explicit indifference judgments as described in Section 5.4.3, the situation becomes more complex.

Although the usage of weak preferences relieves users from the necessity to bring their preferences into a strict order, the application of the  $\succeq$  operator alone to state preferences is not intuitive for all users, especially the mathematically inexperienced ones. Furthermore, the  $\succeq$  operator can hardly express all possible user preferences that have been presented at the beginning of Section 5.3 semantically appropriate. For instance, there might be a clear (i.e., strict) preference between two alternatives and a clear indifference between two other. The application of the  $\succeq$  operator alone decreases the system’s knowledge of the user’s explicit preferences and indifferences that can no longer be exploited or respected by the learning algorithm.

## 5 Machine-based Learning of Personalized CQQL Queries



Bi-directed edges indicate indifference, dashed lines indicate a cycle.

Figure 5.19: Conflicting preference poset; arrows point to the preferred element

Figure 5.19 displays a sample preference that has been input by a user via the user interface expressing the user's preferences between actual documents. By stating  $d_1 \sim d_2$  and  $d_1 \sim d_3$ , the user formulates a cycle shown with a dashed line. The remaining arrows indicate strict preferences. The reflexive and transitive reduction of the graph yielding an invalid Hasse diagram underlines the cycle (see Figure 5.19 (b))<sup>80</sup>.

One way to resolve this conflict is to automatically map  $d_1 \sim d_3$  into  $d_1 \succeq d_3$  as depicted in Figure 5.20. In this case, the transitive reduction of the preference graph results in a Hasse diagram.

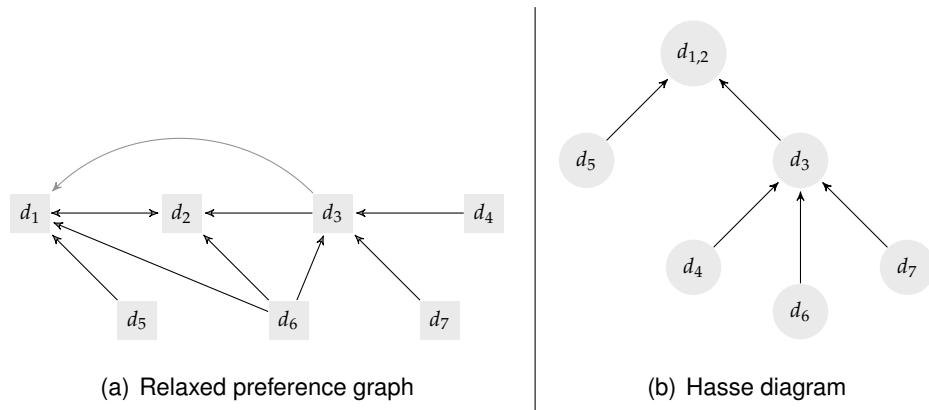


Figure 5.20: Resolved preference poset; arrows point to the preferred element, bi-directed edges indicate indifference

Such a specific treatment of preferences is not provided by the original weight learning algorithm presented in Section 5.4.5 that interprets all preferences as  $\succeq$ . However, this functionality can easily be added.

<sup>80</sup>Because of the equal value of  $d_1$  and  $d_2$  to the user, the documents are combined into one node.

Consider the calculation of the preference utility in Listing 5.2 (line 11). Given that a user has stated an indifference preference between two documents,  $eval(q_\theta, \omega_i, d_1) - eval(q_\theta, \omega_i, d_2)$  is expected to be zero. For  $d_1 \succeq d_2$  it has been shown before that  $eval(q_\theta, \omega_i, d_1) - eval(q_\theta, \omega_i, d_2) \geq 0$  must hold. To express this difference, the algorithm can be easily extended to consider a preference’s type, i.e.,  $\succ$  or  $\sim$ , and to penalize violations of a preference, e.g., by subtracting a penalty factor from the sum.

To cope with relaxation (see above), the algorithm might only accept  $eval(q_\theta, \omega_i, d_1) - eval(q_\theta, \omega_i, d_2) \in [\epsilon, \tau]$ , with  $\epsilon$  being a value slightly greater than zero. The other variable  $\tau$  is a specific threshold value until which the difference of the documents’ scores is still considered insignificant by the user. The actual choice of  $\tau$  depends on the query, the documents, and the user’s sensitivity regarding differences. The interaction between these parameters needs further research, e.g., in form of user studies.

An alternative approach to respect the dynamics of the IN during information seeking is to define a “time to live” (TTL) for the preferences. That is, a time span in which a preference has to be fulfilled. After this time span is reached, the preference can be removed. For instance, the ostensive model [Campbell 2000] and its derivatives use an aging profile for RF judgments. In principle, this approach is also compatible with PrefCQQL. For instance, the TTL of a preference can be decreased every time it takes part in an interactive cycle (see Figure 5.4) or whenever it is not explicitly confirmed (see Definition 5.17). However, the inclusion of preference aging requires an adaption of the evaluation algorithm given in Listing 5.2 in order to reward the fulfillment of the most recent stated preferences.

### Query Incompatibilities

Even if a preference graph is free of conflicts, there is still the chance that it cannot be expressed with a given CQQL query (see Definition 5.25). In order to deal with query incompatibilities, one can either suggest to state a new query or to alter the preference graph to remove inconsistent preferences.

Whenever a query incompatibility with a preference is detected, the system can suggest to manually reformulate the query. Alternatively, machine-based learning can be used to infer a valid query from a set of preferences (see Section 5.5).

#### 5.4.8 Reduction of a Rank

In order to establish a continuous dialog between the cognitive and the information space (see Figure 5.4), the system must be given the chance to communicate its internal state of knowledge about the user’s IN to the user in a comprehensible way. To allow this, Zellhöfer & Schmitt [2010b] suggest to reduce the rank  $rank_{q_\theta}(\omega)$  that has been created on basis of the learned weighting scheme  $\omega$  and a CQQL query  $q_\theta$  to a set of *characteristic* preferences  $P'$ . The set of characteristic preferences  $P'$  is a minimal set of preferences that are needed to produce a *k-equivalent* rank to  $rank_{q_\theta}(\omega)$  using the presented learning algorithm  $learnWeights()$ . If the top- $k$  elements of two ranks are equal and have the same order, these ranks are considered *k-equivalent*.

## 5 Machine-based Learning of Personalized CQQL Queries

That is, the subsequent requirements must apply for  $P'$ :

1. The originally created rank  $rank_{q_\theta}(\omega)$  and  $rank_{q_\theta}(learnWeights(P'))$  are  $k$ -equivalent, i.e.,

$$rank_{q_\theta}(\omega) =_k rank_{q_\theta}(learnWeights(P')).$$

2.  $P'$  is minimal, i.e., there are no preferences that can be removed from the set without violating the  $k$ -equivalence:

$$\neg \exists p \in P' : rank_{q_\theta}(learnWeights(P' \setminus \{p\})) =_k rank(\omega).$$

To derive the preferences in  $P'$ , Zellhöfer & Schmitt [2010b] present a simple algorithm with a complexity of  $\mathcal{O}(n^2)$  depending on  $k$ .

### **Reduction algorithm** [Zellhöfer & Schmitt 2010b, cf. Fig. 12]

1. *Based on the first  $k$  objects of the rank, pairwise preferences of neighboring objects  $\mathbf{o}_i \geq \mathbf{o}_{i+1}$  for  $i = 1, \dots, k - 1$  are derived. The resulting preference set  $\mathbf{P}'$  is already reduced by transitivity and reflexivity. Inconsistent preferences and empty intersections of preference regions cannot occur.*
2. *Removal of useless preferences (see Section 5.25).*
3. *Generation of a minimal  $k$ -equivalent preference set  $\mathbf{P}'$ :*
  - a) *For every preference  $\mathbf{p}$  in  $\mathbf{P}'$* 
    - i. *Test if the removal of  $\mathbf{p}$  from  $\mathbf{P}'$  violates the  $k$ -equivalence.*
    - ii. *Add the preference again in case of a violation.*
  - b) *Repeat step a) until no more preferences can be removed without violating the  $k$ -equivalence.*

Please note that the preferences in  $P'$  are not necessarily the same as in the initial preference set  $P$  that has been used to infer  $\omega$ . However, these preferences correspond to the system's model of the user's IN and are therefore a valuable information for the user supporting the user's understanding of the system's reaction.

An essential parameter affecting the practical usability of the reduction algorithm shown above is  $k$ . If  $k$  is set very high,  $P'$  will also contain a large number of preferences that have to be inspected by the user. On the other hand, the inspection and modification of a high number of preferences can also lead to a faster adjustment of the weights to the user needs, lowering the number of RF iterations.

## 5.5 Learning CQQL Queries as a Special Case of Weight Learning

Although PrefCQQL's weight learning algorithm allows the RF-based personalization of results, it is strictly bound to a given CQQL condition that can be provided by the user or the system. However, as Section 5.4.7 has shown, there is a chance that the user might state inconsistent preferences during the RF iterations. For instance, these inconsistent preferences can be due to erroneous input or because the user's IN has changed drastically and cannot be modeled any longer with the help of the utilized CQQL condition.

To address the latter issue, Schmitt & Zellhöfer [2012] propose a query learning algorithm based on the recently discussed PrefCQQL approach. In accordance with the beginning of Chapter 5, this algorithm can be considered as a *query reformulation* technique because it will search for a newly structured logical query expressing the user's preferences. Because the algorithm is of minor interest in this dissertation, only its main idea is outlined here. For a detailed discussion and additional information such as its runtime and complexity, see the original publication by Schmitt & Zellhöfer [2012].

In order to learn a new query, it is useful to consider the search for a query satisfying the user's preferences as an optimization problem with the following characteristics.

Let  $W$  be all solutions<sup>81</sup> fulfilling the specified preferences. To find a new query, we have to obtain the solution  $\omega \in W$  which fulfills the input preferences  $P$  best (see below).

From the laws of logic we know that every Boolean algebra condition (and thus every CQQL condition  $\varphi$ ) can be expressed in a full disjunctive normal form (DNF) with the following form  $\bigvee_i (\bigwedge_j (\neg) \varphi_{ij})$ . Thus, every condition is bi-directly associated to a subset of  $2^n$  minterms<sup>82</sup>. In consequence, exactly  $2^{2^n}$  conditions exist.

The core idea behind the query learning approach is to assign every possible DNF minterm  $mt_i$  a weighting variable  $\theta_i$ :  $\bigvee_i \theta_i mt_i$ .

Hence, a solution of our learning problem is a weighting scheme  $\omega$  that assigns every weight variable  $\theta_i$  a value out of  $[0, 1]$ . Due to the full disjunctive normal form, between any two different minterms at least one predicate is negated in one and not negated in the other minterm. Thus, the evaluation of the disjunction reduces to a simple weighted sum because of the evaluation of a disjunction in CQQL (see Definition 4.14, Equation 4.11b or Schmitt [2008] for a more detailed discussion):

$$eval(\omega, d) = \sum_{i=1}^{2^n} w_{\theta_i} \cdot mt_i(d), \quad (5.4)$$

where  $w_{\theta_i}$  denotes the value of the weighting variable  $\theta_i$  as assigned by  $\omega$  and  $mt_i(d)$  denotes the evaluation of  $i$ -th minterm against a document  $d$ . Further, every feasible

<sup>81</sup>That is, a CQQL condition to be learnt.

<sup>82</sup>A minterm has the form  $(\bigwedge_j (\neg) \varphi_{ij})$ .

## 5 Machine-based Learning of Personalized CQQL Queries

solution must fulfill all user preferences in  $P$ . The best solution is the solution that maximizes the sum of preference differences:

$$\sum_{i=1}^m (eval(\omega, d_{i_1}) - eval(\omega, d_{i_2})) \quad (5.5)$$

Another presentation of the target of the optimization problem is given by the multiplication of four matrices:  $\mathcal{S} \cdot \mathcal{P} \cdot \mathcal{A} \cdot \Theta \implies \max$  where  $\Theta = (\theta_1, \dots, \theta_{2^n})^T$  contains the weighting variables of a solution, and

$$\mathcal{A} = \begin{pmatrix} mt_1(d_1) & \cdots & mt_{2^n}(d_1) \\ \vdots & & \vdots \\ mt_1(d_k) & \cdots & mt_{2^n}(d_k) \end{pmatrix} \quad (5.6)$$

is a  $k \times 2^n$  matrix containing the minterm evaluation of the  $k$  objects involved in the  $m$  preferences.

$\mathcal{P}$  decodes the preferences as differences. That is, a row corresponds to a preference and a column to a minterm. 1 denotes the larger object and  $-1$  the smaller object.  $\mathcal{S} = (1 \dots 1)$  sums up all  $m$  preference differences. Conditions to be respected are  $\mathcal{P} \cdot \mathcal{A} \cdot \Theta \geq 0$  ( $m$  preference conditions) and  $0 \leq \theta_i \leq 1$  ( $2^n$  weight conditions).

### Finding a Solution to a Linear Learning Problem

This presentation yields the following linear learning problem: consider an example with two conditions  $c_1, c_2$ , three documents and two preferences ( $d_1 \geq d_2, d_1 \geq d_3$ ) with  $mt_1(d_i) = c_1(d_i)c_2(d_i), \dots, mt_4(d_i) = (1 - c_1(d_i))(1 - c_2(d_i))$ .

The target function to be maximized is given by:

$$(1 \ 1) \cdot \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} mt_1(d_1) & mt_2(d_1) & mt_3(d_1) & mt_4(d_1) \\ mt_1(d_2) & mt_2(d_2) & mt_3(d_2) & mt_4(d_2) \\ mt_1(d_3) & mt_2(d_3) & mt_3(d_3) & mt_4(d_3) \end{pmatrix} \cdot \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} \quad (5.7)$$

From the matrix presentation we recognize a linear optimization problem, which can be solved by a simplex algorithm. Linear programming algorithms solve the optimization problem (average case) in weakly polynomial time.

Alternatively, we can require strictness:  $\mathcal{P} \cdot \mathcal{A} \cdot \Theta > 0$ , or restrict weight values:  $\omega(\theta_i) \in \{0, 1\} \subseteq \mathbb{R}$ . The last modification leads to a binary integer problem as a special case of a mixed integer problem that yields an unweighted solution (if it exists). One issue with the optimization problem is the exponential number of weighting variables for  $2^n$  minterms making the optimization hard. As said before, the result can contain  $2^n$  minterms.

To cope with this issue, Schmitt & Zellhöfer [2012] suggest to reduce the problem by starting with  $k < n$  predicates  $\{c_i\} \subset \{c_j\}$ , with  $|\{c_i\}| = k$  and  $|\{c_j\}| = n$ . Using the observation that a solution that is based on  $k$  predicates remains as solution after a new

## 5.5 Learning CQQL Queries as a Special Case of Weight Learning

condition is added, we iterate from  $k = 1$  to  $n$  and stop whenever we find at least one solution in one  $k$ -level. However, for every  $k$ -level we have to test all  $\binom{n}{k}$   $k$ -subsets of the original predicate set.

Thus, the total complexity on level  $k$  is:  $costs(m, n, k) = \binom{n}{k} f(2^k, m + 2^k)$ , where  $f(x, y)$  is the cost of applying the simplex algorithm for  $x$  variables (minterms) and  $y$  constraints (preferences and weight limits). Although the cost of  $f(x, y)$  of a simplex optimization is instance-dependent, it can be roughly estimated by conducting experiments. We make an optimistic assumption that a solution for a small  $k$  can be found. If this is not the case, then the iteration would produce more costs than without iteration:

$$\sum_{k=1}^n costs(m, n, k) \geq costs(m, n, n) \quad (5.8)$$

If we examine  $costs(m, n, k)$  carefully, we see a superposition of exponential and binomial elements. It can happen that two  $k$ -values  $k_1 < k_2$  with  $costs(m, n, k_1) > costs(m, n, k_2)$  exist. In such case, the level  $k_1$  should be skipped from iteration.

A solution to the optimization problem is expressed as weight values for minterms in full disjunctive normal form. In case of discrete weight values from  $\{0, 1\}$ , repeatedly merging minterms, which differ in the negation of just one predicate, simplifies the resulting condition.

Schmitt & Zellhöfer [2012] propose an implementation that stops the iteration  $k = 1, \dots, n$  after the first  $k$ -level contains a solution. The possible results would be:

1. the first found solution of that level,
2. the simplest solution after minterm merging, or
3. all solutions of that level.

Returning all solution helps to see how strict the preference conditions are.

As a result from the query learning algorithm, a new query in DNF that expresses the specified preferences (and thus the new IN) is obtained. Unfortunately, queries in DNF are hardly comprehensible – even for expert users. In consequence, it cannot be recommended to directly modify queries in DNF from a usability point of view. Although there is literature about possible simplifications of the DNF in mathematics and theoretical computer science, e.g., by Andon [1966], these approaches have not been studied further because the learning of CQQL queries falls out of the scope of this thesis.

However, this does not affect the general utility of the query learning, which can be embedded into the search process without confronting users with the learned query directly. Analogously, the learned query can serve as a starting point for further exploration.

To conclude, the utilization of this learning algorithm also enables users to start a retrieval session without the formulation of an explicit query. Instead, users can define their preferences between some QBE documents that will then be used as input for the query learning algorithm.

## 5.6 The Relation of the PrefCQQL Approach to Other Personalization Techniques in Information Retrieval

Figure 5.4 shows that PrefCQQL can be considered as a *preference-based relevance feedback technique*. Moreover, Section 5.4.3 outlined that traditional RF can be regarded as a specialization of the general preference approach. As such, PrefCQQL also integrates RF-based techniques such as the ostensive model [Campbell 2000] by allowing only binary RF. Like other RF approaches, PrefCQQL supports negative RF by placing irrelevant documents as the minimal elements in the preference poset. In addition, PrefCQQL can deal with negative feedback at query level. That is, QBE documents can be negated within CQQL to express negative samples, e.g., as suggested by Assfalg et al. [2000a].

The reliance on inductive preferences of PrefCQQL classifies it as a learning to rank technique (see Section 5.4.4) because the user is given a chance to define a rank for the desired documents, which has to be satisfied by a weight setting and a given CQQL query. Referring to Section 5.4.4, the learning step completes the inductive preferences.

When using solely binary RF, traditional RF is carried out without establishing a well-defined rank of the result documents besides defining the most relevant and irrelevant documents.

Regarding consistent user interactions, the usage of inductive preferences is to be favored because it allows the statement of preferences at document level – a technique that is commonly used in (M)IR, e.g., during query formulation time whenever QBE documents are provided or in the case of traditional RF. In contrast to traditional RF that relies on a query, PrefCQQL's preferences can also be used at query formulation time to learn a CQQL condition as Section 5.5 demonstrated. In other words, the general user interaction with the MIR system can be based on preferences alone. This can take place during the query formulation time when a query is learned or during the information seeking process aimed at personalizing the results.

In comparison to quantitative preferences, inductive preferences also reduce the cognitive workload of the user. Without doubt, the statement of preferences on actual document samples is easier to carry out than quantifying one's preferences in the form of weights. Nevertheless, PrefCQQL allows the direct manipulation of weights in a weighted CQQL query making it a *hybrid preference approach*.

The central idea behind PrefCQQL is the constant dialog between the cognitive and the information space as Figure 5.4 illustrates. Hence, the user has continuous access to the learned weights, the characteristic preferences, and the user-defined preferences. While a typical layperson user might only state preferences and let the learning algorithm infer weight values, an expert user is given full access to the weights. This differentiates PrefCQQL from common IR models that typically do “not include the *cognitive aspects* of the retrieval aspects, such as query negotiation or output evaluation” as van Rijsbergen [2009] put it aptly in a keynote talk at the ESSIR. In PrefCQQL, the user can gain insights into the system's model of IN, which can positively affect the predictability of the system's reactions in return. Eventually, this can improve the overall usability of the MIR system.



## 5.6 The Relation of the PrefCQQL Approach to Other Personalization Techniques in Information Retrieval

To conclude, PrefCQQL is juxtaposed to Kießling's list of desiderata for preferences in databases<sup>83</sup>.

1. *"An intuitive semantics:* Preferences must become first class citizens in the modeling process. This demands an intuitive understanding and declarative specification of preferences. A universal preference model should cover non-numerical as well as numerical ranking methods.
2. *A concise mathematical foundation:* This requirement goes without saying, but of course the mathematical foundation must harmonize with the intuitive semantics.
3. *A constructive and extensible preference model:* Complex preferences should be built up inductively from simpler ones using an extensible repertoire of preference constructors.
4. *Conflicts of preferences must not cause a system failure:* Dynamic composition of complex preferences must be supported even in the presence of conflicts. A practical preference model should be able to live with conflicts, not to prohibit them or to fail if they occur.
5. *Declarative preference query languages:* Match-making in the real world means bridging the gap between wishes and reality. This implies the need for a new query model other than the exact match model of declarative database query languages." [Kießling 2002, p. 312]

*First*, as extensively described before, inductive preferences constitute an intuitively comprehensible mechanism for preference elicitation. The usability and semantic clarity of preferences can even be improved if explicit indifferences and strict preferences are combined as presented before (see Section 5.4.7).

*Second*, PrefCQQL is mathematically well-founded. This includes the preference mechanism that relies on weak preferences (see Section 5.3) and the quantum logic-inspired query language CQQL (see Chapter 4).

To address the *third* desideratum, Section 5.4.3 describes in detail how preferences can be combined in PrefCQQL. Unlike Kießling's approach, PrefCQQL relies on a union of preferences that supports in principle an aging of preferences (see Section 5.4.7) to reflect the dynamics of the IN – a point that is not mentioned in the list.

In contrast to Kießling's *fourth* point, PrefCQQL is not able to "live with conflicts" in most cases. Although there are ways to resolve some conflicts (see Section 5.4.7), conflicts and query incompatibilities are considered as valuable indicators for a change of the current IN. As such, they can be exploited to trigger the learning of a new and better fitting CQQL condition, as outlined in Section 5.5.

The last and *fifth* desideratum falls clearly out of the scope of this dissertation and is due to Kießling's focus on databases. In any case, the practical producibility of embedding PrefCQQL's mechanisms into a declarative query language (SQL in this case) has been shown by Lehrack & Schmitt [2010].

---

<sup>83</sup>This list is included because the first four points are also reasonable in (M)IR.

## 5 Machine-based Learning of Personalized CQQL Queries

## **Part III**

# **Concept and Implementation**



## 6 Design of the Pythia MIR System Prototype

To further investigate the utility of CQQL as an implementation of the principle of poly-representation (see Section 3.2) in the domain of multimedia information retrieval and the effectiveness of the proposed preference approach PrefCQQL, this chapter focuses on the design of a prototypical implementation of the aforementioned principles: the Pythia<sup>84</sup> MIR system prototype.

The development of a prototypical software system is a complex endeavor, which incorporates various tasks<sup>85</sup> that fall out of the research scope of this thesis. For the sake of brevity, this chapter mainly discusses the conceptual model (see Section 6.2) and the agile, user-centered development process of the Pythia MIR system and its interactive components (see Section 6.1). Chapter 9 reports on the usability of the developed prototype.

The actual implementation of the Pythia MIR system prototype at programming or system level, the feature extraction, and the matching (see Section 2.3) are discussed only at a high level of abstraction in Section 6.3. Detailed information at a lower level, e.g., the code documentation or class diagrams, is available as a supplement to this work (see Appendix E).

In order to control the design complexity of the system, the Pythia MIR system is limited to visual perceivable media such as images and (textual) metadata. In particular, the inclusion of the auditive modality, e.g., in form of music, would require the design of additional human-computer interaction input/output channels and mechanisms<sup>86</sup>. Furthermore, supporting this type of media would complicate the technical implementation because more document representations that fall out of the research focus of this dissertation would be needed. To make matters worse, these representations are not necessarily comparable to each other. For instance, it is hardly possible to compare the tempo of a tune to the color of an image. Although one can envision the retrieval of a speech recording on the basis of a textual query or the retrieval of concert recordings of a particular composer based on the provision of their picture, these scenarios are hardly realizable with state of the art MIR systems. Because of these practical reasons, the Pythia MIR system is limited to modalities that can help in a multimodal

---

<sup>84</sup>Named after Pythia, the priestess at the oracle of Delphi in ancient Greece.

<sup>85</sup>For instance, the planning, implementation, and testing of the discussed system and its preceding experimental prototypes took approximately 155 person months.

<sup>86</sup>We acknowledge that audio data can also be visualized, e.g., by plotting its wave form. However, this visualization is not comprehensible for untrained users and is therefore neglected.

CBIR scenario.

Although there are more visual representable media types, e.g., videos, animations, or 3D models, these media types are not addressed separately in this thesis. In any case, the principles presented in this chapter are transferrable to these media to a great extent because they can be visualized in an easily comprehensible manner.

### 6.1 The User-centered Design Principle

It is undeniable that users should be integrated from early on into the development process of a usable computer system. The active integration of users helps to discover underspecified requirements, reveals expectations about the system's functionality and principles of use, and helps to discover usability issues at an early stage of development. In consequence, these issues can be addressed and solved early in order to produce a usable system.

Development processes that follow these principles in the fields of software engineering and interaction design are commonly subsumed under the category of user-centered design (UCD) processes. In contrast to other design techniques in computer science, e.g., the entity-relationship model, UCD approaches usually refer more to a recommendation of similar methods or principles than to a formal methodology [Gould & Lewis 1987]. Roughly speaking, their common principles can be summarized as follows:

1. The establishment of a focus on users and their tasks from early development on,
2. The early reflection on user reactions and performance measurements, and
3. The use of an iterative design process to integrate the obtained findings.

These principles are comparable to the objectives of IIR described in Chapter 3. This relation on a conceptual level has also been discovered by other authors, e.g., White [2004a] or Karlgren et al. [2011].

In extension to the aforementioned works, Zellhöfer [2012f] argues for an integration of UCD techniques into the IIR evaluation because of the similar objectives of user simulations (see Section 8.5.1) and personas. Personas have been originally proposed by Cooper [1999] to assist user interaction designers and software developers to model and understand work tasks and goals of potential user groups. As such, they are still a state of the art technique in interaction design and the related fields, e.g., usability engineering [Dahm 2006; Cooper et al. 2007; Kühn et al. 2011]. Although they are a well-known tool in the UCD process, it is important to note that personas "are not real people, but they are based on the behaviors and motivations of real people we have observed and represent them throughout the design process. They are *composite archetypes* based on behavioral data gathered from many actual users encountered in ethnographic interviews" [Cooper et al. 2007, pp. 75]. In other words, a persona is an artifact that acts as a surrogate for real users from the early development on. This does not mean that they are completely decoupled from real-world users. Instead, Cooper

## 6.1 The User-centered Design Principle

et al. [2007] recommend to enrich the data basis upon which personas are modeled by surveys, research literature, or user interviews to gain a deep understanding of the user needs.

Two sample personas in the scope of MIR are a professional news researcher that uses specific search strategies in a news video archive and a teenager that looks for recreational video clips on the Internet. Obviously, both user archetypes have different motivations and use different techniques to reach their goals. Hence, a usable system will have to address different user needs expressed by the corresponding personas.

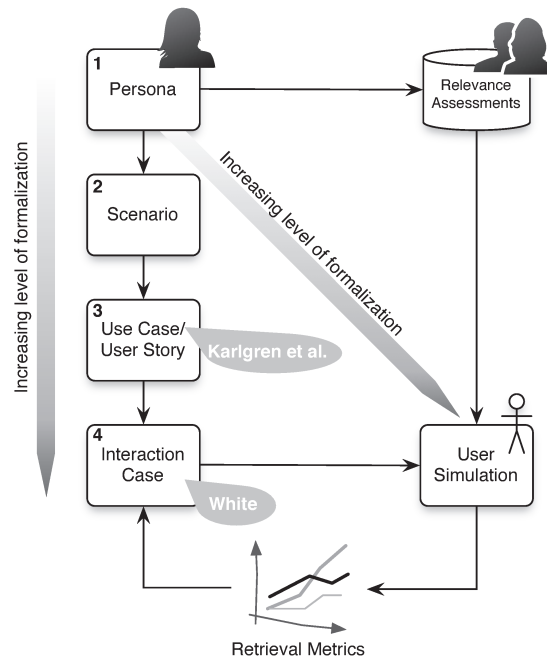


Figure 6.1: Persona-based user-centered system design in relation to user simulations; research foci of White [2004a] and Karlgrén et al. [2011] are indicated [Zellhöfer 2012f, cf. Fig. 1]

The left side of Figure 6.1 illustrates the UCD process used during the agile development of the Pythia MIR system. To some extent, the right part falls out of the scope of this dissertation although Section 8.3.4 contains a conceptual outline of the integration of user simulations and personas into the design and evaluation of IIR systems. Section 8.5.1 presents a simple user simulation used in testing the retrieval effectiveness of the PrefCQQL approach. Generally speaking, UCD can be subdivided into four steps that are discussed in more detail in the next sections:

1. The definition of *personas* is important to obtain clarity about prospected user groups for the MIR system and their work tasks.
2. Based on these personas, *scenarios* can be developed. Scenarios consist of informal work task and context descriptions in a narrative form [Carroll 2000] and can

## 6 Design of the Pythia MIR System Prototype

be extended with storyboards [Landay & Myers 1996]. They should be comprehensible to the stakeholders (i.e., the actual group of potential users) of the MIR system in development in order to discuss them with the potential users.

3. Based on the scenarios, *use cases* that mainly focus on the user interactions with the systems are derived [Jacobson et al. 1992]. Use cases are directly linked to a persona in a way that they describe typical interaction patterns, e.g., the search strategies used to solve a work task. In agile software development, a shorter, more informal form – the *user story* – is used, which is rewritten whenever needed, e.g., if a user is not satisfied with its implementation in the system.
4. In order to specify use cases or user stories, *interaction cases* are used [Schlegel & Raschke 2010]. Unlike use cases or user stories, interaction cases are defined at a fine level of granularity that allows a direct implementation.

### 6.1.1 Personas and Scenarios

The definition of personas in a research project is a challenging task. In contrast to the development of commercial software, research projects are often preceded only with limited market research in order to discover potential user groups on which personas can be based. This is often due to three reasons:

1. the unknown outcome of the research project,<sup>87</sup>
2. the high amount of financial and temporal resources needed for such preparatory studies, and
3. the limited number of available staff.

Nevertheless, potential users have to be kept in mind if a usable MIR system has to be developed.

As no behavioral data for the creation of personas could be gathered in advance because of temporal and financial constraints, the personas used within the UCD process are based on expert interviews and experience. Such “roughly” modeled personas are known as *provisional* or *ad hoc personas* and have been shown to be very helpful during the UCD process [Cooper et al. 2007, cf. pp. 86]. However, these personas should be refined on the basis of additional qualitative data such as interviews, think-aloud protocols, or user observations from real users in their typical context. Furthermore, Cooper et al. [2007] recommended to consult the relevant literature in order include user study results or the like into the persona creation process. This persona refinement is especially important when the prototypical stage of the software development is left.

---

<sup>87</sup>For instance, the research hypothesis could be falsified rendering the whole research pointless or shifting its application area. These dynamics increase the number of risks already present in a normal software development project such as underspecified requirements or lacking funds due to a prolonged development time.



## 6.1 The User-centered Design Principle

In order to receive a manifold feedback on which personas could be based, two half-day workshops with domain experts – mainly from the management level of marketing research, the media industry, and Academia – were organized. The first workshop was held in February 2010 and was visited by 12 participants. In the second workshop in March 2010, 13 participants took part. The minutes of the workshops are available in the supplement to this thesis (see Appendix E). The realization of such workshops is also recommended by others, e.g., Lew et al. [2006] note that “by having closer communication with private industry, we can potentially find out what parts of their systems need additional improvements to increase user satisfaction” [Lew et al. 2006, p. 12].

Both workshops followed the same pattern. After a brief introduction of the theoretical concepts behind multimedia retrieval and CQQL, an early prototype of the Pythia MIR system was shown (see Figure 6.2). The core functionality of this prototype consisted of a QBE image retrieval engine and the preference-based relevance feedback approach PrefCQQL described in Chapter 5. The demonstration was followed by a discussion in form of a stakeholder interview [Cooper et al. 2007, cf. pp. 53] in order to reveal potential usage scenarios of the software, possible user groups (or personas), and recommendable improvements to fit the user needs.

Because the funding of the workshops was made possible by a BMBF<sup>88</sup> supported project, further aspects such as unique selling points of the system or economic feasibility studies were discussed as well. These aspects are not addressed in this work, because they fall out of its research scope.

The open discussion form of the workshops made it hard to differentiate between personas, scenarios, and user stories because the participants used a combination of all three to describe prospective usage domains for the Pythia MIR system. To recapitulate the discussion, the participants of the workshops outlined two basic personas: a layperson or hobbyist user and a media asset-handling professional user. To facilitate reading, the next subsections present the scenarios separated by their corresponding persona. Section 6.1.2 presents a brief summary of the central user stories.

### Usage scenarios for the professional persona

The primary usage scenario of the professional user persona was consistently described in both workshops. This persona works in the field of media or marketing research and searches digital asset libraries on a daily basis. These libraries consist of business-relevant documents such as textual, image, or audio documents. These documents have to be retrieved whenever a specific document is needed, e.g., if a brochure has to be designed. In both workshops, the domain experts emphasized the need for an unsharp or best match retrieval functionality to avoid empty result sets. Furthermore, they underlined that filters should also be non-Boolean because of the dynamics and the vagueness of the information need.

From a marketing research perspective, the experts stressed the importance of searching for the most dissimilar image documents. In this field, the search for similar doc-

---

<sup>88</sup>German Federal Ministry of Education and Research

## 6 Design of the Pythia MIR System Prototype

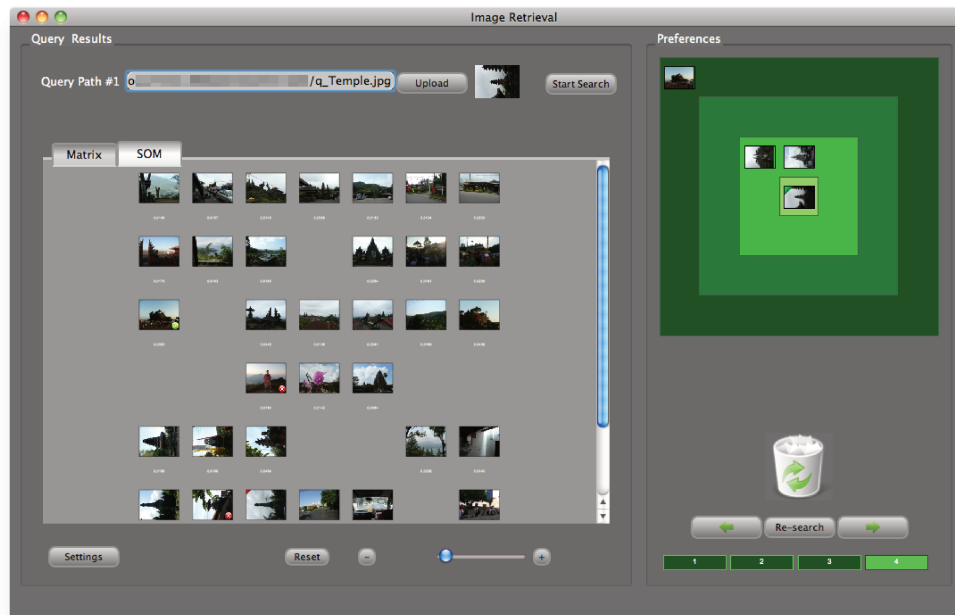


Figure 6.2: Early GUI for the Pythia MIR system [Zellhöfer & Schmitt 2011a, Fig. 1]

uments is not enough to fit the needs of a marketing researcher looking for new developing trends, e.g., in a collection of photographs of different hand bags – or in one participant’s own words: “[...] the ‘weak signal’, the absolute new, not widespread, something that establishes a trend”<sup>89</sup>.

### Usage scenarios for the layperson persona

In contrast to the professional persona, the usage scenarios for the layperson were more versatile. The workshop participants noted that a private user is likely to use the tool for various tasks. A private or hobbyist user is expected to use the software to sort holiday photographs or to satisfy an information need. For instance, one motivation could be to retrieve more information about a tree by taking a photograph of a tree’s leaf. Additionally, the participants described other usage areas such as a typical search for similar images or a functionality comparable to Google’s Goggles<sup>90</sup>. Hence, the MIR system is expected to adapt to the current dynamic user needs.

This application scope differs from the professional persona that has a relatively specific usage domain, for which the system could be optimized.

Complementing the persona’s personal usage, the participants described an assisted product search in an online shop. The search for a product could be started by an

<sup>89</sup>Translation from the German original comment (workshop in February 2010): “[...] das ‘schwache Signal’, also das absolut Neue, noch nicht Verbreitete, das einen neuen Trend ausmacht.”

<sup>90</sup><http://www.google.com/mobile/goggles>

uploaded image and refined by using a self-organizing map or the like (see Section 3.3.2) incorporating visual features and metadata.

### Usage scenarios for the researcher persona

Because of the research focus of this dissertation, a third persona with a research background has to be defined. Unlike the other personas, this persona has no typical end-user related work task. Instead, it uses the aforementioned scenarios to obtain data for the analysis of the retrieval subsystem. This analytical data can be used to fine-tune the utilized algorithms in order to improve the user experience for the two other personas. In consequence, this persona needs means to affect specific parameters of the retrieval engine during a search. Obviously, more knowledge of the system's architecture is required from this persona than the two other persona will need.

The necessity of such a persona was also noted casually in the two workshops and motivated by the need to integrate the described MIR functionality into an existent IT infrastructure consisting of various applications and databases.

### Other requirements

In addition to the persona-related scenarios, the participants of the workshops specified general functional and non-functional requirements for the Pythia MIR system based on the demonstrations of the early prototypes.

Throughout both workshops, five participants explicitly underlined the importance of the adaptivity of the matching functions, e.g., in form of preferences or weights. One participant summarized the general need for an adaptive but controllable system as follows: "Thus, it shall [...] not [be] an autonomous system, but a machine learning from the interaction of the respective user that can be adapted by the user [...]"<sup>91</sup>. This statement includes another user need, the demand for an explicit calibration of MIR system's search parameters depending on the current work task. Furthermore, the participants saw a risk for the user acceptance if the initial search parameters were too "standardized" and could not be adjusted quickly with respect to the user needs.

As mentioned before, three participants stressed the need for a dissimilarity search functionality or at least a supportive visualization of dissimilar documents.

Interestingly, the experts did not expect perfect results from a MIR system. Instead, they asked only for a relatively good result quality. Moreover, they would accept a refinement of the results in the sense of relevance feedback if this interaction would leave the user in control of the action.

With regard to the retrieval process, the participants stressed the importance to include all available representations of a document, metadata in particular. Metadata such as the camera model is expected to reveal usage patterns which are relevant for the current information need.

---

<sup>91</sup>Translation from the German original comment (workshop in February 2010): "Daher dürfe es sich [...] nicht um ein autonomes System handeln, sondern um eine aus der Interaktion mit dem jeweiligen Nutzer lernende Maschine, die sich durch den User anpassen lässt [...]"

## 6 Design of the Pythia MIR System Prototype

Additionally, the workshop participants asked for some kind of fuzzy filter functionality to alter the view on a retrieved result set. Two participants explicitly demanded fuzzy filters because of the often unclear information need and the fear of empty result sets when classical Boolean filters were used. Generally speaking, the domain experts asked to avoid empty results at all costs.

Regarding the retrieved documents, two participants emphasized that the result should contain a variety of documents. In other words, it should be avoided that the results become very homogeneous, i.e., if images of almost the same motif would be displayed at the first 30 ranks of the result list. Further, “resultless” queries<sup>92</sup>, i.e., queries which lead to almost the same results in every relevance feedback iteration, should be minimized.

Not surprisingly, a usable GUI was postulated. Although the participants saw a potential usage area for the MIR system in the professional field, they recommended to adjust the GUI towards the layperson or standard user needs as “the control and the training of the algorithms must function intuitively”<sup>93</sup>. From this perspective, the domain experts appreciated the presence of a drag and drop mechanism during preference elicitation.

### 6.1.2 User Stories

Because of the agile software development process, all user stories underwent constant change, e.g., caused by the iterative user feedback (see Section 6.1.3). For the sake of simplicity, only the finalized<sup>94</sup> central user stories are thematically grouped and listed at this section. These user stories form the basis for the development of the conceptual model of the Pythia MIR discussed separately in Section 6.2.

A separate discussion of the related interaction cases is omitted because of their closeness to the implementation aspects of the system. When adequate, they are discussed along with the different functionalities of the MIR system in Section 6.2.

To improve the readability of the text, the user stories are grouped by their related persona, whereas generally applicable user stories are presented separately. All user stories are structured alike following the often used actor and action form [Cohn 2008, Ch. 2]. They are written from the perspective of a persona and contain the desired action, e.g.: “A bank customer can pay with a credit card”.

Further information about user stories and their creation is available, amongst others, by Cohn [2008].

### General User Stories

#### Search Strategy

**User Story 1 Fast and intuitive adaptivity:** *A user can adapt the underlying matching function intuitively and quickly.*

<sup>92</sup>Translation from the German original comment (workshop in March 2010): “ergebnislose Suchen”

<sup>93</sup>Translation from the German original comment (workshop in March 2010): “[...], dass die Bedienung sowie das Training der Algorithmen intuitiv funktionieren müssen.”

<sup>94</sup>That is, their state at the time of the writing of this thesis.

## 6.1 The User-centered Design Principle

**User Story 2 Controllable relevance feedback:** *A user can refine the results using controllable relevance feedback.*

**User Story 3 Fuzzy filters:** *A user can apply a fuzzy filter on the results that is based on image or metadata features.*

### Result Presentation

**User Story 4 Get an overview:** *A user can get an overview of the contents of the document collection.*

**User Story 5 Diversity of the results:** *A user can inspect the results in a way that a variety of motifs is displayed.*

### Retrieval Model and Language

**User Story 6 Avoid empty result sets:** *A user must be able to retrieve documents even if no perfect matches are present in the document collection.*

**User Story 7 Exploit all representations:** *A user can use all available representations to specify the current information need.*

## **User Stories for the Layperson Persona**

### Search Strategy

**User Story 8 Similarity search:** *A layperson user can find similar images to a given query image.*

**User Story 9 Dynamic information needs:** *A layperson user can change his information need during the current search.*

**User Story 10 Dynamic search strategies:** *A layperson user can change his search strategy during the current search.*

### Result Presentation

**User Story 11 Exploratory visualization:** *A layperson user can upload an image and explore the results with the help of a self-organizing map that relies on visual and metadata features.*

**User Story 12 Similarity grouping:** *A user can group results by similarity or dissimilarity.*

### Other

**User Story 13 Sorting support:** *A layperson user can sort images according to their own criteria.*

**User Story 14 Ease of use:** *A layperson user must find the software easy to use.*

## **User Stories for the Professional Persona**

### Search Strategy

## 6 Design of the Pythia MIR System Prototype

**User Story 15 Refindability of documents:** *A professional user can re-find documents that have been retrieved once.*

### Result Presentation

**User Story 16 Visualize the unusual:** *A professional user can display results in a way that the most outstanding/dissimilar documents amongst the results become visible.*

### Retrieval Model and Language

**User Story 17 Multimodal query specification:** *A professional user can search a library of digital documents by using an image-based or textual query, i.e., using the visual and verbal modality.*

### Other

**User Story 18 Weight representations:** *A professional user can actively affect the impact of certain representations during the retrieval.*

## **User Stories for the Researcher Persona**

### Other

**User Story 19 Weight visualization:** *A research user can visualize the weight setting of the CQQL query.*

**User Story 20 Direct weight manipulation:** *A research user can directly control the weighting of certain representations.*

### **6.1.3 Iterative User Feedback**

Because of the limited availability of the domain experts, other kinds of user feedback had to be used to refine the user stories and to test the GUI prototypes frequently.

Most of the user feedback about the prototypical designs has been obtained by subject matter experts as commonly recommended by a number of authors, e.g., Shneiderman & Plaisant [2005] or Cooper et al. [2007], during conferences and workshops.

A first version of the prototype has been shown at the PIKM workshop at CIKM 2011 [Zellhöfer & Schmitt 2011a]. This demonstration was followed by a presentation during a research stay at the IVA<sup>95</sup> and the spring meeting of the “Fachbereich Datenbanken und Informationssysteme” of the GI<sup>96</sup>. The audience of these demonstrations was predominantly academic. The demonstrations were shown in the form of a walkthrough, i.e., the central functionalities of the Pythia MIR system were presented with the help of a sample work task. This form of demonstration resembles cognitive walkthroughs that illustrate a sample problem-solving process of a potential user [Preece et al. 2002, cf. pp. 420]. However, the demonstration did not include a formal evaluation by the

---

<sup>95</sup>Royal School of Library and Information Science, Copenhagen

<sup>96</sup>The special interest group of databases and information systems of the German Informatics Society

experts. Instead, the feedback on the prototypical design was given informally through direct comments and discussion.

In addition to the mostly non-interactive walkthroughs, different prototypes were presented during demonstration sessions at the ICMR 2012 [Zellhöfer et al. 2012], ImageCLEF 2011, and ImageCLEF 2012. In addition, an advanced prototype was demonstrated to small groups of researchers at the IiX 2012. This version of the prototype is separately described by Zellhöfer [2012a] and is similar to the one described in Section 6.3. During these demonstration sessions, researchers had the chance to work interactively with the Pythia MIR system or to follow a walkthrough.

A comparable prototype to the one used at the IiX was shown in a third half-day workshop in June 2012 to an industrial audience comparable to the one described in Section 6.1.1. For the minutes of this workshop, see Appendix E. The same prototype was demonstrated to an academic audience at the BTW 2013 [Zellhöfer et al. 2013].

These interactive demonstrations and walkthroughs were accompanied by informal user feedback cycles with academic staff, friends, and relatives in order to get frequent and versatile feedback from diverse user groups.

Although one might argue that the approaches for feedback collection being presented in this section can be regarded as “quick and dirty” because they mainly lead to qualitative and subjective judgements about the shown prototype, their feedback into the UCD process is generally considered “an essential ingredient of successful design” [Preece et al. 2002, p. 341]. Comparable accompanying qualitative research is also advocated by other authors such as Cooper et al. [2007].

### 6.1.4 General User Experience Objectives

Besides the usage of contemporary software engineering techniques, the development of the Pythia MIR system is based on an extensive literature review. This literature review is crucial for obtaining requirements that have not been explicitly stated by the domain experts and users but that are indicated by user studies or other research results. Furthermore, the literature review complements the conclusions drawn from the modeled personas and is strongly advised by Cooper et al. [2007].

User Story 14 shows clearly that the workshop participants strive to obtain an easily usable MIR system. Guidelines on how to create a usable software system have been extensively studied by researchers in the field of usability engineering, e.g., by Preece et al. [2002]; Shneiderman & Plaisant [2005]; Cooper et al. [2007], or Nielsen [2009]. These guidelines consist of similar principles with little variation due to the research focus of the respective author.

For instance, Preece et al. [2002] suggest six principles (see below) that need to be addressed in order to build a usable system [Preece et al. 2002, cf. pp. 14], i.e.,

#### Definition 6.1 Usability principles:

1. *effectiveness is determined by measuring how well the system performs in what it is supposed to do,*

## 6 Design of the Pythia MIR System Prototype

2. *efficiency means how well the system supports the users in solving their work task,*
3. *safety subsumes the system's precautions to protect users from getting into dangerous conditions (e.g., losing their work progress) and the system's support in case errors have to be corrected,*
4. *utility means that the system provides the right functionality to solve a work task, while*
5. *learnability states whether the system's functionalities are easy to learn or self-explanatory, and*
6. *memorability expresses if a system can be re-used easily once it has been learnt, e.g., without consulting the manual.*

◇

Because the system's effectiveness is largely affected by its retrieval effectiveness in the context of this dissertation, this aspect is examined separately in Chapter 8.

Another example is the probably best-known guideline by Ben Shneiderman: the eight golden rules of interface design. These rules resemble the aforementioned ones but are not as formalized. Instead, they form a set of best practices that are listed below [Shneiderman & Plaisant 2005, cf. pp. 74]:

### Definition 6.2 **8 golden rules of interface design:**

1. *Consistency A usable system has to strive for consistency, i.e., the user interactions or the used terminology have to follow the same reoccurring patterns.*
2. *Universal usability Different usage behaviors should be supported by the user interface, e.g., an expert user should be allowed to use keyboard short-cuts while a layperson user can rely on a simple mouse-based interface.*
3. *Informative feedback The system has to provide informative feedback about its current state or the consequences of the current user actions.*
4. *Terminating actions Sequences of user actions should terminate clearly. For instance, a dialog should confirm clearly that all information that is required to complete a task has been input.*
5. *Error prevention The system should prevent users from getting into dangerous conditions (see above).*
6. *Reversal of actions Actions should be reversible in case of an error or if the user is not satisfied with the outcome of the action.*
7. *Controllability The user has to keep the system in control, i.e., the user has to be proactive instead of being reactive. In other words, a user should initiate an action which can be supported by the system.*



## 6.1 The User-centered Design Principle

8. Reduction of short-term memory load *That is, the user should not have to rely on the short-term memory to complete the current work task. For instance, crucial information has to be maintained or made accessible during all actions that are needed to complete a task.*

◇

Obviously, these two guidelines are interchangeable to a certain extent. For instance, a consistent and error preventing system is easy to learn, while a safe system has to address error prevention mechanisms and a reversal of actions.

The main difference between the two guidelines is Shneiderman's focus on the varying user needs of different user groups (or personas), i.e., the strive for universal usability. This aspect is also important in the design of the Pythia MIR system as three different personas with sometimes conflicting user stories have been discovered (see Section 6.1.2). For instance, the demand for a system that is easy to use for a layperson (see User Story 14) is hardly compatible with the need to control parameters of the retrieval engine (see User Story 20) – at least at first sight. Section 6.2.2 discusses how this issue is addressed.

Based on these principles taken from (general) usability engineering, Hearst [2009] builds a bridge to the domain of search user interface design by focusing on seven central aspects (the brackets indicate the corresponding rule of Shneiderman's eight golden rules):

1. "Offer efficient and informative feedback, [3]
2. Balance user control with automated actions, [7]
3. Reduce short-term memory load, [8]
4. Provide shortcuts, [2]
5. Reduce errors, [5, 6]
6. Recognize the importance of small details, and
7. Recognize the importance of aesthetics."

[Hearst 2009, p. 28]

The recognition of small details implies the consideration of factors such as the length of text input fields that encourage longer queries or document visualizations that support human perceptive capabilities [Hearst 2009, cf. Sec. 1.12]. As a result, these small details contribute to a positively perceived user experience, although their impact might be hard to measure.

Surprisingly, Shneiderman neglects the impact of the aesthetics on the user experience, although the consistent usage of layout elements can be considered an aesthetic factor. Obviously, an aesthetic interface, i.e., a visual compelling GUI, has a large impact on the user experience. For instance, Ben-Bassat et al. [2006] show that a system's subjectively perceived usability is affected by the evaluation of its usability and aesthetics (and vice versa). Additionally, aesthetics are subject to the user's cultural background. For instance, if the user associates a certain style of layout or color theme with an attribute, e.g., professionalism, this attribute might transfer to other products with similar

## 6 Design of the Pythia MIR System Prototype

aesthetics<sup>97</sup>. Similar phenomena are reported by a number of authors, e.g., by Dahm [2006] or Burmester et al. [2008]. Hence, the user expectations have to be met also in this dimension.

### Results from CBIR/MIR-specific User Studies

Besides these general points that help to refine the usability requirements of the system as postulated in User Story 14, there are further studies that complement the IIR research results presented in Chapter 3, i.e., the dynamic IN and the change of search strategies as reflected by User Story 9 and 10.

The dynamic nature of search strategies is reported by a number of authors in the field of CBIR and MIR. For instance, Urban et al. [2006] conclude from a user study with 18 post-graduate design students that exploratory search motivated users to interact with the examined CBIR system because they got the feeling that the searched images would be in the used collection. Using directed search, the users started to doubt whether a searched image is contained in the collection at all. From this perspective the results resemble the outcome of Kuhlthau's IR user study [Kuhlthau 1991] (see Section 3.1.1). In general, the users preferred the simple explorative UI over the more complex directed search interface. One particularly interesting point of the study is that it reveals that users "could not distinguish between images that have the same color and images that are generally similar (in terms of semantic content, layout, color, etc.)" [Urban et al. 2006, p. 26].

In another study with 20 subjects, McDonald & Tait [2003] investigate user search strategies when using the CHROMA CBIR system [McDonald et al. 2001], which supports directed query-by-sketch/QBE, browsing, and RF. The study clearly shows a general preference of the users for browsing. In contrast, whenever an already known images is searched, directed search is preferred. QBE is not examined separately in the study.

Regarding the acceptance of RF, Urban & Jose [2006] report that users (12 participants with academic and non-academic background) see RF as a useful technique for optimizing their results. Furthermore, the authors emphasize that the grouping of images which are then used as positive examples for RF is considered helpful by the participants of the user study.

Complementing these user studies, Cunningham & Masoodian [2006] analyze 64 image-related information needs that are extracted from a commercial search engine log. They report that browsing is used most often and should be better supported by search engines and interfaces.

Besides general CBIR, some authors explicitly investigate the specifics of personal photo retrieval, i.e., the usage of CBIR in the personal domain by mostly layperson users – a scenario that is also described by the participants of the conducted expert workshops (see Section 6.1.1).

---

<sup>97</sup>This effect is used in advertising and product branding on a large scale.

Cunningham & Masoodian [2007] also examine how users sort and re-find digital images in their personal photo collections. The authors report that “people are more likely to provide descriptions for a set of photos than they are for individual photos” [Cunningham & Masoodian 2007, p. 400]. These sets of photos (albums) are used for sharing and play a central role during the retrieval of a known image (re-finding). Albums are used for a first directed search because the examined users can typically remember a certain attribute describing the album (e.g., where the image was taken). After determining the right album, browsing is used to locate the searched image. During browsing, the users prefer a thumbnail visualization because it allows an efficient location of the searched image. Interestingly, the authors also report that users sometimes browse images only for recreational purposes, i.e., without a distinct information need.

Relying on a six months running user study with 13 participants, Rodden & Wood [2003] make similar observations but also give details of the usage of albums. They note that users organize only photos of good quality in albums. Bad photographs, i.e., ones with bad quality or an uninteresting motif, remain unsorted. Regarding their IN, the observed users mainly search for certain events or groups of photos that have something in common, e.g., the depiction of the same person. The importance of events as a central IN while retrieving images from personal photo collection is also emphasized by Sinha et al. [2009]. This finding is supported by a study accompanying this dissertation (see Section 8.3).

In order to facilitate the navigation through the collection, Rodden & Wood [2003] recommend to sort photographs chronological and to display a large number of thumbnails. The importance of providing a supportive overview of the collection’s content during browsing unfamiliar information spaces is also underlined by Cribbin & Chen [2001]. Concluding a study with 21 student users, the authors argue for grouping images on a common criterion but note at the same time that the support of different search strategies should be prioritized. The authors also advocate the provision of a navigational history of the user’s search progress, e.g., a undo and redo functionality that allows users to restore different stages of their search session. This suggestion links their work to Liu et al. [2009], who studied the search behavior of 17 student users. In their paper, the authors emphasize the importance of a navigation through prior search steps that has to be augmented with a preview of the (old) results in order to reduce the short-term memory load of the users. In addition, Liu et al. [2009] find evidence that different user groups experience the same CBIR system rather differently, i.e., that “age and search experience affect system satisfaction” [Liu et al. 2009, p. 11]. The authors also underline the users’ demand to use drag and drop to give their relevance feedback in an efficient manner.

### **Conclusion**

Besides the more general user experience objectives such as Shneiderman’s eight golden rules of interface design (see Definition 6.2), the user studies reveal additional points that are important in order to build a usable MIR system. It is not surprising that these points coincide with the results from IIR research presented in Chapter 3 to a great

## 6 Design of the Pythia MIR System Prototype

extent:

1. Multiple information seeking strategies (see Section 3.3), i.e., browsing and directed search in particular, have to be supported by the system to increase its utility.
2. Relevance feedback is an acceptable means for query refinement but has to be efficient, e.g., multiple documents should be associated with a relevance level at once or drag and drop should be available to give feedback.
3. The visualization of query results has an impact on the overall efficiency of the system. That is, documents should be arranged by groups based on a commonality or be displayed in a way that a quick overview becomes possible, e.g., by using thumbnails.
4. Different user groups (or personas) experience and use the same system in different ways. This diversity has to be reflected in the UI (see Definition 6.2 [2]).
5. The support of queries that aim at the retrieval of events is crucial for the utility of retrieval systems accessing personal photo collections.
6. As argued in Chapter 3, the user's and system's concept of similarity has to be considered holistically. Otherwise, the gap between the user's mental and the system's conceptual model widens, which results in unpredictable results and user dissatisfaction as Urban et al. [2006] show.

The last point can be considered as the central challenge in usability engineering. Without bridging the *interaction gap* between the mental and the conceptual model, it becomes hard to build a usable system [Cooper et al. 2007]. Amongst others, this issue is amplified in (M)IR because of the user's insecurity about the current information need and the challenge to express it appropriately. In combination with the semantic gap (see Section 2.3.3), which affects both query and retrieved documents, it is likely that users tend to develop a mental model differing from the conceptual model of the system they are interacting with. As a consequence, the success of a search depends mainly on the user's ability to communicate the IN and the system's ability to communicate its conceptual model of similarity. This interplay is also stressed by the cognitive viewpoint on IR described in Section 3.1.2.

In this thesis, similarity is determined by a number of representations, a weighting of these representations, and a logical CQQL query serving as the matching function in the MIR engine. Given a sufficiently complex weighted CQQL function, the actual setting of the weighting variables, their interaction, and eventually their effect on the retrieval result can be hardly foreseen by a user [Zellhöfer & Lehrack 2008] – even if relevance feedback is used. Such unpredictable effects lead to an irritation of the user and a degeneration of the mental model because actions do not cause the expected reactions. Hence, it is important to enable both the user and the system to communicate their current state to keep the interaction gap narrow as elaborated in Section 5.4.2.

### 6.1.5 Summary

To recapitulate, Section 6.1 outlined the three main principles of the UCD approach that have to be considered during software development: 1) the inclusion of the user, 2) the reflection on user reactions and performance measurements, and 3) the utilization of an iterative design process.

Section 6.1.1 addressed the first principle by describing the inclusion of users in the creation of personas and scenarios that were used as basis for the development of the Pythia MIR system. Section 6.1.2 presented how user reactions have been included into the process to refine and to improve the central user stories. Complementing and accompanying iterative user feedback cycles (see Section 6.1.3), the retrieval effectiveness of the system was constantly examined and fine-tuned. This aspect is discussed separately in Chapter 8.

Moreover, an agile software development process has been used. This process allows the continuous integration of user feedback and an interactive refinement of the system functions according to the user needs.

Besides the UCD process, the second pillar on which the development of the Pythia MIR system is founded, is the literature review presented in several parts of this dissertation, e.g., in Chapter 3 or Section 6.1.4.

Additional usability engineering techniques, such as predictive modeling, e.g., in form of Fitts' law [Fitts 1954], have not been used during the development of the prototype because of the constant integration of users into the development process. Fitts' law as a predictor of the time a user takes to reach a target object, e.g., a GUI button of a given size, using a mouse or other pointing devices does not deal with the complex user interactions in a MIR system. Although it can be used to find the optimal location (in other words, the quickest accessible location<sup>98</sup>) for various GUI elements and how these elements should be related to each other, it will not reveal general usability problems. In contrast, screen layout problems can also be revealed by incorporating iterative user feedback.

All these considerations have been taken into account in developing a conceptual model for the Pythia MIR system.

## 6.2 Conceptual Model

This section presents the conceptual model of the Pythia MIR system prototype, which is based on the preference-based relevance feedback approach PrefCQQL (see Chapter 5) and the cognitively motivated principle of polyrepresentation (see Section 4.6).

This part of the dissertation is structured as follows. Section 6.2.1 presents the theoretical basis of the presented conceptual model of the Pythia MIR system. In Section 6.2.2, the general principles of the system's interaction design are described which are meant to deliver the conceptual model respecting the user needs presented in Section

<sup>98</sup>Roughly speaking, Fitts' law states that bigger targets are quicker to reach than smaller ones.

## 6 Design of the Pythia MIR System Prototype

6.1. The subsequent section focuses on important functionalities and relates them to the user stories presented in Section 6.1.2.

### 6.2.1 Theoretical Foundations of a Polyrepresentative User Interaction Model

One of the core ideas of the cognitive viewpoint on IR – the observation that information processing is taking part in both the user and the system – has been presented in Section 3.1.2. Section 3.2 discussed the principle of polyrepresentation (PoP) of the information space, i.e., how documents are represented in the IR system. As mentioned before, there is also an aspect of the principle addressing the user's motivation to interact with an IR system that is often neglected: the principle of polyrepresentation of the cognitive space, which is based on an extensive literature review and analyses of empirical studies by Ingwersen [1996].

Evolving Belkin's ASK hypothesis [Belkin 1980], Ingwersen relates the IN to different cognitive structures: a (mostly) stable social or work context (the context the user is working in) and more dynamic cognitive structures such as the problem space (which equals Belkin's ASK) and the current cognitive state, i.e., "the little known about what is desired" [Ingwersen 1996, p. 15] (see Figure 6.3; right).

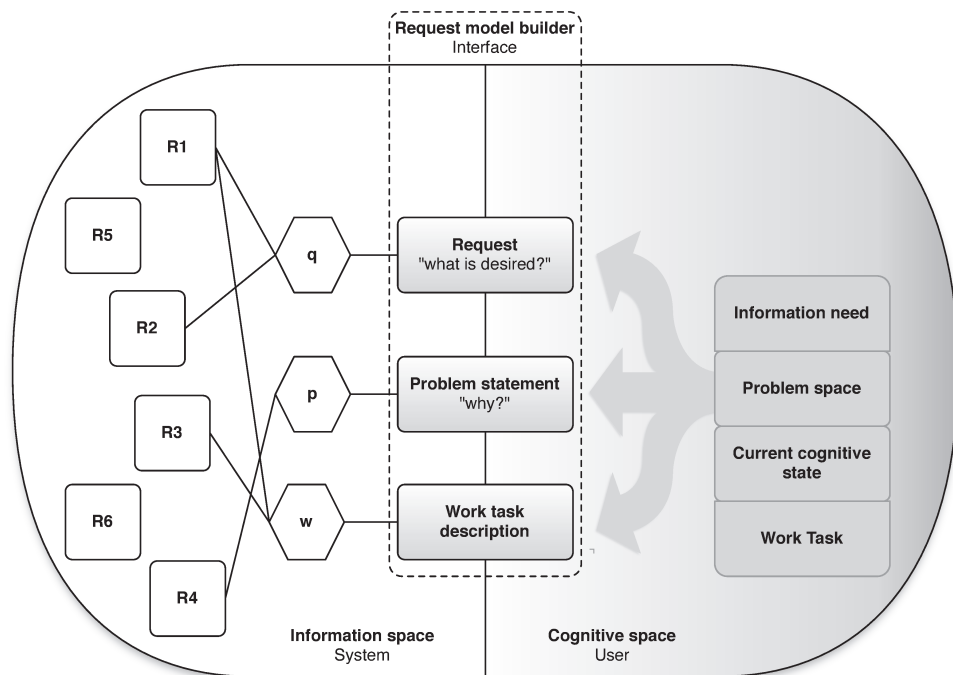


Figure 6.3: The global model of polyrepresentation [Ingwersen 1996, cf. Fig. 8]

The state of these cognitive structures manifests in the intrinsic IN that is subdivided

by Ingwersen [1996] into four extremes cases, which are best supported by different information seeking strategies (ISS) (see Figure 6.4). By decoupling the IN from the other cognitive structures, it becomes possible to explain that the same problematic situation can lead to different IN.

For instance, users with a well-defined and stable IN (see Figure 6.4; 1) are most likely able to assess the relevance of the retrieved documents because of their domain knowledge (reflected by their current cognitive state) and are certain about their current problem space, i.e., their current area of interest [Ingwersen 1996, cf. pp. 15]. As a consequence, these users are able to formulate a confined query in order to retrieve the most informative documents to satisfy their current IN.

	Well-defined	Ill-defined	Supportive ISS
Stable	<b>1 Rich, variable, cognitive state</b> Limited uncertainty Can assess relevance Low curiosity Confined navigation	<b>4 Weak, variable, cognitive state</b> High uncertainty Cannot assess relevance Low curiosity Dead-end navigation	Directed search
Variable	<b>2 Rich, variable, cognitive state</b> Controlled uncertainty Can assess relevance High curiosity Exploratory navigation	<b>3 Weak, variable, cognitive state</b> High uncertainty Cannot assess relevance High curiosity Random browsing	Exploratory search

Figure 6.4: Matrix of the four intrinsic information need extreme cases according to Ingwersen [Ingwersen 1996, cf. Fig. 3]

Another example are users with a variable and ill-defined IN (see Figure 6.4; 3). Such users are uncertain about their problem space and cannot assess the relevance of retrieved documents due to their lack of domain knowledge. On the other hand, they are willing to learn and can therefore move to a different IN state (e.g., # 2) if they have improved their knowledge of the currently searched topic (their current cognitive state) by exploring the information space. These examples clearly show the variability of the first three cognitive structures, while the work task remains stable because it is determined by the user's work or social context.

The examples also emphasize the need for a user interface that supports different information seeking strategies and a transition between them. From this point of view, the IN model suggested by Ingwersen [1996] is consistent with the findings that have been presented throughout the last chapters, i.e., that a MIR system should support different ISS in order to satisfy user needs.

### Linking the Cognitive and the Information Space

According to Figure 6.3, the information and cognitive space have to be connected via an interface that has to fulfill two main tasks in order to establish a dialog between the user and the system:

1. "to assess the proper nature of the [information] need and the underlying cognitive state [... and]
2. to make adequate use of such small request contexts, in particular at the crucial initial phase of IR interaction."

[Ingwersen 1996, p. 21]

Thus, in order to interact with the IR system, the user has to communicate these cognitive structures via an interface or, as it is called by Ingwersen [1996], a request model builder (see Figure 6.3; center). In an ideal situation, three functionally different representations of the user's cognitive space can be extracted at a point in time:

1. "a 'what', i.e. a *request* version which includes what is currently known about the unknown (the wish or desire for information);
2. the 'why', i.e. a *problem statement* as well as
3. a *work task and domain description*."

[Ingwersen 1996, p. 18]<sup>99</sup>

Further, Ingwersen acknowledges that these representations may coincide, e.g., the work task description (3) and the problem description (1). In any case, he assumes that at least the request ("what") and problem statement ("why") should be extractable. In accordance with the description of the dynamic nature of the user's cognitive structures, he further describes the work task description as stable, whereas (1) and (2) change over time.

In consequence, the quality of the extraction of the representations of the user's current cognitive space has a huge impact on the outcome of the IIR process. Based on these representations (see Figure 6.3;  $q$ ,  $p$ , and  $w$ ), the IR system is assumed to discover relevant representations ( $R_i$ ) pointing to documents (omitted in the illustration) or "network[s] of concepts" [Ingwersen 1996, p. 39]. These networks in the information space can be explored along the so-called conceptual paths. In other words, for each representation of the cognitive space there is a corresponding network of relevant concepts (or representations in information space) retrieving different sets of documents. This means that there are paths retrieving relevant documents which would not be utilized by providing only one representation of the cognitive space. For instance, a request ( $q$  path) of a user with an ill-defined and variable IN could be augmented by including the current work task ( $w$  path), resulting in more potentially relevant documents.

Eventually, these different result sets can be fused or combined comparably to the principle of polyrepresentation of the information space (see Section 3.2).

---

<sup>99</sup>The accentuation has been inserted by the author of this text.



Unfortunately, the actual extraction of the representation of the cognitive space remains a serious challenge and requires further studies on the user's cognitive processes during the search and the inclusion of contextual knowledge. Regarding the first point, little research that allows the design of a generally usable UI has been done so far [Larsen 2004, cf. pp. 30]. The second point has gained much more attention over the last years as an increasing number of IR systems incorporate contextual knowledge about the user in form of user profiles or similar data sources, e.g., by including the user's location to optimize search results or by accessing the user's search history.

### A Polyrepresentative User Interaction Model based on CQQL and PrefCQQL

As a consequence, the development of a polyrepresentative user interaction model is needed in order to develop a user interface based on the principle of polyrepresentation. The CQQL- and PrefCQQL-based polyrepresentative user interaction model presented in this section has been originally published by Zellhöfer [2012a]. Some of its core concepts have been discussed in Section 4.6.

Unsurprisingly, the cognitive overlap (CO)<sup>100</sup> forms the central concept of the model. Initially, the CO is formed by an arbitrarily structured CQQL query provided by the user (or any other actor involved in the information seeking process) or a weighted CQQL conjunction of all available representations [Zellhöfer 2012a, cf. Sec. 3.2] if no distinct CQQL query is provided. The choice of a weighted CQQL conjunction as the standard CO matching function is motivated by its retrieval effectiveness during relevance feedback which is discussed in Section 8.5. Moreover, the conjunction is the logical counterpart of the intersection in set theory and thus the "overlap" of different sets in the originally Boolean PoP (see Section 3.2).

**Directed Search in the Model** For illustrative purposes, we will assume the CO to be represented by a weighted CQQL conjunction in a QBE scenario. In other words, the user starts with a directed search.

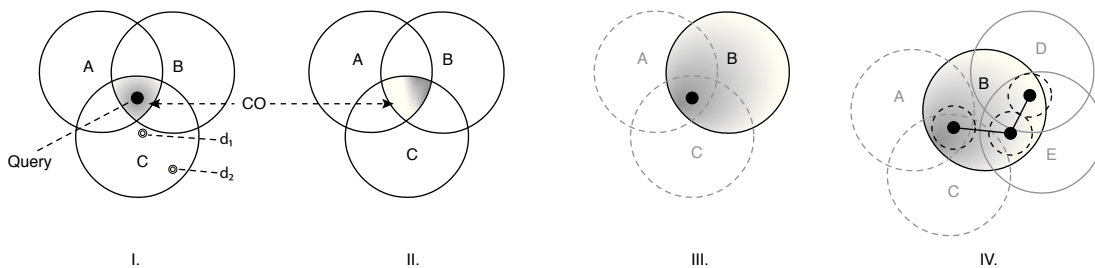


Figure 6.5: Polyrepresentative user interaction model

<sup>100</sup>To be more precise, the model is based on a PCO because of CQQL's best matching semantics (see Section 4.6).

## 6 Design of the Pythia MIR System Prototype

Figure 6.5 (I.) displays the CO formed by three representations in a hypothetical information space. The black point in the center of the CO indicates the best known answer to the current IN, i.e., the specified QBE document. The figure must not be mistaken with a Venn diagram because we are not dealing with crisp sets. Instead, each circle labeled with a capitalized letter shows the region of the information space in which documents that are relevant to the corresponding representation are contained<sup>101</sup>. To reflect CQQL's best matching semantics, the probability of relevance (POR) of the documents regarding, e.g., representation *C*, decreases with an increasing distance to the CO's center. For instance, document  $d_1$  has a higher POR than  $d_2$  regarding representation *C*. Additionally, the intensity of the gray shading indicates the impact or contribution of each representation's relevant documents on the CO. That is, how much the underlying IR model rewards the POR of a document regarding a specific representation during the calculation of its overall POR. Internally, the impact of a representation on the CO is implemented with the help of CQQL's weighting variables. Figure 6.5 (I.) illustrates the equi-weighted case (all  $\theta_i = 1$ ), in which all considered representations contribute equally to the CO, i.e., there is no preference amongst the representations.

**Addressing the User's Cognitive Structures** In this example, the *request* (see Figure 6.3) is obviously represented by the specified QBE document. To incorporate the *problem statement* or the *work task* of the user, the model basically offers three options.

*First*, the CO can be modeled by a pre-defined CQQL query that might be formulated by a domain expert. This CQQL query can have an arbitrary form.

*Second*, the initial CO can be augmented with additional representations. For instance, a request that involves only representations *A* and *B* might be extended with *C* if prior studies have revealed that *C* contains valuable information about the current work task.

*Third*, the weighting of the underlying CQQL query can be altered in order to express the importance of certain representations. The needed data to represent the user's cognitive structures can be derived from user observations, logs, domain experts, or other sources. Figure 6.5 (II.) shows a re-weighted CO that increases the impact of documents which are highly relevant regarding representation *B*. In this scenario, documents that have a high POR regarding *B* will obtain a higher overall score in the evaluation of the underlying CQQL matching function and therefore be ranked in early positions of the result rank. As the initial CQQL query's logical structure has not been changed, the CO remains the same as before, but the gray shading becomes darkest at the *B* corner of the initial CO to express the weighting difference.

**Preferences and Relevance Feedback in the Model** Alternatively, the re-weighting can be caused by PrefCQQL's RF mechanism (see Section 5.4). Here, the weight learning algorithm deduces an appropriate weighting scheme for the CO-modeling CQQL query in order to satisfy the user's preferences. To conclude, users can also modify

<sup>101</sup>Please note that the same limitations regarding the visualization of the information space and CQQL as those mentioned in Section 4.4.1 apply here.

the underlying weighting variables directly if they are able to quantify their (quantitative) preferences between the available representations. To give an example, if a user is only searching for paintings of a specific painter, she might only value documents that belong to the same artist as the QBE document no matter how they look like.

**Explorative Search in the Model** If users do not have this amount of prior knowledge, they can also explore the information space along one or more representations. For instance, if PrefCQQL has shifted the weighting scheme slightly into the direction of *B*, users might inspect more documents that have a high POR regarding this representation. Internally, the model handles this exploration by a dynamic re-weighting of *A*, *C*, and *B*. While the impact of the first two representations on the CO gets lowered, the model increases the impact of *B*.

Eventually, the exploration along one or more representations results in a faceted search (see Section 3.3.2) [Zellhöfer 2012a, cf. Sec. 3.3]. In this case, the initial CO is “replaced” with all documents satisfying the current facet if *B* is Boolean or with all documents that have only a high POR regarding *B* as illustrated in Figure 6.5 (III). Here, documents that are relevant regarding *A* and *C* are no longer of interest. It is noteworthy that this effect is achieved by the utilization of CQQL’s weighting variables only. Apart from the representations associated with the facet, all representations (or CQQL conditions) are disabled using weighting variables, thus leaving the original query (and thus CO) fully intact. Hence, the user can always return to the initial CO. In other words, users can seamlessly switch between directed and explorative ISS. To be more precise, “a facet can be regarded as a weighting scheme for a CQQL query that can be used for filtering the current results *or* to retrieve new documents” [Zellhöfer 2012a, p. 67]. For instance, the presence of persons on a photograph can be used as an intuitively comprehensible facet (see Section 9.1). Depending on the setup of the MIR system, the facet can be used to filter the documents contained in the original CO or used to retrieve all documents satisfying the facet as depicted in Figure 6.5 (III).

An additional advantage of the integration of exploration and faceted navigation is that users are enabled to learn about their IN which can be used to adjust the weight setting during directed search in return. Moreover, if an important representation is discovered, this information can be communicated to PrefCQQL’s learning algorithm as an additional constraint, e.g., to make the corresponding weighting variable constant.

To conclude, browsing in the context of the polyrepresentative user interaction model can be interpreted as follows. Figure 6.5 (IV.) shows the clusters of the most similar documents regarding *B* with the help of dashed black lines. During browsing, the user moves the region of interest (ROI, black point) through *B* inspecting only the documents surrounding the current ROI with the help of an appropriate visualization (see Section 6.2.4). Finally, the user might end up at the rightmost point that describes the new IN best. Consequently, the found document can serve as a new QBE document, which is used to create a new CO of the already known representation *B* and the newly discovered representations *D* and *E*. Please note that browsing does not require a pre-defined CO. Instead, it also works with the full content of the collection. However, for

## 6 Design of the Pythia MIR System Prototype

the sake of consistency, browsing can also be based on a weighted CQQL query with all weighting variables set to zero. The arithmetic evaluation of such a query results in the same POR for all documents in the collection because weighted conjunctions always yield 1 and weighted disjunctions always 0 (see Section 4.3).

**Permeable Information Seeking Strategies** These examples point out the continuous opportunities to switch between different ISS without leaving the polyrepresentative user interaction model. As summarized in Figure 6.4, this change between different ISS is necessary because of the different IN states a user experiences during information seeking. For instance, users with a weak cognitive IN state are enabled to start their search with browsing the collection's contents, slowly moving towards a directed search when first relevant QBE documents are discovered. The retrieved results can then be further personalized with the help of PrefCQQL, which might give them more clarity about certain aspects of their IN that can be subsequently explored. All ISS supported by the presented polyrepresentative user interaction model are based on CQQL and interpreted by an modification of the underlying weighting scheme or the logical structure of a CQQL query representing the current CO. As a side effect, each query or browsing path can be saved and easily reconstructed in order to support the re-findability of documents during subsequent searches. This is not possible with pure browsing approaches, such as the ostensive model [Campbell 2000], which would require the storage of the original browsing path and the fixation of the collection's content (see Figure 3.3). With CQQL, the storage of the weighting scheme and the CQQL query is sufficient to re-find relevant documents regarding an old CO. In case of a change in the collection's content, only the result rank might differ from the original one if documents were added or removed. If a document that is part of the browsing path of the ostensive model would be removed, the path would be irrecoverably destroyed and can no longer be reconstructed.

Furthermore, as postulated by Ingwersen [1996], CQQL offers different means to incorporate the user's cognitive space into the information seeking process. To facilitate the comprehension of this rather theoretical explanation, an illustrated description of a sample interaction with the Pythia MIR system (see Section 6.3) is included in Appendix D.

**Dealing with Explicit Information Need Changes** The ISS described in the earlier parts of this section have in common that they are all reflecting a slowly evolving IN. However, there might be scenarios in which the IN change is more drastic, e.g., in the case of a spontaneous inspiration or the rediscovery of some hidden prior knowledge that has an effect on the user's request.

To react to such drastic IN changes, a specific PrefCQQL indicator based on preference conflicts<sup>102</sup> (see Section 5.4.7) can be used here. Given a set of preferences, the assumed likelihood that a user is inputting conflicts is low as long as the IN changes

---

<sup>102</sup>Note that all of the following actions can also be started manually by the user.

slowly. Furthermore, the risk of conflicts can be minimized if old preferences are removed during the search process similar to the preference aging in the ostensive model [Campbell 2000]. In case a conflict occurs, we have to check if the conflict has been introduced between an old preference and a new one. In this case, we can remove the old preference earlier than expected because we assume the old preference to be outdated, i.e., it does no longer reflect the current IN. The situation differs if preference conflicts occur frequently between recently input preferences. Such conflicts indicate that even recently added preferences are no longer consistent with the IN or that the current query cannot fulfill these preferences. Hence, it is likely that there is another CO that models the current IN in a better way.

To recapitulate, after a change of the IN has been detected or manually toggled, we assume that the last input preferences describe the new IN best. That is, the user provides a set of preferences  $P = \{p_1, \dots, p_m\}$ , whereas each preference  $p_i$  expresses a pairwise preference judgement between two documents:  $d_{i_1} \geq d_{i_2}$ . In order to learn a more appropriate CO from these preferences, a new CQQL query that models the new CO can be learned as described in Section 5.5.

To conclude, the presented model contains an additional indicator that might be utilized to assist the user. As described above, by observing the ISS that are utilized by the user we can draw conclusion about the user's IN stability. For instance, if the user is constantly switching between browsing and relevance feedback without altering the impact of the representations on the CO very much, it is reasonable to assume that the user got stuck. In such case, the MIR system can suggest to learn a CQQL query in order to assist the user in getting new insights into the current IN. However, the usage of such assistance functions is not free from risk because they can cause feelings of control loss for users.

The next section discusses an implementation of the described theoretical model in form of a user interaction model for the interactive Pythia MIR system prototype.

## 6.2.2 Conceptual Model of the Pythia MIR System's User Interaction

“When providing new search functionality, system designers must decide how the new functionality should be offered to users. One major choice is between offering automatic features that require little human input but give little human control, or interactive features, which allow human control over how the feature is used but often give little guidance over how the feature should be best used.”

[White & Ruthven 2006, p. 993]

In this quote, White and Ruthven aptly describe the core problem of the provision of new search functionalities. Their argument is also supported by the results of our conducted user requirement workshops (see Section 6.1.1).

At the same time, they neglect an important factor: the inhomogeneity of the addressed user group. Frequently, the user group is seen an abstract entity: *the* user. In reality, as Section 6.1.2 has already shown, a (M)IR system has to address the different needs of personas. As a result, a usable (M)IR system not only has to find a compromise between manual controllability and automatisms provided by the system; it also has to

## 6 Design of the Pythia MIR System Prototype

address different user needs and usage scenarios adequately.

Another point of criticism is the predefinition of many (M)IR system on one information seeking strategy (ISS). Such systems often can be clearly discriminated in directed or exploratory search systems. This implies that a seeker will only employ one ISS in order to satisfy an IN. It has been shown in various parts of this dissertation that this is not the case (see User Stories 9 and 10 or Section 3). This argument is revisited in Section 6.4.

In conclusion, there is no reason to separate common ISS types in order to develop a usable, interactive MIR system. Hence, this section outlines the general user experience objectives derived from the requirements presented in Section 6.1 and explains how different user needs can be addressed all together in one user interface.

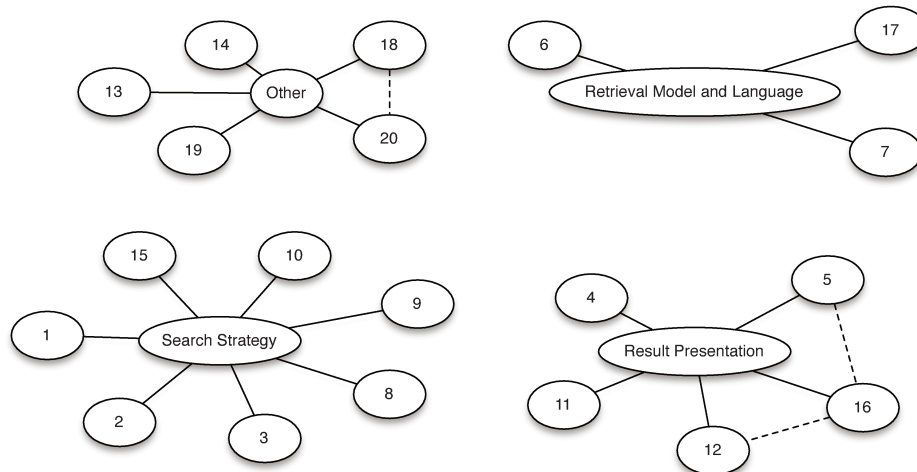
### General User Experience Objectives

User Story 14 postulates the ease of use of the Pythia MIR system's GUI for layperson users. The satisfaction of this requirement is complicated by the fact that the other personas, i.e., a professional user and a researcher, are most likely to perceive a good usability differently. The objective to achieve a good usability for various user groups is typically referred to as *universal usability* in usability engineering and interaction design.

Before we will concentrate on how universal usability is achieved in the Pythia MIR system prototype, a closer examination of the user stories presented in Section 6.1.2 is needed. Figure 6.6 groups all user stories thematically. The visualization makes clear that the user stories are related to three main classes, i.e., the search strategy, the result presentation, and the retrieval model and language. Four additional user stories cannot be easily classified as they are related to non-functional requirements or PrefCQQL specifics. The retrieval model and language behind the Pythia MIR system, i.e., CQQL and PrefCQQL, have already been discussed in detail in Chapter 4; therefore, this section omits the related User Stories 6, 7 and 17.

The remaining groups contain user stories that deal with the demanded search strategy support, e.g., a fast adapting RF mechanism, QBE, or the support of various search strategies during the whole interaction time, as well as the result presentation. As described before, the user stories differ with respect to the persona. For instance, the layperson User Story 11 asks for a result visualization that supports the exploration of a collection's content, while the professional User Story 16 describes the need for a visualization of unusual documents in the results. Similar observations can be made with the ISS: while laypersons might be satisfied with a QBE-based similarity search (see User Story 8), both the professional and the researcher persona wish to directly modify the weighting variables of the underlying CQQL-based retrieval engine (see User Stories 18 and 20).

Although these examples make clear that the expected functionality differs between the personas, the participants of the expert workshops asked for a GUI that is intuitively controllable for a layperson (see Section 6.1.1). They even rendered their argument more precisely by explicitly demanding drag and drop support. While these



Dashed lines indicate semantically equal user stories.

Figure 6.6: Thematical groups of user stories

demands might sound contradictory at first, we believe that they emphasize the users' expectation to obtain a MIR system that can be controlled in a consistent manner no matter if layperson or professional functions are operated. Last but not least, intuitivity, consistency, and controllability are inseparable from a usability perspective. This relation was already shown in Section 6.1.4 with the introduction of Shneiderman's eight golden rules of interface design (see Definition 6.2):

- |                         |                             |
|-------------------------|-----------------------------|
| 1. Consistency          | 5. Error prevention         |
| 2. Universal usability  | 6. Reversal of actions      |
| 3. Informative feedback | 7. Controllability          |
| 4. Terminating actions  | 8. Reduction of memory load |

One interaction design technique that addresses all but the second point is the *direct manipulation paradigm* (DMP) [Shneiderman 1987] described in the next part of this section. Because the Pythia MIR system is developed as a prototypical system, not all of the listed rules will experience the same attention in the conceptual UI model. For instance, informative feedback and the directly related provision of online help facilities are not in the focus of this dissertation. Without doubt, attention to these points is crucial for an end-user-ready software system; however, it is not expected to contribute much to the research focus of this text. How the Pythia MIR system prototype tries to guarantee universal usability is discussed after the introduction of the DMP.

### Interaction Design based on the Direct Manipulation Paradigm

Unlike typical IR search user interfaces that rely on a textual query input and a result list of documents or document abstracts, the Pythia MIR system prototype follows the DMP, which is also used by most current operating systems and other software systems [Shneiderman & Plaisant 2005, cf. Ch. 6]. The DMP relies on metaphoric objects in order to allow users to relate to the internal processes of the (MIR) system and to manipulate these objects directly. It is typical for the DMP that all actions which can be carried out with an object are directly visible and operable, e.g., a file can be moved to the trash bin using drag and drop. The advantage of this approach is that users do not have to confront themselves with technical issues such as freeing sectors on a hard disk in order to remove sequences of bytes that are perceived as a file. Another characteristic of the DMP is that actions can be easily corrected by reverting an action, e.g., by dragging a deleted file from the trash bin to restore it. Moreover, the termination of an action becomes intuitively clear as the aforementioned example illustrates: after the file is restored, it is obvious that no further operations are needed.

Although the advantages of the DMP are obvious and its utility has been shown in countless instances, the approach is not free from problems. For instance, the metaphors and icons have to be implemented consistently and understood by the user. In other words, the user's mental model (i.e., the assumption of the user of how the system is working) has to overlap with the conceptual model of the system. Otherwise, the user will not be able to draw the right conclusions from the UI and consequently will not be able to interact predictably with the system. Hence, Shneiderman & Plaisant [2005] suggest three principles that will be followed in the conceptual model of the Pythia MIR system:

1. "Continuous representations of the objects and actions of interest with meaningful visual metaphors.
2. Physical actions or presses of labeled buttons, instead of complex syntax.
3. Rapid, incremental, reversible actions whose effects on the objects of interest are visible immediately."

[Shneiderman & Plaisant 2005, p. 234]

If one follows these principles, novice users are expected to be able to quickly learn a system as postulated by User Story 14; its operational concepts should be memorizable, error messages are rarely needed, and the action's consequences are visible and can be reversed, e.g., by inverting the performed actions. Eventually, users get the feeling that they are in control of the system (e.g., see User Story 2). Furthermore, the mental load is reduced because users do not have to decompose their task into multiple UI commands [Shneiderman & Plaisant 2005, cf. pp. 234].

Whether the design decisions that are made on the basis of these principles and that are described in the following sections lead to a usable system is discussed separately in Chapter 9.

Following the iterative UCD process used to develop the Pythia MIR system, the system's conceptual model has experienced various evolutionary steps. For the sake



of brevity, a discussion on the different development stages are omitted in this thesis. Instead, the most mature version of the Pythia MIR system's conceptual model is presented. To great extent, this model corresponds to the one of the expert search system outlined in Zellhöfer [2012a]. Previous versions of the conceptual model are discussed in Zellhöfer & Schmitt [2011a] and Zellhöfer & Schmitt [2011b].

### Overview of the Conceptual User Interface Model

As mentioned before, the core of a conceptual UI model based on the DMP is its metaphors. Reconsider van Rijsbergen's quotation given at the beginning of Chapter 2: "In principle, information storage and retrieval is simple. Suppose there is a store of documents and a person (user of the store) formulates a question (request or query) to which the answer is a set of documents satisfying the information need expressed by his question." This quotation underlines the importance of documents during information retrieval. Obviously, the retrieval of a set of relevant documents is the main motivation to interact with an IR system.

As stated at the beginning of Chapter 6, the Pythia MIR system prototype deals with visual perceivable media such as images and metadata. Hence, a document in the scope of this dissertation is typically an image or a photograph. In traditional analog photography, back-lit devices called *light tables* were used to inspect negatives, diapositives, or even photographs. Light tables were used to assess the quality of negatives, to crop images, or for rotoscoping<sup>103</sup>. Nowadays, the *light table metaphor* can often be found in (semi-)professional digital photograph software. For instance, Apple Aperture's GUI (see Figure 6.7) features further typical metaphors from analog photography: a magnification lens (1), a pre-print of the currently processed photograph (2), and a light table (3). In addition, the GUI provides an inspector for metadata that is read from the currently displayed image (4) and various floating dialogs that allow the modification of various technical image processing or filter parameters.

Being clearly focussed on the needs of (semi-)professional digital photography, Aperture's GUI is rather technical and expects the user to have prior knowledge of typical digital photography metadata and the related technical terms. Although the GUI leaves plenty of space for the preview of the digital photograph, it does not focus on the document itself. Instead, it juxtaposes the document and its digital representation (the file format, its color space, its metadata etc.). This is due to the software's task of image processing but might not be adequate for a document-centric MIR retrieval system. Furthermore, the GUI is largely dialog- and menu-driven. Roughly speaking, the only possible direct manipulations of the images are moving them between different folders via drag and drop.

Although the Pythia MIR system prototype's conceptual UI model is also based on the light table metaphor, it is more document-centric in order to acknowledge the user needs during MIR. In principle, the conceptual model of the Pythia MIR system is based on only two metaphors: the *light table* and the *Polaroid-like photograph*. The light table

<sup>103</sup>Light tables are still used in hospitals to inspect x-ray images.

## 6 Design of the Pythia MIR System Prototype



Figure 6.7: Apple Aperture's central GUI elements; version 3.2.4

serves as the space on which result photographs are placed and sorted by various criteria. As in real life, photographs can be freely moved on the light table and dragged to other tools such as the search bar or used to state preferences between them. At the same time, the light table serves as a container of the documents resulting from the retrieval step, which are visualized with photographs regardless of the type of search – directed or exploratory – applied. Basic tasks can be carried out by using drag and drop only, i.e., the definition of QBE documents or the elicitation of preferences in order to give relevance feedback to the system.

Figure 6.8 illustrates the conceptual UI model of the Pythia MIR system prototype. The illustration is only meant to give an overview. The subsequent sections address each component of the GUI in more detail. For the sake of clarity, each of the following sections also references all directly related user stories. The numbers in Figure 6.8 indicate the interface layers which will be addressed later.

As said before, the main components of the GUI are the light table and the Polaroid-like photograph. The photographs act as document surrogates and are the central element of the user interaction. The photographs are used to display relevance information, metadata, and a thumbnail of the image content. Both metaphors are described in more detail in Section 6.2.3. Being central to the user interaction model, the photograph can be freely moved by the user and dragged at different tools in order to execute the associated actions. For instance, a photograph can be dragged onto the search bar, which will be described in Section 6.2.4, to add this photograph to the query. Besides QBE documents, the search bar also handles textual queries.

The search history (see Section 6.2.7) provides a linear redo and undo functionality to the user. Each step in the time bar is associated with the state of the light table at a given point in time. That is, the user has access to different virtual light tables that

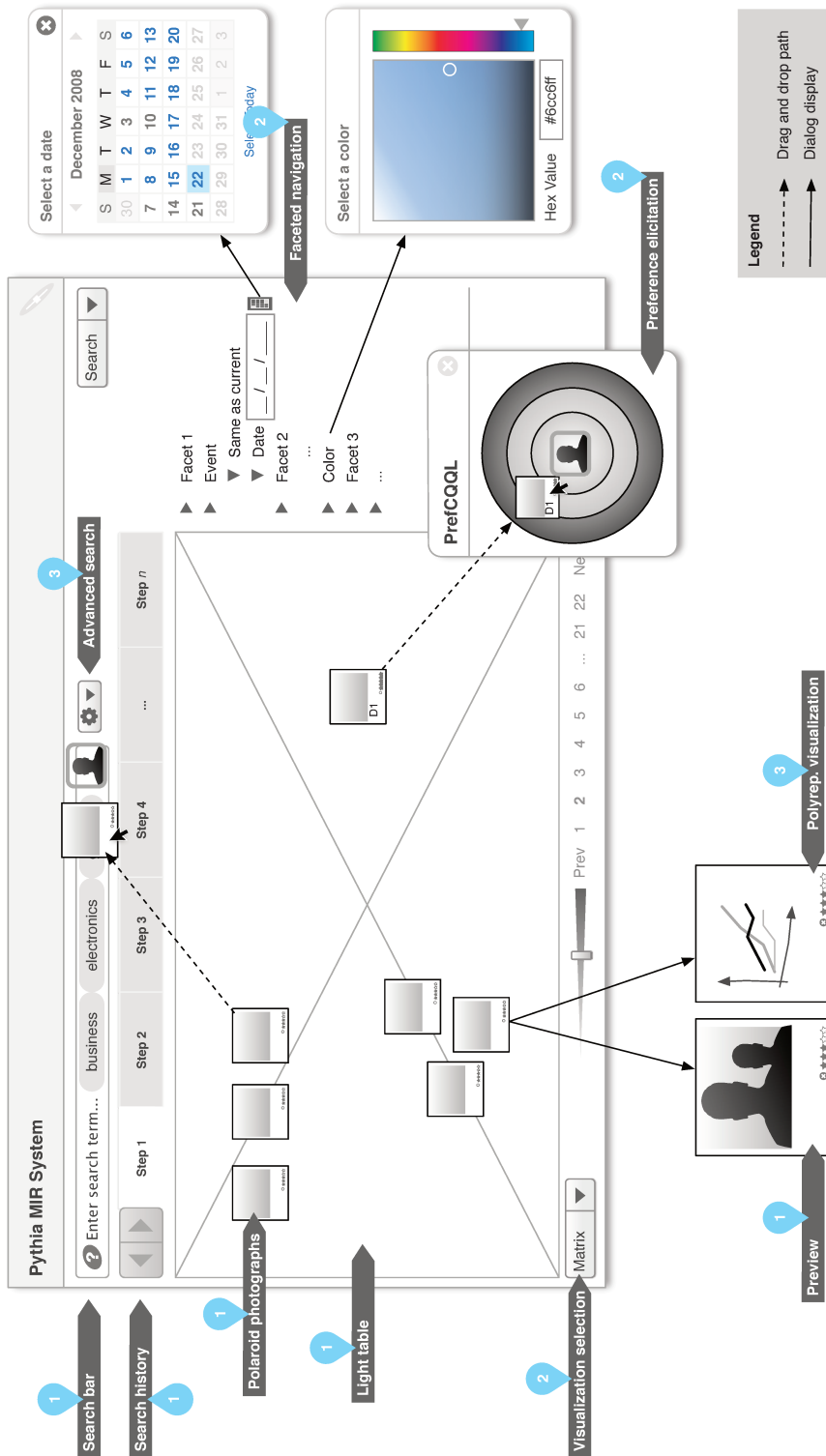


Figure 6.8: GUI mockup of the Pythia MIR system

## 6 Design of the Pythia MIR System Prototype

reflect the user's progress during the information seeking process. As such, the search history also offers a comfortable means to revert actions in case of errors.

Similarly to the search history, the visualization selector serves as a kind of shortcut to "sort" or rearrange the photographs on the light table. Technically, the selection list offers different visualization types for the results displayed on the light table (see Section 6.2.3). To support the sorting metaphor of photographs, changes between the different visualizations are animated to better communicate this change on the light table to the user.

When feasible, the user interface uses animations to convey information; however, animations are never used for decorative purposes in order to avoid detracting users from interacting with the system. For instance, if an action causes a reordering of the ranked result documents and thus the photographs on the light table, the affected photographs are smoothly moved to their new position on the table in order to ensure that the user becomes aware of the new positioning of the photographs.

One way to trigger the retrieval of new documents is to use PrefCQQL's preference-based RF mechanism. In order to elicit preferences, users can drag photographs to the concentric circles of the preference dialog described in Section 6.2.5. To remove preferences, the corresponding photographs can be dragged away from the widget. This mechanism is available for most widgets, e.g., QBE documents can be removed from the query by dragging them away.

Section 6.2.6 completes the description of Pythia MIR system's core user interface by addressing faceted navigation. In extension to the core functionalities, there are advanced search functions that will be described in Section 6.2.8 and visualizations dealing with the principle of polyrepresentation (see Section 6.2.9) that aim at bridging the gap between the user's model of the IN and the system's interpretation.

### **Universal usability within the Conceptual User Interface Model**

Shneiderman and Plaisant define *universal usability* as follows: "Different usage behaviors should be supported by the user interface, e.g., an expert user should be allowed to use keyboard short-cuts, while a layperson user can rely on a simple mouse-based interface." [Shneiderman & Plaisant 2005, cf. pp. 74]. Moreover, universal usability must not be understood as the process of stripping away features of a professional software system until it can be operated by untrained layperson users. Consequently, the cited example must not be considered closing because universal usability also means the support of different usage scenarios of personas as discussed in Section 6.1.1.

One way to achieve universal usability is the provision of a UI that adapts situationally towards the current user needs and offers an appropriate feature set. However, automatically initiated changes of the UI feature set have been shown to trouble users [Shneiderman 2003, cf. p. 2]. Furthermore, such automatisms are seen critical by the experts involved in the UCD process of the Pythia MIR system prototype as they limit the controllability of the system (see Section 6.1.1).

Alternatively, universal usability can be realized with the help of so-called *multi-layer interface designs* [Shneiderman 2003]. Roughly speaking, multi-layer interface design

summarizes the idea to give users control over the number of features that they are able to control and that are needed to solve their current work task. For instance, novice users interact with a system at layer 1 and control only a small amount of features. As the user's expertise rises over time, they can progress to higher layers with more complex feature sets. Regarding the design of the different layers, it is important that the layers are self-contained and allow the execution of useful action sequences which – in the optimal case – motivate users to explore the more advanced layers as well. An ubiquitous example of a simple multi-layer interface design in IR are the standard and advanced search functionalities in commercial Web search engines. Typically, lower layers are not only used by layperson or novice users. Instead, more advanced users also appreciate the usage of lower layers if they enable them to carry out frequently used operations with fewer work steps [Shneiderman 2003, cf. p. 5].

Besides the definition of self-contained interface layers, the multi-layer interface design approach poses an additional challenge for the UI designer: the switch between the layers. If the layer switch has to be triggered by the user, the resulting UI becomes more complex and can – in the worst case – overwhelm the user [Shneiderman 2003, cf. p. 3]. On the other hand, automatic layer switches are subject to the issues described above.

Thus, the Pythia MIR system prototype abandons explicit layer switches. Instead, all advanced functionality in the system is implemented as options. From a theoretical point of view, the Pythia MIR system's UI realizes an expanding multi-layer design with its core functionality on layer one. The subsequent layers 2 and 3 further expand the feature set but are not needed to initiate a simple retrieval. Figure 6.8 depicts the different UI elements and assigns them to the different interface layers indicated by numbers 1-3. Table 6.1 summarizes the findings and juxtaposes the primary targeted personas per layer and the supported main functionalities.

Table 6.1: Multi-layer interface elements of the Pythia MIR system

Layer	Targeted Personas	Contained Elements	Supported Functionality
1	Layperson	<ol style="list-style-type: none"> <li>1. Search bar</li> <li>2. Search history</li> <li>3. Core metaphors</li> <li>4. Preview</li> </ol>	<ol style="list-style-type: none"> <li>1. Directed and exploratory search</li> <li>2. Redo and undo</li> <li>3. Drag and drop input, DMP</li> <li>4. Full-size document preview</li> </ol>
2	Professional, Layperson	<ol style="list-style-type: none"> <li>1. Visualization selection</li> <li>2. Preference elicitation</li> <li>3. Faceted navigation</li> </ol>	<ol style="list-style-type: none"> <li>1. Different visualizations</li> <li>2. PrefCQQL relevance feedback</li> <li>3. Faceted navigation</li> </ol>
3	Researcher, Professional	<ol style="list-style-type: none"> <li>1. Advanced search</li> <li>2. Polyrepresentation visualization</li> </ol>	<ol style="list-style-type: none"> <li>1. Direct CQQL input</li> <li>2. Result explanation and weighting variables visualization</li> </ol>

Interestingly, Table 6.1 also illustrates the progression from elements of the user's

## 6 Design of the Pythia MIR System Prototype

cognitive space at layer 1 to the system's information space as discussed in Section 5.4.2. This transition becomes particularly obvious at layers 2 and 3. While layer 2 operates on preferences, layer 3 also supports the modification of CQQL's weighting variables that are clearly attributed to the information space as depicted in Figure 5.4.

### Conclusion

In order to realize the theoretical model from Section 6.2.1, this section presented a conceptual multi-layer user interface model that is based on the direct manipulation paradigm. Central to the model are two core metaphors, the Polaroid-like photograph and the light table, around which all possible user interactions are arranged. The reduction of the complete IIR process to two metaphors creates simplicity and consistency for the user. In conjunction with the DMP, the consistency of the user interaction is further increased because users can mainly operate the system with the help of drag and drop no matter if they want to alter a query, give relevance feedback, or just manually rearrange photographs on the light table.

Keeping in mind Hearst's suggestion to recognize the importance of small details [Hearst 2009, cf. p. 28], the prototype uses animations to convey such information as the change of the result rank to the user. Furthermore, the general aesthetics of the Pythia MIR system prototype is inspired by common digital photography tools in order to create a familiar user experience. In the opinion of the author, this approach is crucial in order to overcome potential reservations from the user side when they are confronted with novel search functionalities. Although not explicitly mentioned before, the user interface features common UI tropes that are meant to support the learnability of the system [Russell-Rose & Tate 2013], such as inviting textual labels or balloon help for most widgets.

Regarding the user needs of the different personas, the model with its three-layer interface design provides universal usability. Unlike other multi-layer interface designs, the Pythia MIR system does not rely on explicit layer switches. Instead, the different advanced feature sets are made available as options to the system's core functionality, i.e., directed and exploratory search. Already at layer 1, layperson and expert users can utilize directed and exploratory search functions. The following layers provide additional functions such as RF in order to personalize the initial results. Similar design decisions are also advocated by other authors because they encourage users to learn and to confront themselves with more advanced search strategies [Russell-Rose & Tate 2013, cf. pp. 105].

### 6.2.3 Result Visualization and Organization

Addressed user stories: 4, 5, 11, 12, 13, 16
--

The original task of information visualization is perhaps best summarized by Shneiderman's ubiquitous visualization mantra "overview first, zoom and filter, then details on

demand” [Shneiderman 1996]. In the scope of MIR, the information to be visualized is primarily the document result rank consisting of the potentially relevant documents. To be more precise, neither the internal document presentation of the MIR system nor the original document is displayed. Instead, a *document surrogate* that illustrates the most important features of the original document regarding the current multimodal search is used. As mentioned before, all result documents (or the corresponding document surrogates) of the Pythia MIR system are placed on the *light table* regardless of their generative origin, i.e., whether they are, for example, the result of a retrieval step or a filtering process. Hence, the light table forms one of the core direct manipulation metaphors. The second metaphor used for the document surrogate, the *Polaroid and Post-it metaphor*, is described below.

User Stories 4 and 11 explicitly ask for a tool that provides an *overview* over the collection’s content and leverages exploration. To assist the user during the exploration and the result inspection, the light table provides *zoom and pan* controls. For the sake of consistency, panning can also be carried out by dragging the light table in order to move the zoomed view on it.

The *filtering* of the results is supported at various levels as we will describe below. Additionally, the faceted navigation functionality, discussed separately in Section 6.2.6, can also be used to filter the results.

Common to all filter and ISS functionalities is that they affect the contents of the light table directly. In other words, users can interact with the result document surrogates as whole using a self-organizing or self-sorting light table: users can freely move their visual focus on the table, regroup and rearrange the documents, or retrieve a whole new set of result documents.

To address User Stories 5, 16 and 12, the light table can also be used to visualize relations between the documents, e.g., their dissimilarity or similarity. Although these visualizations do not have an impact on the underlying rank that is based on the probability of relevance (POR) of each document, they provide additional insights in the retrieved result. Obviously, such “magic” light table is not available in real life. Nevertheless, we believe that this advanced functionality does not break the light table metaphor as all operations could also be carried out in real life, although with much more effort by sorting the documents manually.

To further support the light table metaphor, all document surrogates on the table can be moved manually as in real life in order to assist users during their search (see User Story 13). For instance, users can rearrange document surrogates to reflect their own sorting categories, label them, or associate a set of document surrogates with a subjective grade of relevance.

To conclude, users can also inspect the documents’ *details on demand* by making use of previews that will be addressed in the next part of this dissertation. Furthermore, the Pythia MIR system also offers means to explain the origin of the results with the help of various tools presented in Section 6.2.9.

### Document Visualization – The Polaroid and Post-it Note Metaphor

As sketched before, the light table is one of the core metaphors of the Pythia MIR system’s direct manipulation UI. The contents of the light table correspond to the result rank generated by the MIR engine. The rank contains document representations, which are visualized with the help of document surrogates in the UI.

The quality and appearance of these document surrogates has a tremendous impact on the usability of the system. Based on the document surrogates, users judge the relevance of a document and decide whether they will inspect it in more detail [Hearst 2009, cf. Sec. 5.1]. Thus, the document surrogate has to visualize all characteristic properties that are relevant to the user’s IN. At the same time, the document surrogate has to be a compromise of the amount of displayed detail information and a waste of screen space (if more than one document surrogate is displayed at once) [Russell-Rose & Tate 2013]. For instance, Cunningham & Masoodian [2007] report on a user study that reveals a preference for a thumbnail display of images since they remain recognizable for a long time – even at small sizes. Unfortunately, the results of this study are not directly transferrable to the Pythia MIR system because our MIR system is multimodal. That is, it supports the retrieval of image documents on the basis of low-level and high-level features such as metadata, which has to be displayed as well by the document surrogate.

To meet this challenge, the UI uses a *Polaroid and Post-it note metaphor* (PPM) to visualize the retrieved documents. The metaphor is consistent with the light table metaphor as it is also borrowed from the world of photography. Moreover, Polaroid photographs feature a frame that has been used for taking notes in the real world and which can be used to accommodate notes or control elements in a computer UI<sup>104</sup>. Figure 6.9 illustrates the application of this metaphor by displaying a video clip, a PDF document, and a photograph with an accompanying tag (from left to right).

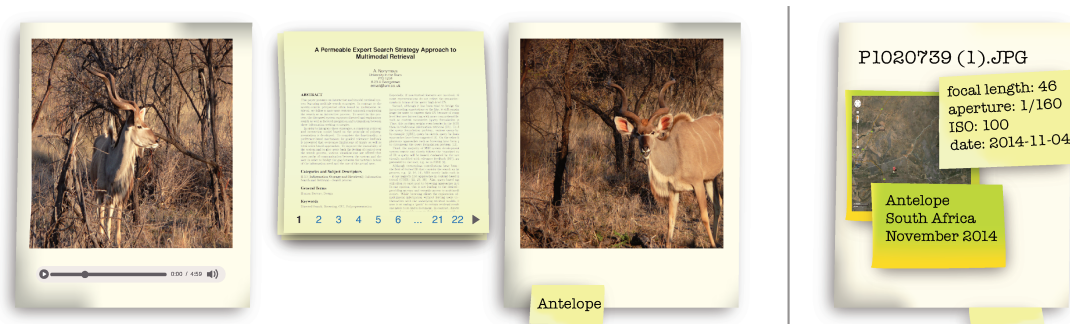


Figure 6.9: The Polaroid and Post-it note interface metaphor

It has to be noted that the metaphor does not have to be implemented in a skeuo-

<sup>104</sup>Alternatively, the metaphor could be based on the plastic frames of negatives and diapositives; however, this option has been discarded because of aesthetic reasons.



morph fashion as Figure 6.9 suggests. The skeuomorph design is for illustrative purposes only.

One advantage of the Polaroid and Post-it note metaphor is that it can be used to display different media types consistently. Although the current implementation of the Pythia MIR system is focussing mainly on multimodal CBIR, it can be easily extended to include videos or textual documents. For instance, in the case of a video the Polaroid's frame easily accommodates the typical video control elements, while a plain Post-it note can be used to display formatted text, including a navigational control widget (see Figure 6.9). Although text can also be visualized with the help of a Polaroid, we suggest to discriminate visually between pure visual and textual documents to provide additional guidance to the user. Consistent with the usage of Post-it notes for texts, the same metaphor reappears on the image document to visualize the tag "antelope". As the right-most Polaroid in Figure 6.9 illustrates, this metaphor can be taken further. Here, we see the "back side" of the aforementioned Polaroid featuring its GPS coordinate displayed by a map, some technical metadata (e.g., the focal length), and all tags that have been added to the document. Please note the two additional tags that appear only on the back of the Polaroid. This indicates that the initial query only contained a QBE document and a textual part containing the key word "antelope" but not the other tags. If the user would like to extend their initial query, e.g., with the key word "South Africa", they could simply drag it to the search bar (see Section 6.2.4).

The usage of the front and back side of the PPM points out an additional advantage of the metaphor. While the front side contains only the most important information in relation to the current IN, the back side features additional detail information. We believe that this is consistent with the real life handling of photograph prints, which often contain additional information, such as the names of the depicted persons, on the back. Thus, the UI supports the inspection of an image document during information seeking in a two step process: 1) getting an overview using the front side, and 2) getting the details on demand from the back side. Furthermore, a full-size preview of the original image is available in addition to visualizations explaining the results (see Section 6.2.9).

Besides the conceptual clarity of the PPM, it is also advantageous when multiple document surrogates are displayed on the light table at the same time and are therefore scaled down. Because of the relatively large content area of the PPM, thumbnails remain recognizable for a long time. This also applies for textual documents. Although their content might become unreadable, their layout impression remains visible. Hence, the document surrogates can still be used for the provision of an overview of the light table's contents. If a document surrogate attracts the attention of the user, they can zoom onto it to make additional information such as tags visible. From this perspective, the PPM is a reasonable and usable complement to the light table metaphor discussed in the next section.

### **Result Visualization – The Light Table Metaphor**

The *light table* serves as the container for the document surrogates and therefore relies on the PPM. Eventually, the light table displays the result of any action that has been

## 6 Design of the Pythia MIR System Prototype

initiated by the user, e.g., a RF iteration, a query resubmission, or the change of the current visualization of the document surrogates. At the same time, the light table is the central UI widget around which all other UI elements are arranged (see Figure 6.8).

Because of the light table's primary purpose to visualize the results of any retrieval process step, it addresses all of the user stories presented at the beginning of this section. To be more specific, its task is to project the linear result rank generated by the MIR model onto its two dimensional light table surface. Although the Pythia MIR system also supports a list-based visualization of the result rank, which can only be reached with the help of a menu, the light table does not support this kind of linear visualization as it would break its metaphor. Additionally, it can be doubted whether the display of the retrieval results in a list really fits the user needs – particularly in MIR. Although the list is the ubiquitous result visualization in IR, none of the presented user stories express a demand for lists. Besides not being very space effective, lists do neither support the visual emphasis of specific properties of documents very well, e.g., distinct discriminative factors in order to group documents as required by User Story 12, nor do they support users in getting an overview over images documents (see User Story 4). Hence, we assume that the provision of more complex result visualization layouts address typical user needs better than a simple result list. A similar viewpoint is taken by Santini [2012]. Moreover, there is evidence that users are able to scan large amounts of images quickly and select the most interesting ones [Hearst 2009, Sec. 12.2.2]. As a consequence, it is reasonable to give users the opportunity to inspect a large amount of image documents at the same time, i.e., to use the screen space efficiently. Please note that this decision also has an impact on the retrieval effectiveness evaluation of the system because typical effectiveness measures in IR assume the presence of a result list. These issues are addressed in Section 8.1.3.

In accordance with the arguments presented before, the light table only supports different two dimensional layouts which are consistent with the real-world arrangement of photographs on a two dimensional table surface. Although a computer UI could in principle also work in three dimensions, e.g., as shown by Nakazato & Huang [2001], we have decided to limit Pythia MIR system's UI to two dimensions for the sake of the consistency of the light table metaphor and to lower the UI's complexity. In case of a 3D visualization, users would have to navigate through three dimensions, which is far more complex than inspecting a 2D surface. At the same time, 3D visualizations are subject to problems that do not occur in 2D, such as the occlusion of objects due to the projection of the 3D visualization onto the 2D computer screen.

All visualization layouts are realized by applying the same metaphor: the sorting of the document surrogates on the light table. The "sorting" is initiated by the user who chooses one of the visualizations from the visualization selection drop-down list (see Figure 6.8). To further support the sorting metaphor, the movement of the document surrogates to new positions on the light table is animated<sup>105</sup>. Moreover, the animation supports users in keeping track of document surrogates that have been inspected before, thus avoiding user irritation. For the sake of consistency, the light table's control

---

<sup>105</sup>For an example, see the sample user interaction video from minute 00:50 on (see Appendix E).

elements (zoom, pan, fit contents to screen, and the visualization selection) are available for all available visualizations and placed beneath the light table for easy access. The same applies to the aforementioned actions that can be carried out with the document surrogates.

The most basic result visualization layout and thus the 2D counterpart of the classical result list is the *2D matrix visualization* that arranges all document surrogates according to their decreasing POR in Western writing direction, i.e., from left to right and top to bottom. This layout provides an overview of the document surrogates as suggested by User Story 4, which is further supported by the aforementioned zoom and pan functionality. Besides giving an overview, this layout does not support other requirements expressed by User Stories 11, 5, 12, and 16, whereas the latter three describe similar user needs (see Figure 6.6).

In order to address User Story 11 in particular, the light table supports a document surrogate layout based on a *self-organizing map* (SOM) (see Section 3.3.2). The SOM-based layout visualizes and arranges the document surrogates in a 2D map with different regions characterized by their relative similarity. Figure 6.10 depicts a SOM based on the documents' color layout representation<sup>106</sup>. The contour lines indicate regions that are relatively homogeneous. As Figure 6.10 illustrates, the SOM-based layout supports users in discovering the visual variance of the documents (see User Story 5). It also displays similarities and dissimilarities by placing similar documents in spatial proximity (see User Stories 12 and 16).



Figure 6.10: Self-organizing map visualization

<sup>106</sup>The color layout has to be seen as a representative representation as the following statements apply to all other representations as well. The color layout is the standard representation used by the cluster algorithms of the Pythia MIR system prototype.

## 6 Design of the Pythia MIR System Prototype

Alternatively, users can utilize a *k-medoid-based cluster visualization* (see Section 6.3.2) that groups documents by their similarity. Unlike in the SOM-based layout, all document surrogates are placed next to the document that forms the center of one cluster. The *k-medoid* cluster algorithm is a classical partitioning clustering technique which clusters the data into a set of  $k$  clusters, where  $k$  has to be set to the desired value. Consequently, the cluster visualization puts an emphasis on the similarity of the documents in a given cluster (see User Story 12), while the clusters are separated by their dissimilarity (see User Stories 5). In particular, the cluster-based layout supports users that search for unusual documents (see User Story 16) because such documents can be easily discovered in cluster with one or only a few members.

The latter two layout techniques support the overview of the content of the light table by grouping the documents on the basis of a common criterion. As mentioned in Section 6.1.4, such approaches have been shown to support users in getting an overview as well as in exploring an unfamiliar information space. In order to meet user expectations and to provide a familiar UI that is also known from Bing or Google's image search, Pythia MIR's UI also provides a classical 2D matrix display, which is also serves as the preset visualization layout to avoid user irritation.

To conclude, the light table is not limited to the visualization layouts presented before. It can be extended with other visualizations without breaking the discussed metaphors as long as a reasonable interaction with the visualization layout is possible using typical 2D actions such as zooming and panning.

### Manual Result Sorting

During longer search sessions or such that are meant to be readopted at a later point in time, user studies [Malone 1983; Nardi 1993] show that users tend to arrange documents into " 'piles' of information arranged by physical location as well as explicitly titled and logically arranged 'files' " [Malone 1983, p. 99]. These manual arrangements of documents help users to structure their work tasks and support the findability of documents. User Story 13 describes a similar need.

To address this user need, the document surrogates can be freely moved on the light table. Moreover, multiple document surrogates are selectable at once in order to arrange them automatically in a group (or "pile"). Sample arrangements consist of stacks or circular arrangements around a center document surrogate. These groups can be tagged by the user or associated with a relevance level to give relevance feedback (see Section 6.2.5).

Furthermore, the UI contains a *collection basket* that is meant to gather documents that are considered interesting by the user during the information seeking process (see Figure 6.20; 13).

Additional sorting criteria, such as the creation time of a document or other metadata, have not been implemented but can be added to the matrix layout without problems because the sorting on a criterion is a trivial programming task. For the other visualization layouts, sorting is a more complex endeavor. In this case, the (sorting) criteria are only useful in form of filters as the actual visual arrangement of the document surro-

gates is defined by the clustering or neural network, which cannot be altered afterwards without destroying the visualizations consistency. However, these criteria can be used for the similarity calculations of the SOM or for clustering to achieve similar results to the sorting of the 2D matrix layout.

### 6.2.4 Directed and Exploratory Search

Addressed user stories: 8, 9, 10, 15, 17

The importance of the acknowledgment of the information need's dynamics has been stressed throughout this thesis. In particular, Section 6.2.1 underlined that multiple ISS have to be offered by a MIR system in order to support the user's different IN states (see Figure 6.4). This theoretical finding is also supported by User Story 10. As outlined in Section 6.2.1, these ISS have to be supported at once. Furthermore, users have to be enabled to freely move between them to satisfy their current dynamic IN (see User Story 9).

This section predominantly focusses on two basic ISS: directed and exploratory search. Additional ISS such as relevance feedback or faceted navigation are addressed in the subsequent sections.

#### Directed Search

Unsurprisingly, the directed search component is designed around CQQL capabilities as a multimodal query language (see User Story 17) in MIR discussed in Section 4.4. Because of the focus of this dissertation, this section will concentrate on directed search in form of *query by example* (QBE, see User Story 8) and largely omit textual query processing. Nevertheless, textual query input is possible and its processing is driven by the Indri textual IR engine [Strohman et al. 2004], which calculates evaluation results for atomic CQQL conditions present in the underlying CQQL query. This approach resembles the utilization of the similarity calculations that have been presented in Section 4.4 in order to incorporate CBIR-based features.

Figure 6.11 depicts the main UI component to specify and control a directed search: the *search bar*. As before, the layout follows Western writing direction, guiding the user through the necessary steps to start the retrieval process.



Figure 6.11: Search bar

## 6 Design of the Pythia MIR System Prototype

The switch on the left side of the search bar (see Figure 6.11; 1) sets the UI in either the directed or the exploratory search mode (see below) and indicates the type of the current ISS.

The second element allows access to the advanced search parameters, e.g., the setting of weight variables or direct access to CQQL.

The text input field (3) accepts textual keyword queries that can be extended with Indri commands. The multimedia query documents box (4) is the MIR counterpart of the text input field. It accepts QBE documents that can be chosen from an OS-provided file selection dialog or dragged from the file system. Furthermore, users can drag document surrogates displayed on the light table onto the box to expand their initial query or to submit a new one. The multimedia query documents box expands automatically if all slots are occupied. To remove QBE documents, they can be simply dragged away from the box to remove them from the query. Additionally, QBE documents can also serve as negative examples with the help of CQQL's negation.

The fifth element limits the number of retrieved results, while the last button (6) is used to start the retrieval process.

The search bar follows the multi-layer interface design paradigm presented before by hiding optional, advanced controls to the user. In principle, a simple directed search can be initiated by dragging a QBE document (or typing in a keyword) onto the search bar followed by pressing the start button. The results can then, e.g., be modified with the help of RF (see Section 6.2.5); alternatively, the initial query can be altered by dragging a result document onto the search bar to resubmit a new query.

Although the current UI accepts only the specification of queries in textual and QBE form, it can easily be augmented with a query by sketch component (see Section 3.3.1). For instance, the needed drawing tools can be accessed via an icon or the contextual menu of the multimedia query documents box. An implementation has been omitted because of the additional complexity of this functionality and its missing relevance for the evaluation of the Pythia MIR system prototype's UI<sup>107</sup>.

The advanced search icon (see Figure 6.11; 2) gives access to interface layer 3 that allows the direct modification of the underlying CQQL query. This functionality enables professional users to state their IN with the full expressive power of a relational complete query language (see Section 4.1).

To conclude, User Story 15 and user studies, e.g., by McDonald & Tait [2003], reveal the need to refind once retrieved documents using directed searches. Although basic refindability during one search session is supported by the Pythia MIR system with the help of a search history (see Section 6.2.7) and the collection basket (see Section 6.2.3), search-session spanning refindability has to be supported by other means. As elaborated in Section 6.2.1, CQQL queries can be stored at every stage of the information seeking process. When restored, the saved query and the corresponding metadata, such as the search session (see Section 6.2.7), is used to set up the search bar accordingly

---

<sup>107</sup>Internally, sketched images or images provided as QBE documents are both treated the same as the needed feature extraction is carried out on their bitmap-based representation. Hence, the inclusion of a sketching facility would only measure the usability of this component but not give further insights into the usability of the overall IIR process.

to resume a past search. Hence, the loading of a past search is delivered with the same UI mechanisms as a newly started search.

### Exploratory Search

Basic exploratory search functionality in the UI is offered by the different result visualization layouts such as the SOM or the cluster-based visualization of the light table's contents (see Section 6.2.3). However, these result visualizations are limited to the results of a retrieval step. Thus, their application requires a preceding directed search.

In order to support users with a variable IN (see Figure 6.4) that do not provide a query to the system, the Pythia MIR system's UI offers a distinct browsing component. Analogously to the directed search, browsing is initiated with the help of the search bar. By choosing to browse, users can retrieve all documents from the current collection at once without the explicit need for a query – although, as described in Section 6.2.1, it might be internally prepared to be used later (see below). In order to facilitate the navigation in the potentially large collection and to give an overview of its contents, the UI supports users by visualizing the collection's content with the help of a  $k$ -medoid cluster algorithm equal to the one described above.

To simplify the exploration of the collection, the light table is divided into two parts during browsing. On its left, the cluster centers, i.e., the documents that are characteristic for a cluster, are displayed. The larger part of the light table is synchronized with the currently inspected cluster center and displays the documents contained in the respective cluster. Figure 6.21 depicts the UI in browsing mode with the contents of the highlighted cluster displayed at the light table in matrix layout. Since the result visualization during browsing is also based on the light table metaphor, the same result layouts and mechanisms as before are available to the user to further explore the content of a cluster. As described in the theoretical polyrepresentative user interaction model (see Section 6.2.1), the user can also drag document surrogates from the light table to the search bar in order to switch to a directed search if a document is found to represent the current IN.

An alternative approach for the browsing in a collection's content is based on an exploitation of a particular characteristic of CQQL: the setting of all weighting variables to 0 effectively disables the calculation of a rank of relevant documents. That is, all documents in the collection obtain the same POR and are therefore retrieved unordered (see Section 6.2.1). In contrast to the approach discussed before, which can be solely based on a visualization technique, this approach maintains an existing CQQL query. The unsorted results can then be explored with the help of the visualizations provided by the light table. If the user decides later to use a directed search or faceted navigation (see Section 6.2.6), this query can be directly used to form a CO. Alternatively, users can further explore the information space along one or more representations by using the polyrepresentation visualization dialog (see Section 6.2.9) or by directly modifying the weighting variables of the underlying CQQL query (see Section 6.2.8).

One further application area in which CQQL can be used to assist during the exploration of a collection's content is *polythetic clustering*. Here, the CQQL query that repre-

## 6 Design of the Pythia MIR System Prototype

sents the CO is used to calculate the difference between all documents in the collection. This approach is essentially different from a directed search, in which each document's POR is calculated with respect to the CO. A polythetic clustered visualization enables users to get new insights into a collection. For instance, highly relevant documents that lie outside the current CO and are therefore not contained in the retrieved result rank are placed in different clusters that are characterized by their inter-document similarity. The inspection of these clusters might lead to a modification of the initial CO, e.g., by using a disjunction of some relevant CO that are represented by different polythetic clusters to cover completely different parts of the information space with a new query. Figure 6.12 shows a mockup of CQQL-based polythetic clustering where each enlarged document surrogate depicts a cluster center surrounded by the cluster's associated documents [Zellhöfer 2012a, cf. p. 66].

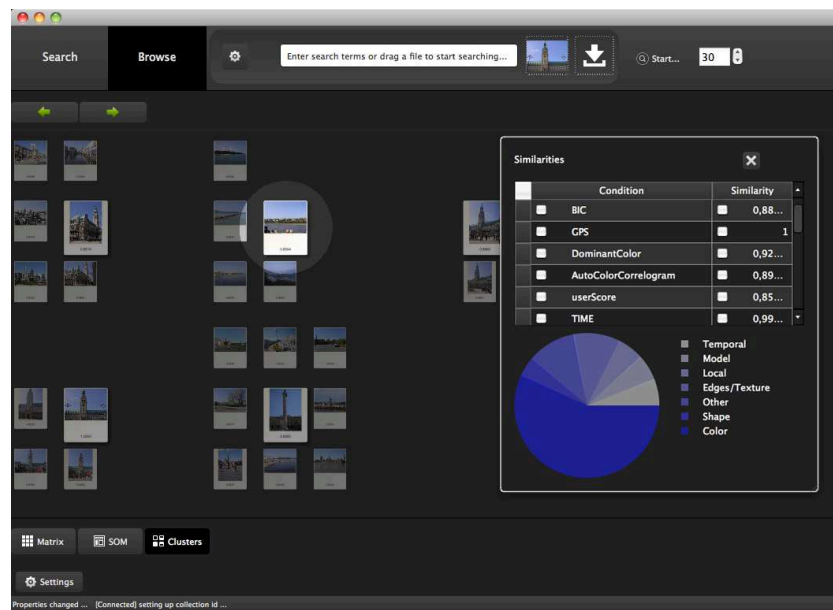


Figure 6.12: Mockup of CQQL-based polythetic clustering

To conclude, automated query learning functionalities and the related PrefCQQL-based indicators that have been outlined in Section 6.2.1 are not supported in the current UI because of the users' specific demand for a fully controllable MIR system (see Section 6.1.1).

### 6.2.5 PrefCQQL – Preference Elicitation and Relevance Feedback

Addressed user stories: 1, 2, 9

Section 5.4 presented the preference-based RF approach PrefCQQL. Relevance feedback as a common personalization technique in IR that addresses the dynamics of the user's



IN (see User Story 9) has been covered in detail in Section 5.1.

User Story 2 explicitly demands a controllable RF mechanism to refine the retrieved results. Furthermore, the users expect this mechanism to adapt fast and to work intuitively (see User Story 1). The effectiveness of PrefCQQL's adaption towards the user's IN is addressed in Section 8.5. This section focusses on the provision of a controllable PrefCQQL-based RF mechanism that uses the DMP to remain consistent with the metaphors presented in Section 6.2.3.

The main problem with RF is that users have to be motivated to give it at first place. Hence, the RF mechanism has to be usable [Ruthven & Lalmas 2003], consistently embedded into the other user interactions, comprehensible, controllable, and easily executable. Section 5.4.1 argued for the utilization of inductive preferences in order to improve the comprehensibility of the RF mechanism.

### Interpretation of the User Input

Section 5.4.3 showed how preferences can be interpreted as graphs and how such graphs can be built or modified on basis on a result rank of retrieved documents  $\mathcal{R}$ . However, from a usability perspective, it is also important to consider how documents that do not participate in a user-defined preference relation but that are displayed on the light table are interpreted. Basically, there are two ways to interpret the roles of the retrieved documents within the preference graph [Zellhöfer & Schmitt 2010a].

*First*, the list semantics of the result rank is neglected in favor of the user input. That is, the user choses a small group of documents on which preferences are defined. Documents that are not involved in any form of user interaction are ignored and are not added to the preference graph. Because of the user's disinterest in extensive interaction with a system, it can be assumed that the latter group of documents will form the majority [Shneiderman & Plaisant 2005]. In other words, the user will only interact with a few preferences, which form the exclusive source of information regarding the user's current preferences.

*Second*, the list semantics of the original rank  $\mathcal{R}$  is maintained and serves as additional information regarding the user's preferences. Given a list of retrieved documents ordered by their POR, the user only defines preferences between the documents that do not conform with the expected rank, which would reflect the current IN. That is, a more relevant document succeeding a less relevant document would be stated in form of a preference expressing the inverted positioning. To give another example, the most relevant document would be preferred over all other documents in  $\mathcal{R}$ . The main difference from the first approach is that documents which have not been touched by the user keep their position in the result rank and are accordingly included in the preference relation graph. That is, documents that have not been interacted with are interpreted as preferences in conjunction with the explicitly stated ones.

To better illustrate the difference between the two approaches, consider the following retrieved rank:  $\mathcal{R} = \{d_2, d_3, d_1, d_5, d_4\}$ . The optimal rank would be ordered by the document indices. To give RF, the user states one preference  $P = \{d_1 \succeq d_2\}$ .

## 6 Design of the Pythia MIR System Prototype

In the first approach, only the preference  $P$  is passed to PrefCQQL'S learning algorithm that learns a weighting scheme resulting in  $\mathcal{R}' = \{d_1, d_2, d_3, d_4, d_5\}$  because the weighting scheme also causes a reordering of the last two documents.

The second approach maintains the list semantics of  $\mathcal{R}$  and extends the preferences accordingly:  $P' = \{d_1 \succeq d_2 \succeq d_3 \succeq d_5 \succeq d_4\}$ . As a result, the weight learning algorithm yields  $\mathcal{R}' = \{d_1, d_2, d_3, d_5, d_4\}$ .

Although this example results in a bad outcome for the second approach as it also satisfies one unwanted preference  $d_5 \succeq d_4$ , it can also be argued that it can in some cases provide additional valuable information to the learning algorithm, e.g.,  $d_2 \succeq d_3$ . Nevertheless, we propose the first alternative, which takes into account only those documents that have been part of the user interaction. Moreover, it can be assumed that users, who try to avoid superfluous interactions, will not state preferences amongst documents they are not interested in. An automatic augmentation of the user-defined preferences would add potentially unwanted preferences between documents the user is not interested in. Furthermore, users would be forced to manually remove such preferences in the second approach in order to avoid unwanted preferences such as  $d_5 \succeq d_4$ . This additional work might easily overstrain users if large sets of preferences have to be inspected at every RF iteration. Roughly speaking, the second alternative would reward highly placed results using a mechanism similar to pseudo relevance feedback. We believe that this automatism is likely to irritate users as they can only gain full control over their preference input if they define preferences over the complete retrieved rank. In addition, depending on the size of the result rank, the second approach might create a huge amount of preferences and therefore restricts the parameter space of the learning algorithm too much. As a result, no or only a few new documents would appear in the new rank  $\mathcal{R}'$  after one learning step because the learning algorithm is bound to satisfy the preferences that almost perfectly describe the initial rank  $\mathcal{R}$ . This concern is strengthened by the experimental results of a supervised thesis, which reports none or little change in the top- $k$  results during RF when the second approach is used [Käppler 2009]. In contrast, the first approach leaves enough freedom for the learning algorithm to effectively change the result rank during the RF iteration as Section 8.5 will elaborate.

In conclusion, the first approach is to be preferred due to two main reasons. *First*, it is intuitively comprehensible because only user-defined preferences serve as input to the learning algorithm. Documents that have not been involved in any user interaction do not participate in any preference relation and therefore obtain an "unknown" level of relevance. Additionally, the first approach is more transparent than the second particularly if the user works with result visualizations that do not map directly to the linear result rank, e.g., the SOM-based visualization (see Section 6.2.3). In this case, the deduction of the additional preferences in the second approach is hardly predictable for the user as they cannot conclude the automatically generated preferences from the result visualization.

*Second*, the first approach leaves enough "space" for the learning algorithm to find a weighting scheme that supposedly affects the new result rank  $\mathcal{R}'$  more notable than when the second one would be used. As a result, it becomes more likely that new

documents are presented to the user which are more relevant than the ones contained in the initial rank  $\mathcal{R}$ .

### Preference Elicitation and Conflict Handling

Section 5.4.7 described two main problems of the PrefCQQL approach during preference processing: preference conflicts and query incompatibilities. The mentioned section outlined that the preference elicitation process is potentially error-prone because preference conflicts can be easily input – in particular if complex preference graphs are constructed during the interaction with the system. This situation is problematic from a usability perspective as the chance of providing erroneous input should be generally minimized because error-tolerant or error-avoiding interfaces improve the user satisfaction and the ease of use of a system (see User Story 14).

Obviously, the error-proneness is due to the complexity of the modification of the preference graph. Thus, the complexity has to be lowered to enable users to elicit preferences easily (see User Story 1) – at least at the lower interface layers. Another factor that affects the complexity of the preference elicitation is the actual amount of actions that have to be carried out by the user to define a preference. For the sake of consistency, we suggest to base the preference elicitation on the document surrogates and the corresponding drag and drop functionality. That is, users can state preferences by directly interacting with the document surrogates respecting the DMP.

To limit the complexity of the preference elicitation, the UI makes use of the *concentric circles metaphor* (CCM). As the name suggests, the CCM consists of a number of concentric circles placed around a center. The innermost circle contains all stated QBE documents (if the query contains such documents) to illustrate that these documents form the currently best known answer to the IN (see Section 6.2.1). The innermost circle is surrounded by three concentric circles (or rings) that visualize the decreasing levels of relevance, which can be expressed with the help of preferences<sup>108</sup>. In order to state a minimal preference, at least two documents have to be placed on two different rings. For instance, a document  $d_1$  that is placed on the first ring is preferred to a document  $d_2$  on the second ring etc. Figure 6.13 illustrates the principle.

The elicitation of preferences using concentric circles is simple: users can drag document surrogates from the light table to the ring reflecting the desired level of relevance. After the document is dropped, a mirrored copy of the document surrogate is placed on the ring and the original document surrogate remains on the light table to avoid confusion. In order to remove a document from a preference, it can be dragged away. To modify preferences, document surrogates can be freely moved between the rings. To avoid ambiguous input, each document surrogate can only be dragged once to the concentric circles. If it is dragged again, it will replace the document surrogate that has been placed before at one ring. This prevents users from stating preferences such as  $d_1 \succeq d_1$  or  $d_1 \succeq d_2 \succeq d_1$ . While the first preference is obviously pointless, the latter one

<sup>108</sup>The limitation to three circles has been chosen deliberately as a compromise between complexity and expressive power. In principle, any number of concentric circles is possible.

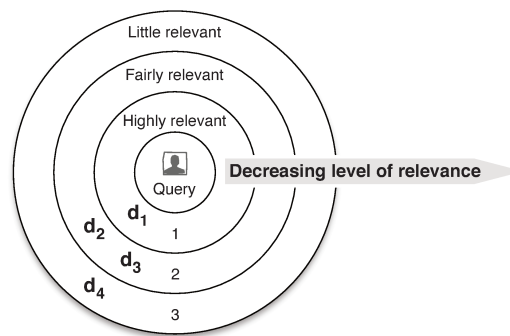


Figure 6.13: Concentric circles for preference elicitation

requires a little explanation. Let  $d_1$  be placed at ring 1,  $d_2$  on ring 2, and  $d_1$  again on ring 3. Following the explanation from above, the increasing number of a ring's index corresponds to its decreasing level of relevance. That is, a user would expect to have stated  $d_1 \succ d_2 \succ d_1$ . Although this preference is obviously contradictory, it would be interpreted internally by PrefCQQL as  $d_1 \succeq d_2 \succeq d_1$ . Furthermore, the prevention of multiple instances of a document within the concentric circles also circumvents the input of preference conflicts caused by cycles (see Section 5.4.7). However, this CCM approach also limits the expressive power of preferences that are stated with its help. For instance, if two documents  $d_2$  and  $d_3$  are placed on one ring and a third document  $d_4$  on the next outer ring (see Figure 6.13), the system must conclude that  $d_2 \succeq d_4$  and  $d_3 \succeq d_4$  holds. Moreover, the number of concentric circles limits the “depth” of preferences, i.e., users cannot state preferences such as  $d_4 \succeq d_i$  if the example given in Figure 6.13 is followed. For more details about Pythia MIR system's interpretation of preferences stated with the help of the CCM refer to the sources available as a supplement to this dissertation<sup>109</sup>. Nevertheless, the usability of an early implementation of the CCM for preferences has been examined in a supervised thesis by Böckelmann [2009] and is further investigated in Chapter 9.

For the sake of completeness, a similar approach to the CCM, presented in Blanken et al. [2007, Chap.11], has to be mentioned. The presentation by Blanken et al. [2007] only sketches the idea to visualize weighted RF using different rings but leaves the actual implementation completely open.

In order to communicate irrelevant documents to the system, the user is offered two different ways. First, as described in Section 6.2.4, users can state negative QBE documents. Second, irrelevant documents can be expressed in form of a preference. We interpret irrelevant documents as being less preferred than the last document in the retrieved rank. Due to CQQL's evaluation, the retrieval engine knows the POR of each document in the collection (see Section 4.4). Hence, the documents with the lowest POR  $D_{low}$  are also known. To express irrelevance in form of a preference that can be passed

<sup>109</sup>The implementation in C++ is available in `dbis::gui::preference::PreferencePlugin::poset()`.

to PrefCQQL's learning algorithm, each document  $d_i$  in the set of irrelevant documents  $D_{irrelevant}$  is given less relevance than each document  $d_l \in D_{low}$ , i.e.,

$$\forall d_l \in D_{low}, d_i \in D_{irrelevant} \mid d_l \succeq d_i.$$

These preferences can be automatically constructed if  $D_{irrelevant}$  is known, e.g., if these documents are indicated by the user. To support the simple elicitation of irrelevant documents, the UI makes use of the *trash bin metaphor*. In analogy to most modern operating systems, document surrogates can be dragged to the trash bin to mark them as irrelevant. To remove them, they can also be dragged out of the trash bin. As described above, documents in the trash bin are interpreted automatically and passed to PrefCQQL's learning algorithm along with the other explicitly stated preferences. Although this mechanism involves some automatism, we believe that it will not irritate users as it expresses a notion of irrelevance.

To improve visual clarity, the fact whether a document is participating in a preference relation is indicated by an icon at the frame of each document surrogate's frame. By default, irrelevant documents are hidden from the light table to make more room for potentially relevant documents. In order to inspect the irrelevant documents, the trash bin features a distinct light table in a separate dialog, with which the user can interact in the same way as with the main light table.

To conclude, the CCM avoids the input of different conflicting or erroneous preferences, while it provides a comprehensible visualization of stated preferences. However, it cannot prevent query incompatibilities as they are caused by a combination of the used matching function and the stated preferences (see Section 5.4.4). Thus, errors due to query incompatibilities still have to be communicated to the user for further decisions, e.g., whether a new matching function should be used or a preference has to be removed.

To recapitulate, the CCM in conjunction with the trash bin offers a simple, controllable and consistently usable means for preference elicitation as demanded by User Story 2. Its controllability is further increased by a search history functionality that allows to undo and redo recent actions (see Section 6.2.7).

Because preferences form an additional information about the user's IN and therefore are not directly part of the retrieval result, the UI separates the concentric circles UI component from the light table. Due to their relation to the personalization of retrieval results, the CCM component is placed in a separate personalization dialog, which also contains the trash bin, the collection basket (see Section 6.2.3), and the faceted navigation (see Section 6.2.6). Figure 6.14 depicts the personalization dialog and illustrates how document surrogates can be dragged from the light table to the concentric circles in order to elicit preferences.

As outlined in Section 5.5, preferences can also be used to start a new directed search. This functionality could be offered at the same interface level as the preference elicitation because it is equally complex.

Advanced preference handling that allows the utilization of PrefCQQL's expressive power can be realized using Hasse diagrams and the corresponding interactions on

## 6 Design of the Pythia MIR System Prototype

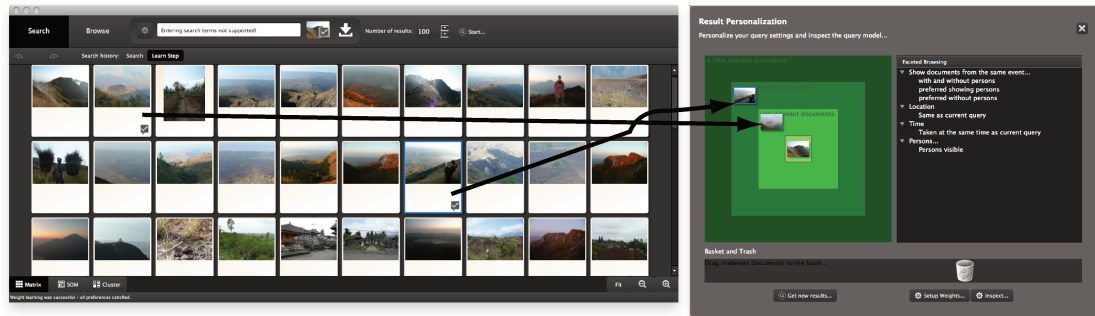


Figure 6.14: Preference elicitation per drag and drop and visual relevance feedback

such graphs (see Section 5.4.3). As the modification of Hasse diagrams is a complex and potentially error-prone task, we suggest to offer this functionality at the third user interface layer. If this functionality is provided to the user, the CCM has to be extended by some sort of clearly visible indicator that shows if a stated preference graph is too complex to be displayed with the help of the CCM. In this case, the CCM can only visualize an approximation of the complex preference graph, which might irritate users as the visualization at interface layer two is not fully consistent with the system's internal model of the IN.

### 6.2.6 Faceted Navigation

Addressed user stories: 1, 3, 6, 9

Section 3.3.2 reported that faceted navigation (FN) is frequently combined with keyword search. In this case, the keyword search acts as a filter on the retrieved documents that are the explored with the help of FN. The Pythia MIR system extends this approach by including QBE documents and results of the collection's exploration. In the latter case, the last inspected document serves automatically as a point of reference for the FN if needed. That is, the last inspected document becomes part of the query in order to enable specific facets such as location- or event-specific facets. As such, FN supports the dynamic IN of users (see User Story 9).

Figure 6.8 illustrates a sample FN widget. For the sake of clarity, the facets are thematically grouped. Each facet group can contain an arbitrary number of facets that can be used for the exploration of the collection's content. While facets such as color do not need a reference in form of a QBE document (or an inspected, browsed document), advanced facets, such as event-related ones, need a reference in order to determine if documents describe the same event as the query. Internally, the affiliation of a document to an event is modeled with the help of the creation date and time as well as the GPS coordinate (see Matching Function 12). This exemplary facet will also be used

during the evaluation of the user experience (see Section 9.1.5). Alternative facets that are implemented consist of documents taken at the same event with or without persons, temporal or spatial proximity, or documents that depict persons. Although facets could be created automatically, the Pythia MIR system's UI only supports manually created facets, of which many are directly related to the used matching function. This limits the flexibility of the retrieval engine but ensures that the facets are comprehensible to the user [White & Roth 2009] making them an intuitively comprehensible means to adapt the MIR system to the user's IN (see User Story 1).

Whenever a user chooses a facet, the newly retrieved results appear on the light table in consistence with the sorting/filter semantics presented before. Alternatively, FN can be used in the sense of a filter on the already retrieved documents (see User Story 3). In this scenario, which is currently not supported by the prototype, the facets act as fuzzy filters as demanded in Section 6.1.1. Here, the weighting scheme corresponding to the given facet is re-evaluated on the retrieved documents, yielding a new rank that can be presented to the user. Because of CQQL's best matching semantics, this filter operation is least likely to return an empty result set, which, as postulated in User Story 6, has to be avoided.

For the sake of simplicity, all facets in the Pythia MIR system prototype are mutually exclusive. This limitation is not due to technical restrictions. Instead, it is meant to lower the complexity of the UI and to make the results more predictable for the user.

### 6.2.7 Search History

Addressed user stories: 2, 4, 15
----------------------------------

The search history's primary purpose is to simplify the reversibility of user actions and to improve the controllability of the system. The need for these features has been expressed, amongst others, by User Story 2. It provides the central undo and redo functionality and allows users to navigate between different stages of their current search session in order to correct mistakes or to re-find documents (see User Story 15). Furthermore, the search history visualizes the current state of the search session. The need for undo and redo functionalities has been shown by different user studies, e.g., by Cribbin & Chen [2001] or Liu et al. [2009].

The search history of the Pythia MIR system updates itself after every ISS change or RF iteration automatically. Internally, the state of the search session is saved in form of the CQQL query and the corresponding weighting scheme. Additionally, the type of the current ISS is saved and displayed in the search history widget in order to communicate the "navigational state" [Russell-Rose & Tate 2013, p. 197] of the search session. This approach is user-friendlier than simply saving the search history in form of enumerated search steps because it lowers the chance to confuse users as the changes between the ISS are clearly indicated and can therefore be recognized easier. This is in particular true if users skip several prior search session steps in order to resume their

## 6 Design of the Pythia MIR System Prototype

search from a more distant search stage. If the search history is only used to correct erroneous input, the verbose labeling of the search steps is not expected to improve the usability significantly. Figure 6.15 shows the search history widget with the undo, redo, and search history buttons (from left to right).

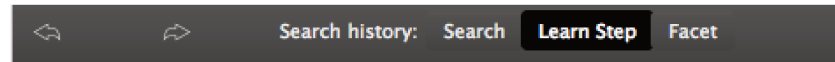


Figure 6.15: Linear search history widget

The search history for the Pythia MIR system is designed in a linear fashion. That is, a user can only navigate through the stages of the search session backwards or forwards to restore the state of the light table at the selected time. If the user changes the ISS from the current search stage (as indicated by the black background in Figure 6.15), all subsequent search stages are overridden. Although a branching search history is technically possible, we have decided against it because linear histories are common in nowadays software systems. Furthermore, a branching search history adds complexity to the UI, which might easily overstrain users. Nevertheless, a branching search history is an interesting concept as it enables user to utilize the search history to explore a collection's content because it encourages trial and error or browsing search strategies. Figure 6.16 illustrates a branching search history. While in a linear search history there must be only one step 2 or 3, a branching search history enables users to utilize multiple search strategies from different stages of their search session on. For instance, a user might explore the information space in step 2a and 3a, then jump back to step 1, and restart with a directed search resulting in 2b etc.

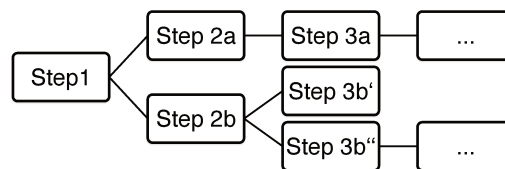


Figure 6.16: Branching search history

Although a basic variant of this search history-supported browsing is also possible with the linear search history in order to satisfy User Story 4, the approach is less flexible than a branching search history as the change of the ISS in a past search step will override all subsequent steps in the history eventually erasing all redo opportunities.

To conclude, the search history can also be easily extended to include prior search sessions in order to support the findability of documents that have been retrieved before (see User Story 15). The inclusion of prior search session requires only the storage of the stated query, of the corresponding weighting schemes, and of the metadata that describes the search session.



### 6.2.8 Manual Weight Setting – Advanced Search

Addressed user stories: 1, 9, 18, 20

The manual weight setting is clearly an expert functionality as it directly operates on the weighting variables of the CQQL-based matching function. It is expected to be used only by the professional and the researcher persona (see User Stories 18 and 20) and is not mentioned by any of the layperson's user stories. Consequently, it is placed at the third interface layer and can also be operated by laypersons if enough time is spent on training. The manual weight setting is inseparably linked to the result explanation functionality presented in Section 6.2.9.

The manual weight setting as part of the advanced search enables users to directly modify the weighting variables' values using a graphical dialog (see Figure 6.20; 15<sup>110</sup>). Because of the direct effects of these modifications on the retrieval system as required in Section 6.1.1, this functionality also addresses User Stories 1 and 9.

### 6.2.9 Result Explanation – Advanced Search

Addressed user stories: 19, 20

The result explanation components are inseparably linked to the manual weight setting described in Section 6.2.8. Eventually, the result explanation components are a means to bridge the gap between the user's cognitive and the system's informative space as requested by the global model of polyrepresentation (see Section 6.2.1). In other words, the explanation components are meant to allow an adjustment of the user's mental model with the system's conceptual model of the current IN. Being expert functions, the result explanation components, which consist of the *polyrepresentation visualization* and the *weight inspector*, are placed on the third interface layer (see Figure 6.8). The weight inspector is part of the advanced search functionality of the UI.

**Polyrepresentation Visualization** The *polyrepresentation visualization's* main purpose is to illustrate the similarity of an inspected document regarding its representations with respect to the current query, e.g., a QBE document. This component is motivated by studies revealing frequent user misconceptions about how the IR system determines the relevance or similarity of a document with respect to a query. For instance, Urban et al. show that users typically "could not distinguish between images that have the same color and images that are generally similar (in terms of semantic content, layout, color, etc.)" [Urban et al. 2006, p. 26]. Furthermore, the component supports

<sup>110</sup>Please note that the dialog displays only weighting variables at the same structural level, e.g., as present in the weighted conjunction (see Matching Function 2). Higher structured queries require more complex interfaces such a tree-based visualizations of the weighting variables as shown by Schmitt [2007].

## 6 Design of the Pythia MIR System Prototype

users in understanding the internal representations of the MIR system. Figure 6.17 shows an exemplary polyrepresentation visualization dialog. Besides listing the POR (labeled with “similarity”) of each representation (labeled with “condition”), the dialog also visualizes the impact on the calculation of the document’s overall POR (labeled with “userScore”) per thematic representation group (e.g., color or shape) with a pie chart. The thematic groups are meant to simplify the UI by giving user insights into the general characteristics of the inspected document that lead to its current relevance judgement. In the given example, the similarity judgement mainly depends on color-based representations. As a consequence, one can conclude a visual similarity between the inspected document and the QBE document which is primarily determined by the common color palette.

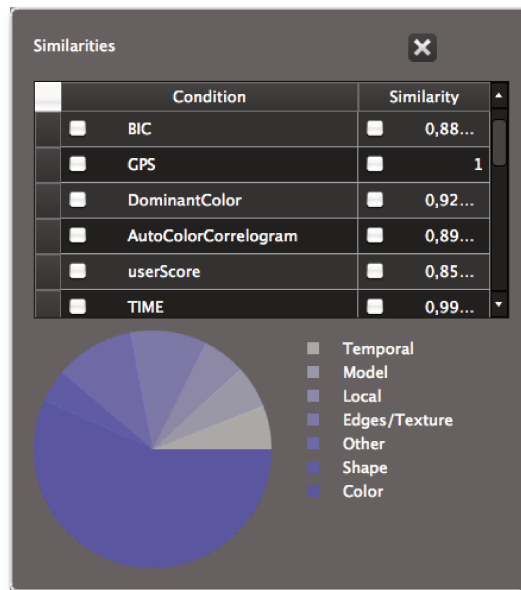


Figure 6.17: Polyrepresentation visualization

By inspecting the POR of each representation, users can decide to explore the information space along one or more representations by clicking on them or to learn about the MIR system’s model of their IN. In conjunction with the manual weight setting functionality, users are further enabled to modify the underlying CQQL query representing their IN according to their needs. For instance, they might learn that their current IN is best represented by the similarities regarding the creation time of a document and its GPS coordinate in case they are looking for documents that depict the same event, e.g., a birthday party.

**Weight Inspector** However, the polyrepresentation visualization does not consider the different impact of the representations on the CO, which is demanded by User Story 19. This user story is directly related to User Story 20 as a comprehensible visualiza-

tion is a prerequisite to modify weights in a reasonable way. A visualization of the representations' impact on the CO is provided by the *weight inspector* (see Figure 6.18).

The weight values (or the impact) of each representation onto the CO can be visualized with the help of a bar chart, where the bar's length illustrates the impact of the corresponding representation (see Figure 6.18). With the help of this visualization, users can learn how their interactions affect the system's model of the CO. For instance, users can examine how their stated preferences change the weighting scheme during the different RF iterations. This knowledge can then be used to directly modify weighting variables or to explore the information space along representations that attracted a user's attention.

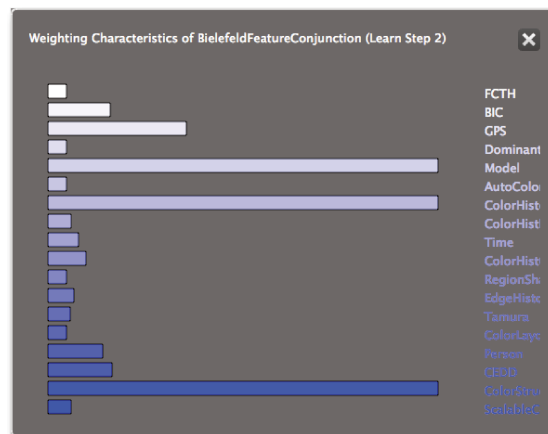


Figure 6.18: Weight inspector

**Characteristic Preferences** Section 5.4.8 suggested the utilization of a rank reduction algorithm to communicate the system's internal model of the IN to the user with help of *characteristic preferences*. In order to be consistent with the UI concept, these characteristic preferences should be displayed in the preference rings widget, which forms the main UI element for the handling of user-input preferences (see Section 6.2.5). Unfortunately, the set of derived characteristic preferences is not guaranteed to contain all user-stated preferences (see Section 5.4.8) and might therefore conflict with the user input. In other words, each RF iteration could alter the preferences in the preference rings if the widget would be used to display the characteristic preferences.

In order to conform with the user expectations, we have decided not to display the characteristic preferences because we believe that such an automatic intervention into the user-initiated preference elicitation process would cause major irritations since recently stated preferences might disappear or change.

Nevertheless, the internal state of the MIR system has to be communicated to the user. This important communicative task is supported by the polyrepresentation visualization and the weight inspector described above. In principle, the Pythia MIR system prototype's UI can also be extended with a display of the characteristic prefer-

## 6 Design of the Pythia MIR System Prototype

ences based on the rank reduction algorithm. Because of the presented arguments, this type of widget has to be placed on interface layer 3 to avoid user irritation and to not disturb the normal preference elicitation process. For instance, this functionality could be combined with the advanced preference graph modification tools.

### 6.3 Prototypical Implementation

The prototypical implementation of the Pythia MIR system is based on prior work that has been developed in Java at the Chair of Database and Information Systems of the Brandenburg University of Technology, which has been supervised by the author of this text.

The CBIR feature extraction and similarity calculation core GOLEM<sup>111</sup> [Zech 2008] was originally based on LIRE<sup>112</sup> [Lux & Chatzichristofis 2008] and has been continuously extended by the author of this dissertation, e.g., to support CQQL evaluation and the weight learning algorithm of PrefCQQL (see Section 5.4.5). In addition, the basic GUI functionality including the described preference elicitation mechanism [Böckelmann 2009] and a SOM-based result visualization [Uhlig 2010] was developed.

The independent parts of the system have been ported to C++, consolidated, and largely redesigned to become transformed into a functional MIR system within a research project funded by the BMBF<sup>113</sup>. In particular, the general system architecture has been revised in order to exploit parallelization features of multi-core CPUs, the feature extraction has been widely re-implemented, the weight learning algorithm has been upgraded to support parallelization, Indri-based textual IR [Strohman et al. 2004] has been added in order to support PDF documents, and an evaluation toolkit has been programmed to name some improvements over the basic Java implementation. The full implementation in C++ is provided as a supplement to this thesis (see Appendix E). The contributors of source code are stated in every code file.

Because of the complexity of the implementation (see Table 6.2), the next sections only outline the architecture of the central components of the Pythia MIR systems. For a more detailed information about the implementation, see the supplement that includes a fully searchable documentation, including complete caller graphs, inheritance diagrams, and the used coding guidelines.

Table 6.2: Lines of code of the Pythia MIR system, revision 1813

Language	No. of Files	Comment Lines	Code Lines
C++	577	23,651	107,855
C/C++ Header	508	30,799	36,826
	<i>Sum</i>	54,450	144,681

Calculated by CLOC version 1.60; <http://cloc.sourceforge.net>

<sup>111</sup>Gadget Obfuscating LIRE's Employment on Metadata

<sup>112</sup>Lucene Image REtrieval

<sup>113</sup>German Federal Ministry of Education and Research

### 6.3.1 Basic System Architecture

The Pythia MIR system is implemented in C++ using the Qt framework<sup>114</sup>. It supports the major operating systems: Mac OS X, Linux, and Windows.

The implementation strictly follows the object-oriented programming paradigm and relies heavily on the model-view-controller architectural pattern [Gamma et al. 1995]. Generally speaking, the design of the Pythia MIR system follows the typical structure of (M)IR systems (see Figure 2.1). In order to control the complexity and the dependency of the subsystems, the facade pattern [Gamma et al. 1995, pp. 185ff.] is commonly used for the central retrieval components.

#### Central Components

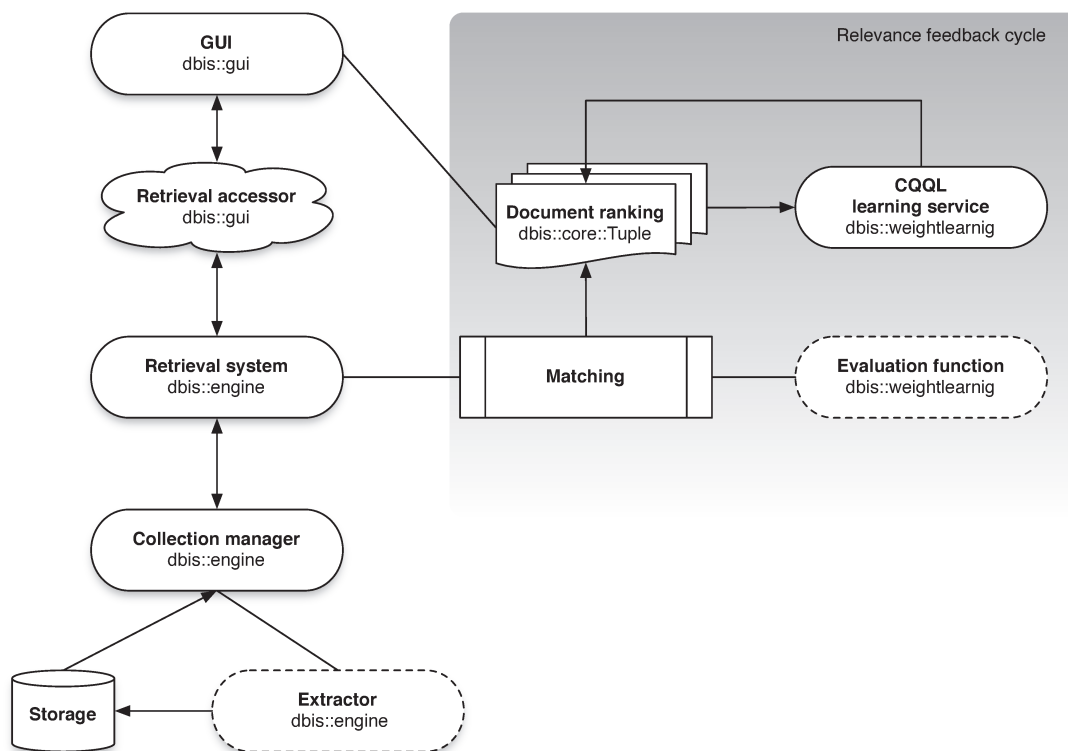


Figure 6.19: Central system components of the Pythia MIR system; dashed lines indicate system components using a dynamic plug-in mechanism

Figure 6.19 displays the central components of the Pythia MIR system and the main control flow between the subsystems. Because of their importance, the RF-relevant

<sup>114</sup>Qt is a C++ framework for multi-platform development; <http://qt-project.org/>

## 6 Design of the Pythia MIR System Prototype

components that link the GUI to the retrieval system (see Section 5.4) are shown separately.

The figure also lists the corresponding namespaces used in the implementation. For the sake of completeness, the most important namespaces and their content are listed in Table 6.3.

Table 6.3: Central namespaces of the Pythia MIR system, revision 1813

Namespace	Content
dbis::core	Core classes and data structures
dbis::engine	The retrieval main engine that combines all retrieval-relevant subsystems
dbis::extraction	Feature extraction (low-level and metadata)
dbis::gui	Main GUI client for preference elicitation
dbis::network	Network communication
dbis::retrieval	Retrieval services including index functionality and data access
dbis::textsearch	Namespace of the Indri format extension
dbis::weightlearning	Nelder/Mead based weight learning implementation and services, matching functions

**Graphical user interface** The GUI constitutes the main user interface of the retrieval system. Because of its importance it is separately discussed in Section 6.3.2. For instance, the GUI gives users access to the relevance feedback cycle that controls the retrieval system's operation. Moreover, it allows the configuration of the retrieval system or the creation of document collection. Because such functionalities have only administrative character, a further discussion is omitted in this thesis.

The GUI is indirectly connected to the retrieval system via the retrieval accessor.

**Retrieval accessor** The retrieval accessor decouples the GUI from the actual retrieval system. In the current implementation, a local pseudo-network connection to the retrieval system is used for the sake of simplicity. However, the GUI can also be connected via a network using the *dbis::gui:NetworkAccessor* that provides a socket-based communication path.

The decoupling of the GUI and the retrieval system enables future implementations to substitute the current GUI, e.g., with a Web-based interface. In addition, it allows a headless operation of the retrieval system to automate long-running experiments.

**Retrieval system** The retrieval system is the main facade of the retrieval engine, i.e., the feature extraction and matching, the storage management, the system's configuration management, and index functionality, to give some examples. It serves as a hub providing all necessary functionality to retrieve documents from a collection and to personalize the results.

**Collection manager** The retrieval system uses the collection manager to access documents and document representations. It also controls the extraction of high- and low-

level features with the help of extractors. In the case of PDF documents, it passes visual parts to the CBIR subsystems and textual parts to the Indri-based IR component [Strohman et al. 2004]. Furthermore, it offers caching functionalities to improve the access performance to query results. To recapitulate, it abstracts the storage management for the retrieval system in order to simplify later changes in the storage or feature extraction functionality.

**Matching** The matching part of the retrieval system uses the strategy pattern [Gamma et al. 1995, cf. pp. 315ff.] that defines a general matching algorithm whose actual implementation varies depending on the evaluation function (i.e., a matching function as listed in Appendix B.1) and whether RF is used. The matching creates a ranking of documents, which is presented to the user and on which preferences can be defined.

**Document Ranking** Internally, each document is represented as a tuple of the similarity scores for a given query. The similarity scores are determined by the retrieval system. The representation in tuple form is handy because it also supports solely DB-based queries and is natural to CQQL. If the user decides to state preferences, the preferences serve in conjunction to the affected tuples as input to the CQQL learning service. The CQQL learning service acts as a facade to the weight learning algorithm that has been presented in Section 5.4.5.

For the sake of flexibility, the Pythia MIR system supports dynamic loadable plug-ins for the extractors and evaluation functions. For instance, this allows the addition of evaluation functions without the need to recompile the program.

The GUI supports statically-linked plug-ins that implement various search strategies and visualizations. This simplifies the control of the provided functionality of the GUI, which is crucial for the conducted usability study (see Chapter 9) that relies on the same code base for each GUI variant. In principle, GUI plug-ins can be made dynamically loadable, although the current implementation forgoes this functionality in order to lower the programming complexity.

### 6.3.2 Overview of the User Interface and Implemented Functions

The conceptual model behind the PrefCQQL-based interaction design has been explained in Section 6.2. This section presents the central user interface parts implemented in the prototypical Pythia MIR system. In principle, the core functionality is described by Zellhöfer [2012a]. Though, the implementation accompanying this dissertation further contains access to PDF and other textual documents with the help of Indri [Strohman et al. 2004] and advanced support for explorative search strategies.

Although the GUI is in principle capable of handling the direct input and modification of CQQL (see Section 6.2.4), this functionality is not implemented because a fully capable CQQL interpreter was not available at the time of the development of the Pythia MIR system prototype. That is, the current UI can only deal with a number of

## 6 Design of the Pythia MIR System Prototype

pre-defined matching functions listed in Appendix B.1. As a consequence, the prototype also does not support negations unless they are part of a pre-defined matching function. Moreover, CQQL-based polythetic clustering is not available at the current stage of development.

One further limitation affects the handling of PDF documents. Although they can be used as QBE documents, their textual contents are ignored by the prototype in favor of the contained images. The current implementation extracts all images that surpass a certain size threshold in order to discard small icons or decoration with minimal semantic contribution (e.g., stylized bullet points). The extracted images are then used as QBE documents. We acknowledge that this interpretation, which eventually attributes all images in a PDF with the same importance for the IN, is not usable in real-world; instead, it is meant as a mere technical proof of concept for an extraction process of composite multimedia documents with Indri. How such multimedia documents should be treated in a QBE scenario remains an area for future research.

Acknowledging the fact that an interactive MIR system can only be inadequately described in textual form, the supplement of this thesis contains multiple videos demonstrating its usage in real life (see Appendix E).

Figure 6.20 displays the main GUI elements of the Pythia MIR system in directed search mode. Arrows indicate paths between the different dialogs. A legend to this figure is available in Table 6.4, which also contains references to the related conceptual design.

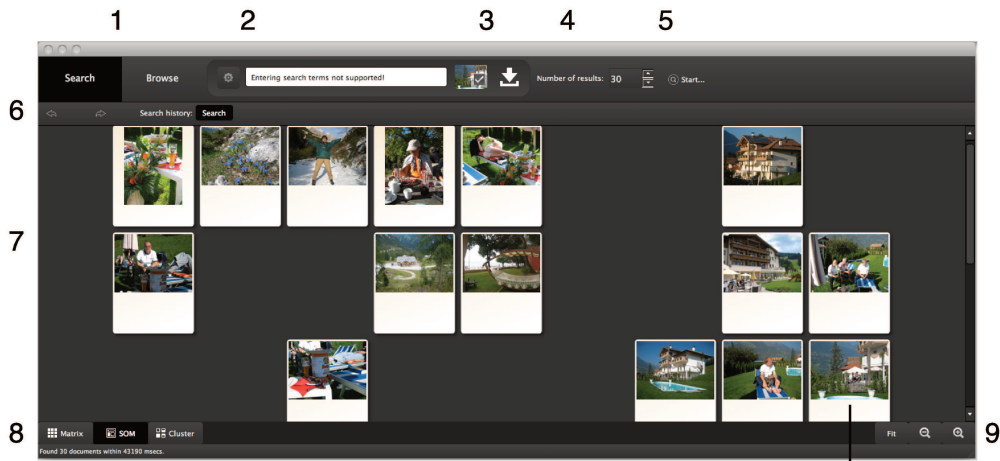
Figure 6.21 shows the GUI in exploratory search mode with the cluster representing documents on the left. As in the directed search mode, the result visualization panel is placed on the right. In order to support the exploration, a  $k$ -medoid clustering approach is used to summarize the collection's contents. The clustering is based on the distances between the color structure representations of the documents. Color structure has been chosen because of its retrieval effectiveness that is discussed in Section 8.4.2, Table 8.9.

Table 6.4: Main GUI elements of the Pythia MIR system (Legend)

Number	Title	See Section...
1	Search strategy indicator	6.2.4
2	Textual query box	6.2.4
3	Multimedia query documents box	6.2.4
4	Search result limiter	6.2.4
5	Search submission	6.2.4
6	Breadcrumb search history navigation	6.2.7
7	Result visualization panel	6.2.3
8	Result visualization selector	6.2.3
9	Result visualization zoom	6.2.3
10	Result personalization window	(Structural element)
11	Preference rings	6.2.5
12	Faceted navigation panel	6.2.6
13	Collection basket and trash bin	6.2.5
14	New search submission	6.2.4, 6.2.5
15	Manual weight setting dialog	6.2.8
16	Polyrepresentation visualization	6.2.9
17	Weight inspector	6.2.9



Main window



Separate dialogs

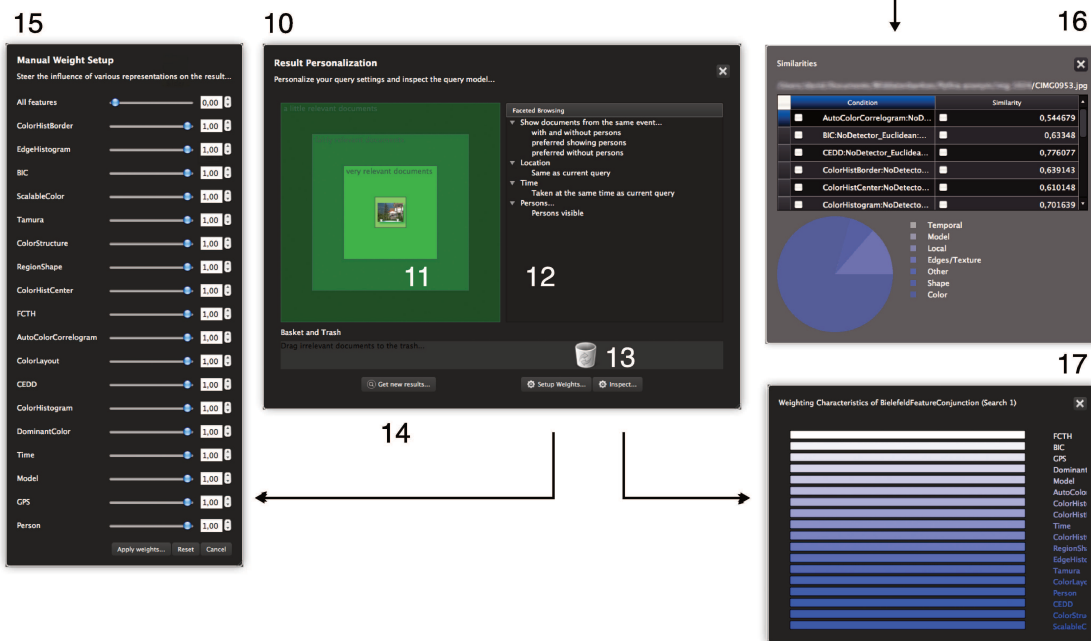


Figure 6.20: Main GUI elements of the Pythia MIR system; the full size preview dialog is omitted

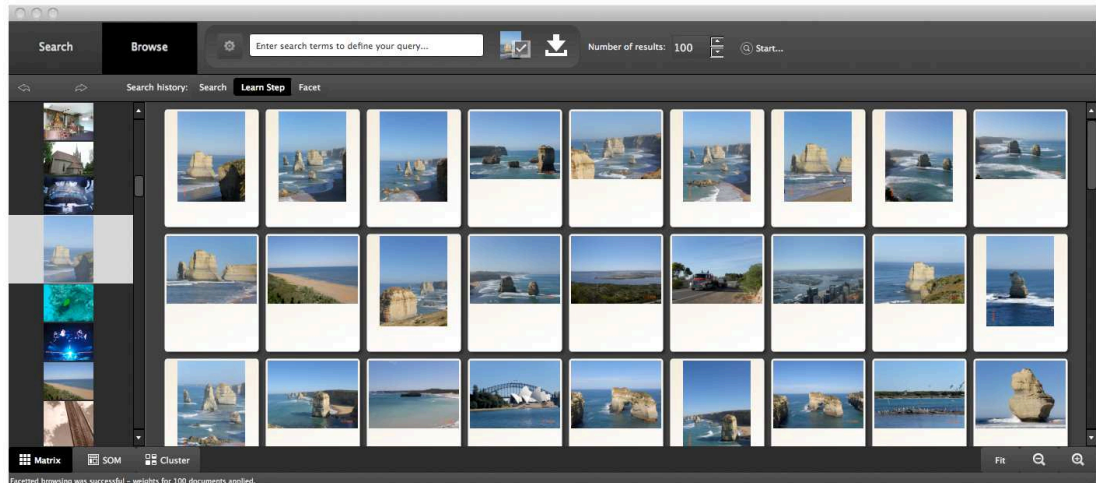


Figure 6.21: Screenshot of the browsing functionality in the Pythia MIR system

## 6.4 The Relation of Pythia MIR to other Interactive MIR Systems

Since the days of the early QBIC system [Flickner et al. 1995], a multitude of CBIR and MIR systems has been developed. Because of the large amount of proposed systems, this dissertation cannot provide a comprehensive discussion on their functions and interaction mechanisms. Instead, this section focusses on characteristic representatives of modern CBIR and MIR systems. Further overviews of such systems, besides the ones given in Section 3.3.2, are available by Rodden & Wood [2003], who concentrate on digital photo management software, or by Hearst [2009], who also presents video retrieval systems. Some textual IR system have been discussed briefly in Section 3.1.1.

As mentioned in Section 3.3, the probably first end-user CBIR system *QBIC* [Flickner et al. 1995] supports only directed searching based on the QBE approach. *QBIC* already combines different features to determine the relevance of image documents. Its GUI is awkward and does not meet current UI design standards, which is not surprising considering the age of the system. Further systems that evolved from *QBIC* are presented in Section 3.3.1. Their main contribution is to include negative QBE documents.

Acknowledging the user need of supporting different ISS, Santini & Jain [2000] suggest a combination of both directed and exploratory search (in form of browsing) for the image database system *El Niño*. In addition, the system tries to integrate visual and textual queries using a unifying query algebra. The *CHROMA* system [McDonald et al. 2001] offers QBE, query by sketch, browsing, and supports RF. However, the system can only exploit a dominant color-based representation to calculate the relevance of the retrieved documents. Both aforementioned systems have in common that the supported ISS are offered separately, i.e., a seamless transition between them is not possible.

Another example for systems combining various ISS is the *Flamenco project* [Yee et al.

## 6.4 The Relation of Pythia MIR to other Interactive MIR Systems

2003], which supports directed keyword search and faceted navigation. In this system, faceted navigation is used to “navigate along conceptual dimensions” [Yee et al. 2003, p. 401] formed by metadata in a fine arts database. The retrieval engine of the Flamenco project relies on the aforementioned metadata. The UI is optimized on the basis of usability studies and seamlessly integrates exploratory search with directed search.

In contrast to the presented multi-ISS systems, Urban et al. [2006] propose a MIR system that only supports explorative information seeking. Their system extends the *ostensive model* [Campbell 2000] by utilizing content-based low-level features in addition to text-based features of the original model. The relevance feedback given in the system “ages” in order to express the dynamics of the user’s IN (see Section 3.3.2). In other words, older relevance judgments receive a lower impact on the RF algorithm than recently stated ones. The development of the systems was accompanied by user studies showing that users perceive the ostensive RF as a useful means to interact with a CBIR system. However, the ostensive model, by definition, does not support the explicit provision of negative feedback.

The reliance on a theoretic model links the systems implementing the ostensive model to the CBIR system variants proposed by Liu et al. [2009, 2010], which are based on *information foraging* [Pirolli & Card 1995; Pirolli 2007] (see Section 3.1.1). Liu et al. [2010] attempt to formalize the information seeking interactions by bringing information foraging theory to CBIR. By enhancing their prior work [Liu et al. 2009], the authors suggest a RF-driven CBIR system called *uInteract*, which is based on QBE. The system also incorporates a temporal component in its RF mechanism, i.e., relevance judgments age similarly to the ostensive model.

Concerning the UI, both the ostensive model-based systems and Liu et al. [2010]’s prototype introduce new concepts. While the former offers a relatively cluttered graph-based explorative navigation through a document collection (see Figures 3.3 and 3.4), Liu et al. follow a more conservative layout and user interaction approach. As Figure 6.22 clearly shows, the design results in a crowded UI without clear boundaries between active and passive parts that breaks with usability best practices, e.g., by not providing informative and concise feedback (see Definition 6.2).

Similar problems can also be observed with other systems, e.g., the *agileviews* system by Marchionini et al. [2000], which relies on different views on the information space aiming at assisting users during exploratory information seeking. The views range from a search history, result presentation to document previews and are visible simultaneously. As a result, the UI becomes very crowded [Zellhöfer 2012a, cf. p. 68]. We attribute these issues to the fact that these systems were not developed following the user-centered design principle (see Section 6.1).

### Conclusion

The most obvious difference of the Pythia MIR system from the aforementioned systems is its support of inductive preferences relying on PrefCQQL. From an (M)IR point of view, the graded PrefCQQL RF approach can be regarded as an extension of the RF approach of *uInteract* [Liu et al. 2010] as well as of other RF approaches (see Section

## 6 Design of the Pythia MIR System Prototype

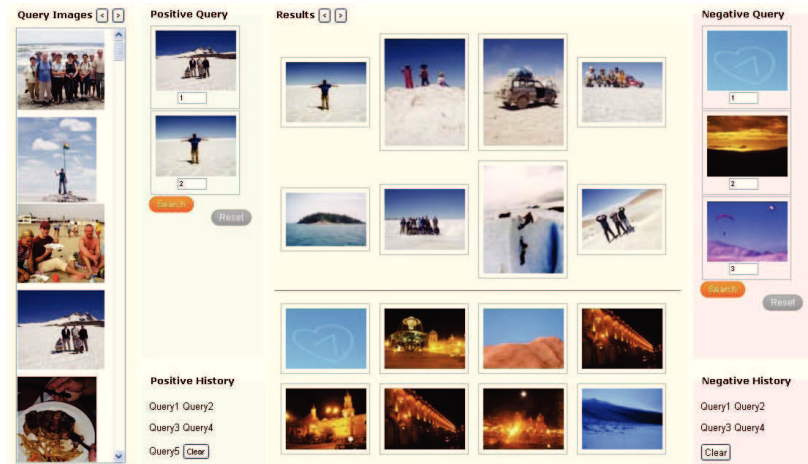


Figure 6.22: The ulteract interface [Liu et al. 2010, Fig. 1]

5.4.3). In Liu et al.'s approach that overcomes binary relevance feedback, which is limited to decisions of relevance or irrelevance, users can provide RF as a total order of relevant and irrelevant documents. Unfortunately, this implies that users actually can decide whether a document is more relevant than another or not. As indicated before, we believe that this decision is not possible in every scenario. Be it that a user considers two or more documents equally relevant or the user does not care about their relevance. This thought can also be found in the FIRE CBIR system [Deselaers et al. 2005], which allows neutral RF in addition to the traditional binary RF input. In contrast, PrefCQQL relies on weak posets allowing users to input graded RF judgments including the equality between documents. Other poset-based approaches could not be found in the literature, although, as mentioned before, Blanken et al. [2007, cf. p. 305] roughly sketch a somewhat similar “weighed relevance feedback” approach without clearly stating its semantics or implementation.

Another main feature of the Pythia MIR system is its utilization of CQQL, which can combine conditions based on different data access paradigms and extends the relational algebra (see Section 4.1). A similar functionality is offered by the El Niño system [Santini & Jain 2000] that combines visual and textual features. However, the El Niño system does not take the different data access paradigms of the features or representations into account. Furthermore, it only supports directed searches and browsing in contrast to the various ISS supported by the Pythia MIR system (see Section 6.2).

The reliance of the Pythia MIR system on CQQL enables it to exploit the full expressive power of logic. Amongst others, this allows the formulation of queries incorporating positive and negative QBE sample documents (see Section 4.4). Thus, negative feedback can also be given at query formulation time and not only during RF as it is the case with other interactive systems such as the ones suggested by Kherfi et al. [2003]; Deselaers et al. [2005], or Liu et al. [2009]. As a consequence, users can retrieve relevant documents potentially faster and more precise.

## 6.4 The Relation of Pythia MIR to other Interactive MIR Systems

Moreover, the fact that the Pythia MIR system is based on the PoP and CQQL gives it both a mathematically and cognitively theoretical sound foundation. This holistic theoretic point of view on the information seeking process makes the system unique in MIR. Although the uInteract system Liu et al. [2010] is also based on a theory, namely information foraging, its theoretic background only focusses on the user interactions. In contrast, the theory behind the Pythia MIR system spans both the user interactions and the retrieval model. Comparably, the ostensive model [Campbell 2000] primarily focusses on the retrieval model, or in particular, the idea of aging relevance judgements. Its UI is only designed to give access to the retrieval model. Although aging preferences are not available in the current prototype of the Pythia MIR system, Section 5.4.7 roughly sketches how such an aging could be implemented with PrefCQQL.

Regarding the supported ISS, many CBIR/MIR system consider directed and exploratory search utilities as concurrent functionalities between users are not expected to change frequently. As elaborated in Section 6.2.1, this assumption is arguable. We believe that the seamless transition between different ISS is crucial for user-centered interactive MIR system. From this point of view, only the Flamenco project [Yee et al. 2003] is comparable to the Pythia MIR system because it combines multiple search strategies and allows a seamless transition between them in order to support the search process. On the other hand, Flamenco lacks support of QBE input and a flexible query language such as CQQL.

Furthermore, the support of multiple ISS in conjunction to the direct access to a query language is necessary to meet the different needs of layperson and expert users. As pointed out in Section 2.3.3, expert users might want to directly access retrieval parameters, such as the weighting variables within CQQL queries or the choice of representations used for the relevance assessments of the retrieved documents. Although this example might be valid for comprehensible representations like textual descriptions, colors, or shapes, it is obvious that even expert users will reach their limit of comprehension (sooner or later) during the interaction with a MIR system (see Section 5.4).

As presented in Section 5.4.2, the Pythia MIR system tries to tackle this problem by bridging the (semantic and query formulation) gap between the user's cognitive and the system's information space. By establishing a dialog between the user and the system, the user is enabled to learn about the system's notion of relevance and vice versa. From a user's point of view, this dialog is supported by various visualizations and preference input mechanisms (both qualitative and quantitative) described before in Section 6.2. Eventually, this bridging is also important to limit the uncertainty of inductive reasoning (see Section 5.4.1) by continuously providing interactive means to negotiate the dynamic IN, e.g., in form of the PrefCQQL RF approach, between the user and the system.

To come to an end, the Pythia MIR system is the only interactive MIR system that is based explicitly on the PoP. As said before, other CBIR/MIR systems mostly apply feature fusion on an ad-hoc basis. To the knowledge of the author, the only other interactive system that relies on the PoP is the implicit RF approach by White [2006], which only supports the polyrepresentation of textual documents.

## 6 Design of the Pythia MIR System Prototype

**Part IV**  
**Evaluation**





## 7 General Considerations

When one brings a human actor ('the user') into the information retrieval setting, all standardization of evaluation disappears. There is no single experimental design to follow."<sup>115</sup>

*Kalervo Järvelin; 2011*

Chapters 2 and 3 presented two somewhat complementary approaches towards IR. While the *system-centric* viewpoint focusses on the IR system, its retrieval model, and algorithms in order to improve its effectiveness, the *user-oriented* viewpoint moves the user into its focal point. This subdivision is also present in this dissertation. Chapters 4 and 5 primarily discuss the query language CQQL and the accompanying machine-based learning algorithm, while Chapter 6 describes how this query language can be used to implement an interactive MIR system based on the principle of polyrepresentation.

These perspectives on IR are also reflected by the split evaluation in the following two chapters. First, the retrieval effectiveness of the CQQL approach is examined. Second, the usability of the presented prototype is investigated. This presence of characteristics of the two major perspectives on IR is common in IIR evaluation [Kelly 2009, cf. p. 17]. In combination, both evaluation approaches give insights into the overall utility of the presented interactive MIR system.

One central problem with the evaluation of IIR or adaptive IR systems is the high number of uncontrolled or hidden variables<sup>116</sup> [Borlund 2003; Voorhees 2008] in contrast to the Cranfield-based approach discussed in Section 7.1. For instance, the familiarity with a given topic (see Section 7.1) of some participants in a study will affect their interaction with the system. On the other side, a weakly performing IR system will affect the perceived usability of a GUI and its overall user experience.

In order to provide both control and realism, IIR evaluations are often based on *simulated work task situations*. Borlund defines a simulated work task situation as a "short 'cover story' that describes a situation that leads to an individual requiring to use an IR system. The 'cover story' is, semantically, a rather open description of the context/scenario of a given work task situation." [Borlund 2003, p. 5]. This task-driven evaluation is also common in usability engineering, where test persons judge the usability of a system based on a defined task (i.e., usually a task the system was designed for [Preece et al. 2002; Nielsen 2009]). Consequently, Borlund also links this approach to the notion of task in human-computer interaction (HCI) that refers to the required actions to achieve a goal using a specific tool [Borlund 2003, cf. p. 7]. It is not sur-

<sup>115</sup>Järvelin [2011, p. 124]

<sup>116</sup>In the sense of experimentation and statistics.

## 7 General Considerations

prising that task-based evaluations have been carried out in multimedia retrieval too, e.g., by Urban et al. [2006], or Liu et al. [2009, 2010]. Nevertheless, a standard design for the evaluation of IIR systems comparable to the Cranfield paradigm could not be established yet [Järvelin 2011].

Although simulated work tasks are a viable means of IIR evaluation, they cannot make clear assertions about the contribution of the IR system's effectiveness and the user interface design on the solution of the task. Furthermore, it is obvious that retrieval effectiveness and the actual user interaction design do both have a huge influence on the user experience. Hence, it becomes complicated to isolate statements about a system's usability and its retrieval effectiveness.

To address this issue, the evaluation presented in this work is subdivided into two parts.

1. A Cranfield-inspired evaluation that is extended to better fit the requirements of the evaluation of an adaptive IR system is presented in Chapter 8.
2. Based on the retrieval effectiveness evaluation, three alternative prototypes are evaluated by means of usability engineering in Chapter 9 using the same simulated work task, query, and retrieval model.

According to Kelly's interpretation [Kelly 2009, cf. p. 15], this separation moves the evaluation described in this thesis into the system-centric direction because it tries to establish more control over the experimental variables. Nevertheless, this does not mean that the evaluation is only relying on quantitative methods. Instead, it is complemented by qualitative studies of the user experience (see Chapter 9).

In order to relate this evaluation design to the state of the art, the following sections discuss the current evaluation practices with a focus on adaptive multimedia retrieval.

### 7.1 The Traditional Cranfield Approach

Unlike the evaluation in the DB domain, which is mainly focussing on *performance*, e.g., query execution time or transactions per second<sup>117</sup>, and supported functionality (SQL standard compatibility, maintenance utilities, programming interfaces etc.), the main objective of IR evaluation is the measurement of *effectiveness*. Although IR systems can differ in their functionality, e.g., by supported document types or the power of their query language, the field traditionally compares systems to each other by assessing how well they perform in satisfying a user's IN. In other words, the result quality of the IR systems is evaluated.

In order to establish an experimental setup that guarantees control over the experimental variables and repeatability, Cleverdon [1962] proposed three basic principles for the evaluation of IR systems that are nowadays known as the *Cranfield paradigm*. According to the Cranfield paradigm, an experimental setup consists of the following mandatory parts:

---

<sup>117</sup>These measurements can be obtained by running standardized benchmarks, e.g., provided by the Transaction Processing Performance Council (<http://tpc.org/>).

## 7.1 The Traditional Cranfield Approach

**Definition 7.1 Document collection:** *A defined immutable document collection that is used during the test.* ◇

**Definition 7.2 Topics:** *Several sample information needs that are represented by one or more queries.* ◇

**Definition 7.3 Relevance assessments:** *Separate relevance assessments for all documents of the collection with respect to the pre-defined topics.* ◇

It is important to mention that users and their interaction with an IR system are excluded from Cranfield-based evaluations. This is to increase the control over the experimental variables and to simplify the execution of the experiment(s). In the end, this results in lower costs of such experiments in terms of time and money in comparison to the evaluation of adaptive systems (see Section 7.2) that directly involves users [Voorhees 2008]. The omission of users in Cranfield-based experiments links this approach to the system-centric viewpoint introduced in Section 2.2. Figure 7.1 illustrates the connection by extending the original Figure 2.1, where the shaded items denote the controlled experimental variables in the Cranfield paradigm.

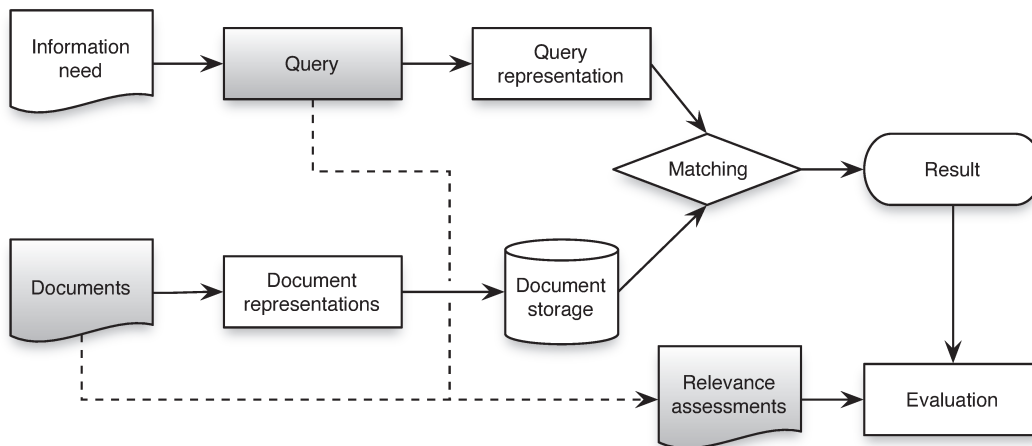


Figure 7.1: Schematic illustration of the Cranfield laboratory setting; relevance feedback is omitted for clarity

In order to setup a Cranfield-based experiment, a document collection has to be defined. Additionally, topics that abstract information needs of (potential) users are constructed. Once specified, the topics remain static throughout the experiment. Based on these topics, each document of the collection (or a representative subset) is assessed according to its relevance to the given topic. These relevance assessments are often done by professional users (assessors), whose individual assessments are then averaged. Traditionally, the relevance of a document is on a binary scale, i.e., it can either be relevant with respect to a topic or not. The combination of topics and relevance assessments is

## 7 General Considerations

called the *ground truth* of the collection. The appeal of this experimental design lies in its clarity and simplicity. Once a collection and a ground truth are defined, alternative IR system designs can be tested and compared to each other. Given that one has access to all tested IR systems in addition to the collection and the ground truth, the results of an experiment are reproducible.

The Cranfield paradigm has been shown crucial in advancing the field of IR [Ingwersen & Järvelin 2005; Voorhees 2008]. In principle, today's major evaluation initiatives such as TREC (Text REtrieval Conference) [Voorhees & Harman 2005] or CLEF<sup>118</sup> (originally Cross-Language Evaluation Forum, now Conference and Labs of the Evaluation Forum) are all based on the Cranfield paradigm that has been successfully used in IR for decades.

Being a more recent field of research, MIR cannot look back on an evaluation tradition originating in the 1960s. Nevertheless, prominent evaluation initiatives such as ImageCLEF [Müller et al. 2010] are also based on the Cranfield paradigm. Although some researchers, e.g., Wang et al. [2001], published their document collections and ground truths from early on, the establishment of an evaluation initiative was for a long time out of the focus of mainstream MIR research [Müller et al. 2010]. Starting in 2003, the retrieval of video data was introduced as the TRECVID [Smeaton et al. 2006] subtask to TREC. Nowadays, ImageCLEF can be considered as one of the most prominent evaluation initiatives with a focus on MIR. This is in particular true for the subtasks relying on the MIRFLICKR collection [Huiskes & Lew 2008b; Huiskes et al. 2010] that have remained relatively stable over the past years. The MIRFLICKR collection contains 200,000 documents from Flickr<sup>119</sup>. Its ground truth ranges from topics dealing with an image's quality such as its blur to visual concepts such as beaches. Alternative ImageCLEF subtasks used images and texts from Wikipedia<sup>120</sup> or x-ray images.

Other recently established initiatives leave the traditional topic-based context of evaluation and focus on the retrieval of specific events (e.g., concerts or sport events). One example is the MediaEval benchmark<sup>121</sup>, which is based on VideoCLEF [Larson et al. 2008]. For instance, MediaEval's *social event detection task* mirrors the current discussion that photos mainly illustrate events, e.g., observed by Amarnath & Jain [2011]. The task itself is based on 70,000 photos from Flickr that include metadata such as date and local information. The accompanying ground truth is binary.

### 7.2 Evaluation of Adaptive Retrieval Systems

Voorhees defines *adaptive information retrieval* as "a search process process that adapts toward the user's needs and context" [Voorhees 2008, p. 1]. This definition can be regarded synonymous with the interpretation of interactive information retrieval (see Section 3.1) used in this thesis. As described in the last section, this adaption toward

---

<sup>118</sup><http://www.clef-initiative.eu/>

<sup>119</sup><http://www.flickr.com>

<sup>120</sup><http://www.wikipedia.org/>

<sup>121</sup><http://www.multimediaeval.org/>

the user's needs falls out of the scope of the Cranfield paradigm. In consequence, the Cranfield approach provides little support for research focussing on adaptive IR. This point of view is also shared by advocates of the Cranfield tradition such as Voorhees [2008].

The main criticism of the Cranfield paradigm is that it highly abstracts the actual users' information needs and their context in order to keep the experimental variables controllable. In fact, the interaction forms and cognitive processes described in Chapter 3 are completely neglected [Borlund 2003; Ingwersen & Järvelin 2005].

Borlund [2003] further punctuates her criticism by pointing out three revolutions originally presented by Robertson & Hancock-Beaulieu [1992] that are not addressed by the Cranfield paradigm, i.e.:

1. The cognitive revolution
2. The relevance revolution
3. The interactive revolution

The first two revolutions require more realism regarding the utilized topics representing INs and the nature of the relevance assessments. In accordance with Section 3.1, the dynamic nature of INs and relevance assessments have to be taken into account. That is, topics and relevance should be judged against the user's current situation. This subjectivity of the perceived relevance is also discussed by Voorhees [2008], who similarly to Borlund [2003] criticizes the binary relevance scale in the Cranfield paradigm. Acknowledging the importance of the user interaction, Borlund [2003] includes an observation of the user's searching processes into her proposal of an IIR evaluation model consisting of the following three components:

- "Part 1. A set of components which aims at ensuring a functional, valid, and realistic setting for the evaluation of IIR systems;
- Part 2. Empirically based recommendations for the application of the concept of a simulated work task situation [1]; and
- Part 3. Alternative performance measures capable of managing non-binary relevance assessments."<sup>122</sup>

[Borlund 2003, p. 4]

Roughly speaking, the core idea behind the first two parts is to include potential users in the evaluation and to collect data. In order to maintain control over the experimental variables, Borlund [2003] advocates the usage of simulated work tasks – a concept proposed, e.g., in Ingwersen [1992, 1996] and extended in Ingwersen & Järvelin [2005]. Work tasks allow the inclusion of situational factors and the user's context into the evaluation, improving the realism of the experiment – especially if potential users are involved. On the other hand, simulated work task situations define a controlled setting, in which the interaction with the IR system can be examined.

<sup>122</sup>The footnote [1] refers to "With respect to the traditionally employed performance measures of recall and precision." [Borlund 2003].

## 7 General Considerations

The third part of the model deals with the actual analysis of the collected data. It requires alternative IR performance metrics that support non-binary relevance assessments, e.g., as presented in Borlund [2003]. One of those measures used in this dissertation is presented in Section 8.1.2.

Although simulated work tasks offer a certain degree of experimental controllability, the inclusion of real users reduces the repeatability of such experiments. Furthermore, users are affected by tiring and learning effects. One potential evaluation technique that addresses these issues is the usage of *user simulations*. The focus of such simulations is on the simulation of user interactions with an IR system [Azzopardi et al. 2011]. Although user simulations do not necessarily represent a full simulated work task, they provide a subset of controllable experimental variables (a factor that is endorsed by some authors, e.g., by Voorhees [2008, cf. p. 1884]) and introduce reproducibility to the experimental setup. Additionally, they can reduce the overall cost of the experiment and scale well without violating financial and temporal constraints – the main factors limiting the practical feasibility of tests involving users [Voorhees 2008]. After a SIGIR workshop on the simulation of interaction in 2010 [Azzopardi et al. 2011], publications at ECIR 2011 [Baskaya et al. 2011], SIGIR 2012 [Baskaya et al. 2012], and a broad usage at IiX 2012 [Kamps et al. 2012], user simulations are currently gaining acceptance in the context of IIR evaluation [Zellhöfer 2012f].

To recapitulate, the focus of classical MIR evaluation remains system-centric, although some benchmarks suggest the inclusion of users into the experiments, e.g., at ImageCLEF [Müller et al. 2010; Zellhöfer 2013]. Nevertheless, these interactive evaluation initiatives have not yet created a persistent momentum for IIR experiments in MIR.

## 8 Evaluation of the Retrieval Effectiveness

[...] the time may not be very remote when it will be understood that for complete initiation as an efficient citizen of one of the new great complex world-wide States that are now developing, it is as necessary to be able to compute, to think in averages and in maxima and minima, as it is now to be able to read and write.

---

*H. G. Wells, Mankind in the Making; 1903*

Although the entry quote is often (and wrongly<sup>123</sup>) summarized as “Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.”, the next chapters follow its core idea and expect a certain understanding of basic statistics. That is – for the sake of brevity – basic notations of statistics such as scale types will not be introduced as there are plenty of good introductions available, e.g., by Mendenhall et al. [1999] or Miller & Miller [1999].

### 8.1 Comparing the Effectiveness of IR Systems

Since the early days of IR and its systematic evaluation in the 1960s and 1970s represented by works of Cleverdon [1962] and Lesk & Salton [1968], numerous evaluation metrics have been suggested and used for measuring the retrieval performance of IR systems. At present, the de-facto standard tool for retrieval metric calculation, *trec\_eval*<sup>124</sup> lists over 30 different metrics that became relevant during two decades of TREC [Voorhees & Harman 2005, Ch. 3]. Clearly, not all available metrics can be presented in this work. Thus, we focus here on some representatives of system-centric (see Section 8.1.1) and user-centric measures (see Section 8.1.2) that are directly related to this thesis.

Detailed overviews of the commonly used metrics in both IR and MIR are available in the basic textbooks, e.g., van Rijsbergen [1979]; Croft et al. [2009]; Manning et al. [2009], or Baeza-Yates & Ribeiro-Neto [2011]. Their usage and utility in the scope of TREC is discussed by Voorhees & Harman [2005], while Müller et al. [2010] take a similar perspective on ImageCLEF. More user-centered aspects of the evaluation of IR are covered by Kelly [2009, Ch. 10] or Järvelin [2011].

---

<sup>123</sup>An interesting investigation on how the quote has been changed throughout various citation cycles to finally fit the need of statisticians is available by Tankard [1979].

<sup>124</sup>Version 9.0 with “-m all\_trec” parameter as obtained from [http://trec.nist.gov/trec\\_eval/](http://trec.nist.gov/trec_eval/) at October 6th, 2012 and compiled with GCC 4.2.1 (build 5658) under Mac OS X 10.6.8.

### 8.1.1 System-Centric Retrieval Metrics

System-centric retrieval metrics are tightly coupled to the IR mapping presented in Section 2.2 and Cranfield tradition (see Section 7.1). The following metrics are often called system-centric because they mainly focus on the general quality of a result ranking, e.g., how many relevant documents are retrieved. Other factors affecting the performance of an IR system such as the user interface, the query context, or particular user preferences are not considered by these measures [Baeza-Yates & Ribeiro-Neto 2011]. Nonetheless, these measures have been shown to be useful during the last decades. They associate a certain number (or a set of numbers) with a given result that can easily be used to compare the retrieval performance of multiple IR systems.

Although the following two metrics – *precision* and *recall* – are not directly used in this thesis, they are essential for an understanding of IR evaluation.

**Definition 8.1 Precision:**

$$Precision = \frac{|R_q^+ \cap R|}{|R|}, \quad (8.1)$$

where  $R_q^+$  are the relevant documents with respect to a query  $q$  and  $|R_q^+ \cap R|$  denotes the number of retrieved relevant documents regarding a query  $q$  divided by the number of retrieved documents  $R$ .  $\diamond$

**Definition 8.2 Recall:**

$$Recall = \frac{|R_q^+ \cap R|}{|R^+|} \quad (8.2)$$

In contrast, recall divides the retrieved relevant documents by the total amount of relevant documents  $R_q^+$  regarding query  $q$  in the document collection.  $\diamond$

Usually, precision is plotted against recall at standard levels, i.e., the interval  $[0, 1]$  in increments of 0.1, to illustrate the effectiveness of an IR system. In order to obtain the precision values at the standard recall levels, precision values are interpolated [Croft et al. 2009].

Precision and recall have in common that they are set-based. That is, an examined result is considered unordered. In contrast, Section 2.2.2 pointed out that modern IR systems are based on totally ordered rankings, e.g., sorted by the probability of relevance of the result documents.

To cope with this development, other measures have been proposed, e.g., *precision at n* (see Section 8.1.2) or *mean average precision* that has become the metric “most often used in IR research to represent the overall effectiveness performance of a system.” [Voorhees & Harman 2005, p. 59].

**Definition 8.3 Average precision:** The average precision ( $AP_i$ ) for a query  $q_i$  is defined as

$$AP_i = \frac{1}{|R_q^+|} \sum_{n=1}^{|R_q^+|} Precision(R[n]), \quad (8.3)$$



## 8.1 Comparing the Effectiveness of IR Systems

where  $|R_{q_i}^+|$  is the number of total relevant documents for the query  $q_i$ . Let  $R[n]$  be the reference to the  $n$ -th document in the examined ranking. Thus,  $\text{Precision}(R[n])$  is the precision of the  $n$  best documents after a relevant document is retrieved. If the document at  $R[n]$  is not relevant,  $\text{Precision}(R[n])$  is taken as zero.  $\diamond$

**Definition 8.4 Mean average precision:** Then, the mean average precision (MAP) over a set of queries  $Q$  with  $q_i \in Q$  is defined as

$$MAP = \frac{1}{|Q|} \sum_{i=1}^{|Q|} AP_i. \quad (8.4)$$

That is, the arithmetic mean over all  $AP_i$  for each query  $q_i \in Q$  [Voorhees & Harman 2005].  $\diamond$

### Averaging Methods

When retrieval metrics are averaged over multiple queries, one can basically decide between the utilization of *macro-* or *micro-averaging*. MAP is a typical representative of macro-averaging. Generally speaking, macro-averaging is used most in IR evaluations [Croft et al. 2009, p. 319].

In *macro-averaging*, the measurement to be averaged is computed for each query; thereafter, these per query measurements are averaged. In literature, this approach is often called “user-oriented” in contrast to the “system-oriented” *micro-averaging*, which first combines the results of all queries and then calculates the average on these values. In other words, macro-averaging weights all queries equally (no matter how many results are contained), while micro-averaging weights all results equally. Thus, queries with a large number of results gain a higher impact on the computed averaged measurement when micro-averaging is used.

### 8.1.2 User-Centric Retrieval Metrics

In contrast to the system-centric metrics discussed in the last section, user-centric measures try to incorporate subjective notions of relevance or the direct utility of a system for a user into the evaluation. For instance, the utility can be assessed by the amount of relevant documents retrieved at early positions of the ranking, or by the novelty of the results. That is, how many unseen documents are retrieved by the system, e.g., after a relevance feedback iteration. While the first can be measured by using *precision at n*, the latter might be examined utilizing rank correlation coefficients.

Besides the fact that system-centric measures are based on the idea that the set of relevant documents is the same for all users, they are also highly affected by outliers, i.e., relevant documents retrieved at a late position in the ranking [Baeza-Yates & Ribeiro-Neto 2011]. For instance, the precision of a system that contains 10 relevant documents at the first 10 positions of the ranking containing 100 documents in total gets the same precision as a system ranking the same relevant documents at position 91-100. Without

## 8 Evaluation of the Retrieval Effectiveness

doubt, the user would prefer the first system. This finding is reflected by the following metric.

**Definition 8.5 Precision at n:** Precision at  $n$  ( $P@n$ ) calculates the precision assuming the result set contains only the first  $n$  elements of the ranking whereas  $n$  is also called the cutoff. That is,  $\mathbf{R} = \{\mathbf{R}[1], \dots, \mathbf{R}[i]\}$ ; with  $i \leq n$ . Given  $r^+$  documents are relevant in  $\mathbf{R}$ , then

$$\text{Precision @ } n = \frac{r^+}{n}. \quad (8.5)$$

For comparison purposes, cutoffs are typically set to standard levels as 5, 10, 20 etc.  $\diamond$

Because different IR systems might return rankings of different sizes for a query, cutoffs are also used for other measures such as MAP. Please note that the expressiveness of  $P@n$  is limited for large cutoff values. In this case, the same limitations as discussed above apply.

With reference to the argument to incorporate non-binary relevance assessments in IIR evaluations (see Section 7.2), the aforementioned measures become of limited utility because all of them assume a binary relevance scale. One of the metrics supporting non-binary relevance is the *discounted cumulated gain* (DCG) [Järvelin & Kekäläinen 2002] presented below. Alternative measures can be found, e.g., in Borlund [2003] or Carterette [2011]. The decision to use DCG in this thesis is based on three main reasons. *First*, it supports non-binary relevance assessments. *Second*, DCG has gained acceptance in the IR community. For instance, a recent performance evaluation of different IR effectiveness measures presented by Carterette [2011] at SIGIR 2011 shows the utility of DCG as a user-centered measure in terms of its realism regarding the actual user interaction and its robustness. *Third*, the metric is based on an intuitively comprehensible model of user demands presented below.

**Definition 8.6 Discounted Cumulated Gain:** The discounted cumulated gain (DCG) vector [Järvelin & Kekäläinen 2002] accumulates the non-binary relevance score of each retrieved documents in the order of their position in the ranked result list. Additionally, it applies a discount factor to each document's relevance score based on its position in the list. This discount factor is meant to devalue relevant documents that are retrieved late and are therefore unlikely to be seen by a user reading down the result list. This discount factor is complemented by the assumption that the user is most interested in highly relevant documents at an early position. In order to calculate the DCG vector, the cumulated gain (CG) vector [Järvelin & Kekäläinen 2002, Eq. 1] has to be introduced first.

$$\text{CG}[i] = \begin{cases} G[1], & \text{if } i = 1 \\ \text{CG}[i - 1] + G[i], & \text{otherwise.} \end{cases} \quad (8.6)$$

$G[i]$  denotes the relevance score of the document at the retrieved rank position  $i$  as well as the CG vector's component.

To devalue lately retrieved but relevant documents, DCG uses a rank-based discount factor that effectively limits their contribution to the cumulated gain [Järvelin & Kekäläinen 2002, Eq.

2].

$$DCG[i] = \begin{cases} CG[i], & \text{if } i < b \\ DCG[i-1] + \frac{G[i]}{\log_b i}, & \text{if } i \geq b. \end{cases} \quad (8.7)$$

In this equation,  $\mathbf{b}$  denotes the base of the logarithm. The logarithm is used to allow a progressive reduction of the relevance score in favor to a steep reduction, e.g., as the division by  $\mathbf{i}$ . The discount factor is not applied for positions lower than the logarithm's base to avoid boosting of the relevance score, which would distort the results. Furthermore, it is not used for the first rank position to avoid a division by 0 ( $\log_b 1 = 0$ ).

The increase of the logarithm's base smooths the effect of the discount factor in order to model a patient user that is willing to scan through a (potentially) large number of result documents.  $\diamond$

Although Järvelin & Kekäläinen [2002] suggest to use four relevance levels encoded by 0 to 3, where 0 denotes irrelevance and 3 full relevance, DCG is not limited to this scale. For instance, Spink et al. [1998] rely on a scale with relevant, partially relevant, and irrelevant documents that could also be used with DCG. Binary relevance assessments can also be used with the metric, although it will then not express the aforementioned assumptions about the user preferences.

In order to compare multiple IR systems to each other, one has to normalize the DCG vector using an *ideal gain vector* (iGV).

**Definition 8.7 Ideal Gain Vector:** The ideal gain vector (iGV) [Järvelin & Kekäläinen 2002] is the theoretically best possible order of relevance scores. That is, a list of all document relevances ordered descending by their magnitude. In the case of a relevance scale ranging from 1 to 3, where  $\mathbf{k}$  denotes the number of documents at level 1,  $\mathbf{l}$  the documents at level 2, and  $\mathbf{m}$  the number of fully relevant documents, it can be constructed as follows [Järvelin & Kekäläinen 2002, Eq. 4]:

$$iGV[i] = \begin{cases} 3, & \text{if } i \leq m, \\ 2, & \text{if } m < i \leq m + l, \\ 1, & \text{if } m + l < i \leq m + l + k, \\ 0, & \text{otherwise.} \end{cases} \quad (8.8)$$

$\diamond$

This iGV is then used to normalize the actual DCG vector to allow the comparison of multiple IR systems.

**Definition 8.8 Normalized Discounted Cumulated Gain:** The normalized discounted cumulated gain (nDCG) vector is the result of the division of each component  $\mathbf{i}$  of the DCG vector by the related iGV component, i.e.:

$$nDCG[i] = \frac{DCG[i]}{iGV[i]}. \quad (8.9)$$

$\diamond$

### 8.1.3 A Brief Critique of the Rank-based Measures

Although DCG incorporates a user model, it is still inherently rank-based. Rank-based result visualizations (lists) are still the predominant presentation form in end-user IR systems but have deficits in MIR as outlined in Chapter 6 or by Santini [2012].

In the field of MIR, supportive user interfaces provide different visualizations, leaving the list-based paradigm in favor to (at least) 2D matrix visualizations (e.g., Campbell [2000]; Urban et al. [2006]; Liu et al. [2009]) or even 3D visualizations [Nakazato & Huang 2001]. Hence, the assumption of the DCG measure that there is a 1:1 mapping between the order of the result rank and the presented form in the UI is violated in the scope of this thesis. In other words, a document with the second highest probability of relevance is not necessarily placed in spatial neighborhood of the first placed one, e.g., if a cluster-based visualization is used. As a result, a high DCG value obtained by the MIR system might not correspond to a subjectively perceived good retrieval effectiveness.

These arguments are also true for other rank-based effectiveness measures or rank correlations metrics (see Baeza-Yates & Ribeiro-Neto [2011, Ch. 4]). Furthermore, rank correlation coefficients are much harder to interpret because they often fall into the interval  $[-1; 1]$  and have no clear boundaries that allow statements how well an IR system performs with respect to another system. This problem is formulated accurately by Krippendorff: "Except for perfect agreement, there are no magical numbers, however. The ones suggested here should be verified by suitable experiments." [Krippendorff 2004, p. 422]. This puts an additional burden onto the researcher and explains why rank correlation coefficients are seldom used for the evaluation of IR systems.

Although DCG and nDCG have the aforementioned shortcomings, the latter will be used for the evaluation of CQQL and PrefCQQL's retrieval effectiveness. This approach is state of the art in the evaluation of IR systems but should not be mistaken with the subjectively perceived user satisfaction, which is also affected by a system's usability as elaborated in Chapter 9.

### 8.1.4 Testing Statistical Significance

In the last sections, exemplary retrieval metrics for the comparison of multiple IR systems were presented. These metrics generate numbers that can be used for such comparisons, but they do not show whether the difference between two IR algorithms is meaningful in a statistical sense. That is, whether the observed effect is not due to chance. Thus, *significance tests* are needed.

Significance tests are based on a *null hypothesis*. The null hypothesis in the scope of this dissertation states that there is no difference between the retrieval metrics of a baseline system *A* and a tested system *B*. However, we are interested whether the test system performs better than the baseline, i.e., the *alternative hypothesis*. Roughly speaking, significance tests are a means to indicate if the null hypothesis can be rejected in favor of the alternative hypothesis. If so, one can tell that there is a statistical significant difference between the retrieval effectiveness of both systems.

## 8.2 Overview of Cranfield-based Test Collections

A widely used significance test in IR is the *Student's t-test* [Croft et al. 2009]. Although the t-test makes certain assumptions about the test statistic (the examined sample), e.g., that it is approximately normally distributed, it is known to be robust against violations of this assumption [Mendenhall et al. 1999, cf. p. 386]. This observation is supported by experimental results in IR reported by Smucker et al. [2007]. A more thorough discussion of significance testing in the field of IR is available in Blanken et al. [2007, Ch. 13] or Croft et al. [2009, Ch. 8].

Which significance test is appropriate for the experiments described in this thesis is discussed in Section 8.4.3.

## 8.2 Overview of Cranfield-based Test Collections

Comparably to IR, Cranfield-based test collections are the predominant benchmark of retrieval quality in MIR. Starting with hundreds of images, the collections have been constantly enlarged and now contain thousands to millions of images associated with a diversity of topics. Because of the amount of available collections with differing objectives, this dissertation cannot discuss all of them in detail. While Section 7.1 gave an overview of their development, this section focusses on the collections that are used to evaluate the retrieval effectiveness of the CQQL approach (see Section 8.4). This approach is strongly advocated by Voorhees who postulates that “a candidate retrieval technique [should] be consistently better than the alternatives on a variety of test collections before concluding the effect is real.” [Voorhees 2008, p. 1884].

In accordance with this requirement, the presented collections have been chosen to represent different usage scenarios in order to draw conclusions about the general performance of CQQL as an implementation of the principle of polyrepresentation. Additional collections are presented in Blanken et al. [2007, Ch. 13] and Müller et al. [2010].

### 8.2.1 Caltech 101 and 256

The Caltech 101 test collection [Fei-Fei et al. 2004] consists of 9,197 bitmap images in color and grayscale. The name of the collection is due to total amount of 101 available topics. The choice of topics is diverse and includes topics such as faces, dinosaurs, airplanes, gramophones, yin-yang symbols, and various animals. It is arguable whether all topics really mirror daily-life retrieval tasks, e.g., the infamous prehistoric subset consisting of images depicting stegosaurus, brontosaurus, and trilobites. Figure 8.1 illustrates typical image types found in the collection ranging from illustrations, photographs, drawings, paintings, and scans. Often, the documents themselves are rotated, scaled, or modified in a way that their background has been replaced by an artificial one (see Figure 8.1, first two images in the first row). Obviously, these modifications are due to the collection's origin in object recognition as they can be used to measure the sensitivity of feature extraction algorithms regarding invariances etc.

Based on this collection, Griffin et al. [2007] propose a successor – the Caltech 256 collection, which features 30,607 documents subdivided into 256 topics sharing roughly

## 8 Evaluation of the Retrieval Effectiveness

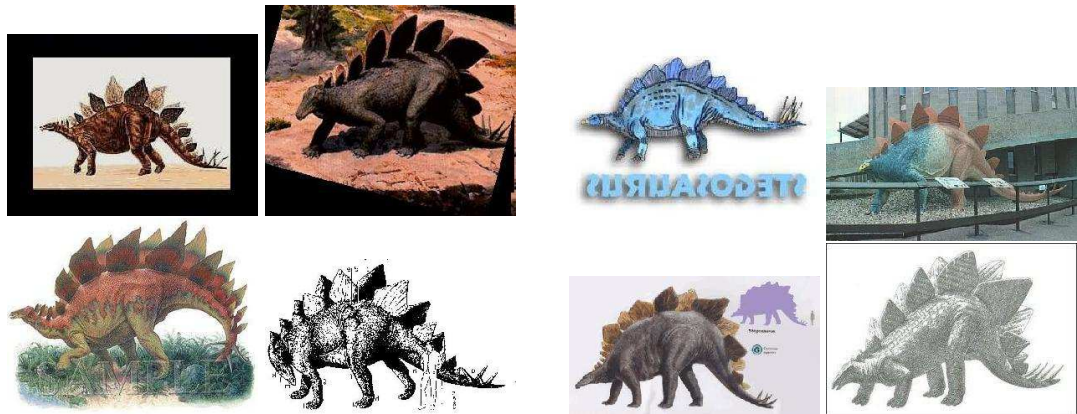


Figure 8.1: Samples from Caltech 101's "stegosaurus" topic

the same characteristics with Caltech 101.

The Caltech collections have been chosen deliberately to be used in this dissertation in order to include collections without a typical CBIR background. The main motivation to include also objection recognition collections is to investigate whether the predictions made by the principle of polyrepresentation can be generalized as described above.

**Generation of the Ground Truth** The collections are distributed with a binary ground truth that is used for the experiments described below.

### 8.2.2 MSRA-MM

The MSRA-MM collection [Wang et al. 2009] contains 65,443 thumbnail images taken from Microsoft Live Search (now replaced with Microsoft Bing). The accompanying sample queries are based on Microsoft's search logs. Unlike the collections discussed before, the collection features relevance assessments with 3 relevance degrees, i.e., very relevant, relevant, and irrelevant. As argued in Zellhöfer [2012c], this would enable MSRA-MM to be used with the DCG metric. Because of copyright issues, the original images are omitted in the collection. To compensate this problem, seven global low-level features accompany the thumbnails. Nevertheless, new features cannot be extracted from the original images limiting the collection's general utility if one assumes an effect due to the scaling of images on the feature extraction.

For each topic, there are 1,000 QBE documents, of which 100 were randomly chosen to be used in this thesis due to computability considerations (see Section 8.4).

Comparable with Caltech 101 and 256, there is also a revision of the dataset<sup>125</sup> containing ca. 1 million images and 1,165 queries. The revision of MSRA-MM is not examined in this thesis because of memory space and computability constraints (see Section 8.4).

<sup>125</sup><http://research.microsoft.com/en-us/projects/msrammdata/>

**Generation of the Ground Truth** The collection is distributed with a graded relevance ground truth that is used for the experiments described below.

### 8.2.3 UCID v2

The UCID v2 [Schaefer & Stich 2004] collection has replaced a predecessor of the same name that will not be used in this thesis. The UCID collection consists of 1,338 uncompressed TIFF images, of which only 904 could be read properly due to compatibility issues (see Table 8.1). The objective of this collection is to provide a benchmark of personal photographs. All images are taken with the same camera model and automatic settings, i.e., special effects such as Sepia coloring or the like are not present. The topics range from persons or natural scenes to man-made objects such as houses. The images are shot indoors and outdoors.

**Generation of the Ground Truth** The collection is distributed with a binary ground truth that is used for the experiments described below.

### 8.2.4 Wang

The Wang collection is a subset of the Corel stock photography collection<sup>126</sup> first used by Wang et al. [2001]. It contains 1,000 images from professional photographers. The images are both taken indoors and outdoors and belong to 10 topics. Sample topics are “sports and public events”, “beach”, or “dinosaurs”. One interesting point about this collection is that the topics consist of events and motifs unlike the aforementioned collections that are focussing on motifs alone, e.g., a photo of an elephant.

**Generation of the Ground Truth** Although the original publication [Wang et al. 2001] contains a comparison of different features, the ground truth has not been made fully available. Instead, a table [Wang et al. 2001, Tab. 1] listing 10 topics is given without an association between topics and images. Hence, the ground truth had to be reconstructed manually on the basis of this table. To obtain the ground truth, the relevance of all images to the given topics was judged by one assessor using a binary scale.

### 8.2.5 Summary

To conclude, Table 8.1 summarizes the characteristics of the presented collections. Regarding the content of the collections, MSRA-MM, UCID, and Wang can be considered as more realistic in terms of their real-world content than both Caltech collections. Nevertheless, MSRA-MM and Wang are extracts from images found on the Web or from a professional collection, i.e., a set of images composed by a user possibly after image

---

<sup>126</sup>The full collection cannot be purchased any longer.

## 8 Evaluation of the Retrieval Effectiveness

processing and after some images have been discarded. The decision to discard an image can be based on different reasons: poor quality of the image, i.e., a blurred or backlit image, a duplicate motif, bad composition, or personal reasons [Zellhöfer 2012c].

Findings from a conducted survey show that such collections do not necessarily mirror the contents of a personal photo collection (see Section 8.3 and Zellhöfer [2012c]). Regarding this argument, UCID provides the most realistic data in comparison to the other collections. Nonetheless, the choice of collections reflects the current fields of primary research in CBIR and MIR, i.e., object recognition, large scale image retrieval, and image retrieval from personal photograph collections.

In total, 107,151 documents that are associated with 697 topics<sup>127</sup> are available for the evaluation discussed in Section 8.4. Depending on the collection, the topics can be disjoint or not. That is, the topics are considered disjoint if a document can only be regarded as relevant with respect to one topic. For each topic, its relevant documents are chosen as QBE documents for the same topic. All resulting QBE documents (Table 8.1, number of QBE Docs.) are then tested against the full collection. With this approach, a high number of queries is used that is most likely to vary in terms of query difficulty. Section 8.4.1 describes the experimental setup in more detail. As a consequence, 45,165 distinct queries against 107,151 documents are available for the evaluation.

Table 8.1: Characteristics of the Cranfield-based test collections

Collection & Collection Size	No. of QBE Docs.	Mean QBE Docs. per Topic	Standard Deviation QBE Docs. per Topic	No. of Topics	Disjoint Topics	Collection Type
<b>Caltech 101</b> 9,197	9,197	90.17	125.35	101	✓	Object recognition
<b>Caltech 256</b> 30,607	30,607	119.14	85.89	256	✓	Object recognition
<b>MSRA-MM</b> 65,443	3,400	100	0	68	✓	Web Image Sample
<b>UCID</b> 904	904	2.45	2.48	262	✓	Personal photos
<b>Wang</b> 1,000	1,057	105.70	16.75	10	✗	Stock photography
<i>Total:</i> 107,151	45,165			697		

### 8.3 Design of a Test Collection for Adaptive Retrieval Systems

Section 7.2 outlined some major issues of experiments based on the Cranfield paradigm when used in an adaptive IR scenario. Although, as suggested by Voorhees [2008], the discussed collections cover different usage scenarios, they do not support non-binary

<sup>127</sup>Please note that there might be overlaps between the documents contained in the collections. The same holds for the topics. Potential duplicates could not be removed automatically because a fully functioning duplicate detection algorithm was not available to the author. In consequence, all collections are evaluated separately in order to avoid any distortion of the results.



ground truths if one neglects the MSRA-MM collections. Unfortunately, the utility of the MSRA-MM collections is limited because of the missing original documents.

Besides the obvious subjectivity of relevance judgements, the support of a gradual relevance scale by the collection and the effectiveness metrics is particularly interesting in conjunction with the poset-based learning algorithm that is used in this dissertation. As described in Chapter 5, the learning algorithm tries to find a weighting scheme for a given CQQL query based on a user-defined preference poset. This poset organizes multiple documents at different relevance levels and has to be fulfilled by the new query. As a result, the new rank might contain a re-ordering of already relevant documents in a form that highly relevant document precede less relevant ones as well as new documents. Such a re-ordering is most likely appreciated by the user [Järvelin & Kekäläinen 2002]; however, it cannot be measured by metrics such as precision @  $n$  or MAP because they are only sensitive to the change of the amount of relevant documents in the examined rank.

The suggestion by Borlund [2003] to include users in the evaluation is desirable but is also subject to some practical restrictions. The inclusion of users in the actual evaluation is costly from both a monetary and temporal perspective. Furthermore, it limits the repeatability of an experiment and the amount of maximum user interactions with the system, e.g., how many relevance feedback iterations can be carried out by a person (see Section 7.2). To overcome these limitations, a test collection should support user simulations based on gradual relevance assessments.

In order to address these issues, Zellhöfer [2012c] proposes a test collection for the evaluation of adaptive, visual MIR systems: the *Pythia collection*. The following sections describe the core characteristics of the collection and its supplementary material. The subset of the collection used as a pilot task in ImageCLEF 2012 [Zellhöfer 2012e] is not discussed in this dissertation<sup>128</sup>.

#### 8.3.1 Creation and Origin of the Pythia Collection

In Section 8.2, the lack of realism of the presented collection was criticized in comparison to the contents of a personal photo collection. In contrast to the other collections, the Pythia collection is directly sampled from end-users. In other words, it contains images of poor quality, e.g., blurred or backlit ones; or duplicate motifs and bad compositions that are usually not part of other collections because most photographers are not likely to publish works in low (technical) quality.

The collection consists of 5,555 images that have been sampled from the hard disks of 19 photographers. In addition to taking random samples from their personal collections, each contributor had to complete a survey, which aimed to examine their photograph taking behavior, demographics, and computer usage (see Appendix C.1). Figure 8.2 lists the contribution of each photographer to the collection. In order to provide a variety in photographic motifs and style, the photographers have been chosen based on their different demographic background. The full demographic information

<sup>128</sup>The collection is also part at ImageCLEF 2013; see <http://www.imageclef.org/2013/photo>.

## 8 Evaluation of the Retrieval Effectiveness

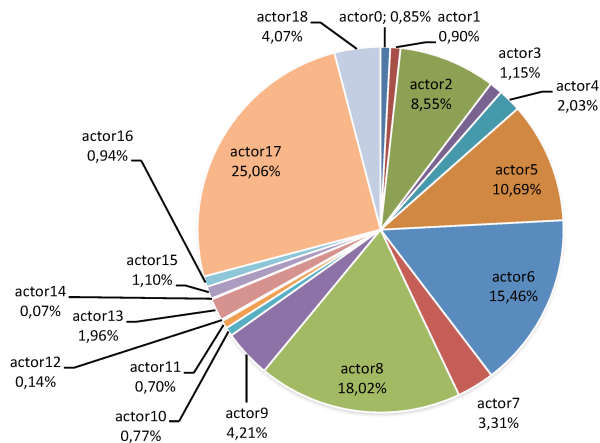


Figure 8.2: Contribution to the Pythia collection per photographer [Zellhöfer 2012c]

of the contributors and their answers to the survey are listed in Table C.1 in the appendix to this dissertation. Their demographics can be recapitulated as follows.

Roughly 50% of the contributors are female. Furthermore, most of them are fully employed and take photographs only at special events. They were born between 1944 and 1985 (median: 1982.5; mean: 1976.5). In consequence, “one can interpret the content of the collection as a mirror of a photographer’s lifespan with typical changing usage behaviors, cameras, topics, and places” [Zellhöfer 2012c, p. 3].

None of the photographers believe that they are colorblind. Only three photographers have taken classes relevant to the field of MIR, i.e., MIR or IR courses. Though, 47.37 % report that they have little knowledge of the principles of CBIR. As a result, the contributors are not biased in terms of an over-representation of “(M)IR-savvy” people. Nevertheless, computer and related scientists form the largest group amongst the contributors with 31.56 %, closely followed by economists with 26.32% (see Table C.1). The other contributors have varying academic and non-academic backgrounds. To conclude with, the photographers use the Internet between 91 and 120 minutes a day and know websites with a photograph sharing functionality such as Flickr or Facebook<sup>129</sup>, but use these websites seldom to share personal photographs.

### 8.3.2 Characteristics of the Pythia Collection

The Pythia collection contains images that have not been image processed extensively after being copied from the digital camera or mobile phone. That is, special effects such as sepia filters that are a built-in function of the used device or the like can occur. The Exif data of the original images has been preserved and is available for 100 % of the images, while 81.85 % also contain GPS data.

The collection contains a large amount of “tainted” images (ca. 16.8 %), i.e., images

<sup>129</sup><https://www.facebook.com/>

### 8.3 Design of a Test Collection for Adaptive Retrieval Systems

with a low quality or *motif duplicates* (see Table 8.2). A motif duplicate (MD) is defined as “a photograph that has been taken twice or more subsequently with the photographer’s intention to depict the same motif. These MD are characterized by the fact that the photographer mostly did not move but took the same motif again by correcting the rotation, translation, shutter speed, or the like of the camera because the composition did not look well.” [Zellhöfer 2012c, p. 4].

Table 8.2: Motif duplicates and their type [Zellhöfer 2012c, cf. Tab. 3]

Type	No. of Docs.	%
1. Motif Duplicates (total)	379	6.82
<i>Unmodified</i>	15	8.38
<i>Translated</i>	71	39.66
<i>Format Switch</i>	21	11.73
<i>Zoomed</i>	75	41.90
<i>Rotated</i>	26	14.53
<i>Sharpened</i>	18	10.06
<i>Altered (Lighting or Effect)</i>	12	6.70

Although the collection is relatively small in comparison to other MIR test collections (see Table 8.1), it is well understood that a comparison in terms of size alone is inappropriate when setting system-centric and test collections for adaptive MIR systems side by side. This is mainly due to the fact that collections which support the evaluation of adaptive MIR systems have to control more experimental variables [Borlund 2003; Voorhees 2008] and require much more expenses – in particular during the obtainment of the ground truth as shown in the next section.

To recapitulate, the last section has shown that the photographers contributing to the collection seldom upload their photos to the Internet. Thus, “one can expect to find different photos than on Flickr or the like – the common origin of most other available collections – in the presented data set” [Zellhöfer 2012c, p. 3]. Hence, the Pythia collection complements the other collections in terms of content and usage scenario.

#### 8.3.3 Obtainment of the Ground Truth for the Pythia Collection

The ground truth of the collection consists of two parts: a typical set of topics describing *visual concepts* and an *event-based description* of the images’ contents .

**Ground truth for visual concepts** To obtain the visual concept-based ground truth, 42 assessors were asked to judge each image’s relevance with respect to a topic (see Table 8.3) using a scale of 0 to 3, where 0 denotes irrelevance and 3 full relevance. A visual concept refers to the content of a photo depicting an object or a distinct scenery, e.g., a close-up of a flower or a market place. The choice of the topics is based on an analysis of the collection’s content [Zellhöfer 2012c]. In order to acquire relevance assessments from different user groups, each image’s relevance with respect to a topic was assessed by multiple assessors, i.e., each image’s relevance is judged by 2.69 assessors in average (standard deviation: 1.60). To maintain the origin of a relevance assessment with respect to a demographic or user group, each assessor completed the same survey as

## 8 Evaluation of the Retrieval Effectiveness

the contributors to the collection (see Appendix C.1). The results of this survey are available in Table C.2 in the appendix.

The characteristics of the assessors can be summarized as follows. The year of birth of the assessors ranges from 1979 to 1991 (median: 1987; mean: 1986.29). 33.33 % of the 42 assessors are female. Figure 8.3 illustrates the professional and mostly academic background of the assessors. The largest group consists of economists and is followed by IT-related fields of study.

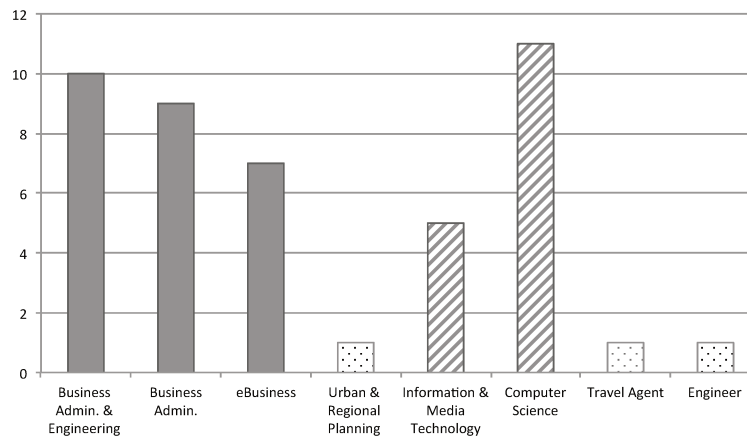


Figure 8.3: Job types of the assessors; economic background is marked in solid gray; IT is denoted by diagonal lines [Zellhöfer 2012e, Fig. 2]

On average, the assessors have only little knowledge in the fields of CBIR and MIR, with a slight trend towards being “informed outsiders”. Nine assessors have visited a class on multimedia retrieval, and 11 visited a lecture about the text-focused field of IR. They use the Internet between 91 and 120 minutes a day, use online photo sharing facilities less than a month, and take photos mostly at special events.

To sum up, the assessors can be considered IT-literate but not educated in the scientific fields related to this dissertation. In other words, they are (potential) end-users of MIR applications such as a CBIR similarity search but non-experts regarding CBIR and MIR techniques. Thus, one can assume their relevance assessments are not particularly biased towards the algorithmic interpretation of images based on low-level features.

**Ground truth for events** Acknowledging the recent discussion about the fact that images depict or illustrate events more often than visual concepts (e.g., a squirrel) that is reflected by the presented studies (see Tables C.1 and C.2 in the appendix); the literature, e.g., Amarnath & Jain [2011]; and the aforementioned MediaEval social event detection task (see Section 7.1), the Pythia collection includes an event-based ground truth. In contrast to the visual concept-based ground truth, the relevance assessments of the depicted events have been obtained directly from the contributing photographer. Because of the original photographer’s ability to associate a photo with a specific event, a binary relevance scale is used.

### 8.3 Design of a Test Collection for Adaptive Retrieval Systems

Table 8.3: Topics (visual concepts) of the Pythia collection

Name	Name
1. Beach and Seaside	17. Still Life
2. Street Scene	18. Church (Christian)
3. Statue and Figurine	19. Art Object
4. Asian Temple & Palace	20. Cars
5. Landscape	21. Ship / Maritime Vessel
6. Hotel Room	22. Airplane
7. People	23. Temple (Ancient)
8. Architecture (profane)	24. Squirrels
9. Animals	25. Sign
10. Asian Temple Interior	26. Mountains
11. Flower / Botanic Details	27. Monkeys
12. Market Scene	28. Birds
13. Submarine Scene	29. Trees
14. Ceremony and Party	30. Abstract Content
15. Theater / Performing Arts	31. City Panorama
16. Clouds	32. (Christ.) Church Interior

[Zellhöfer 2012e, cf. Tab. 7]

Table 8.4: Event distribution in the Pythia collection

WordNet Event	%
A. Conference	0.65
B. Event	0.36
C. Excursion	7.23
D. Flight	1.78
E. Holiday	77.86
F. Jubilation	0.49
G. Party	1.33
H. Rock Concert	8.70
I. Scuba Diving	1.03
J. Soccer	0.04
K. Visit	0.54

[Zellhöfer 2012e, Tab. 5]

Table 8.4 lists the main event classes (using terms from the WordNet<sup>130</sup> lexical database to avoid ambiguities) and their frequency of occurrence in the collection. The discrimination of events ranges from event class instances, e.g., a U2 concert at a given time, to general classes such as a rock concert. The associated events are not necessarily chronologically connected and might reoccur, e.g., a holiday trip to London can be found more than once. It is noteworthy that the bias towards holiday trips illustrated by Table 8.4 is not freely chosen but determined by the contents of the randomly picked images from the personal photo collections [Zellhöfer 2012c].

#### 8.3.4 Support for User Simulations

Section 7.2 outlined the utility of user simulations for the test of IIR systems but left their actual implementation open. This dissertation cannot describe all arguments and application techniques for user simulations. Instead, it sketches the design of a user simulation based on the data provided by the Pythia collection (see Section 8.5.1). A more thorough discussion of user simulations can be found in Azzopardi et al. [2011] and the aforementioned literature.

Although the inclusion of real users into the evaluation of IIR systems is – without doubt – desirable, the actual execution is often impractical [Mulwa et al. 2011]. As argued in Section 7.2, this is due to the high consumption of monetary and temporal resources, hardly controllable variables such as tiring and learning effects affecting the test subjects, and their limited repeatability. To compensate these issues, the Pythia collection offers persona-based ground truths that can be used to construct different user simulations.

<sup>130</sup><http://wordnet.princeton.edu/>

## 8 Evaluation of the Retrieval Effectiveness

As described in Section 6.1, the *persona* approach originated in the field of interaction design and was first presented by Cooper [1999]. The core idea behind personas is to assist interaction and software designers in comprehending the needs and objectives of the targeted user groups of the system in development. When transferred to the MIR domain, the most interesting characteristics of a persona are its information seeking behavior and strategies in conjunction with its relevance assessments.

By maintaining the demographic information of the assessors along with their relevance assessments (see Section 8.3.3), it becomes possible to define different ground truths based on commonalities between the assessors, e.g, their professional background and their CBIR expertise that can be eventually used to construct a persona. Roughly speaking, the resulting persona serves as a filter for all available relevance assessments. The derived persona-based ground truth can then be used as input for a user simulation as Figure 8.4 illustrates. In the presented scenario, the user simulation acts as a driver for the adaptive IR system and relies on the gradual relevance judgments of the persona to provide relevance feedback. A more detailed description of the utilization of this approach with data from the ImageCLEF 2012 pilot task on personal photo retrieval [Zellhöfer 2012e] is available in an article by Zellhöfer [2012f].

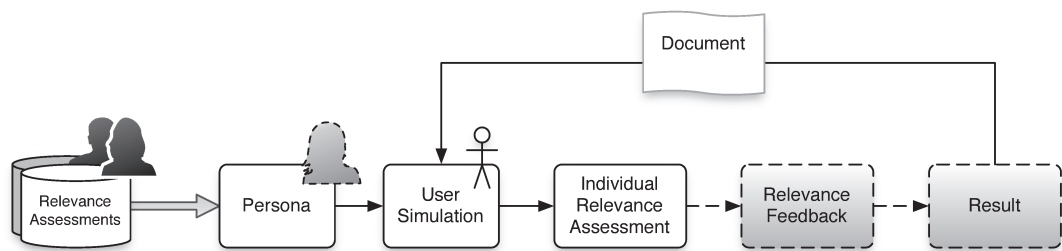


Figure 8.4: Position of a persona-based user simulation in the relevance feedback loop

In consequence, the derived user simulations are based on a model of real users as postulated by Cole [2011]. This realism includes erroneous relevance judgements that can occur during the interaction with an IIR system, e.g., if a user judges an irrelevant document accidentally as relevant without noticing the mistake. This discriminates the approach from other user simulations – namely Baskaya et al. [2011] and Baskaya et al. [2012] – that define artificial probabilities expressing how often a user simulation provides fallible relevance feedback. In other words, the approach of Baskaya et al. uses a simulation of erroneous RF based on a model that is derived from literature whilst the presented approach relies on a realistic sample of fallible RF, whose likelihood varies over the different topics and personas [Zellhöfer 2012f].

We agree with Tunkelang [2011] that the simulation of RF based on the collection's ground truth remains only a "rough approximation to reality" [Tunkelang 2011, p. 7]. The presented user simulation is limited to one aspect of a persona in the sense that it does not incorporate cognitive aspects of the user and different uses of the IIR system

such as alternating information seeking strategies (see Section 3.3). In fact, each persona in the context of this dissertation is reduced to a relevance judging entity in the simplest case<sup>131</sup>. However, the definition of a model addressing these aspects is not trivial and has not been accomplished yet [Cole 2011]. Thus, we believe that the discussed user simulation approach can be justified because it allows the evaluation of an IIR system using simple personas (i.e., multiple user groups) and their subjective relevance assessments that are, e.g., described by Borlund [2003] or Voorhees [2008].

### 8.3.5 The Collection in Relation to Adaptive IR Evaluation

Section 7.2 presented three major criticisms of the Cranfield tradition leading to three postulated revolutions, i.e., the cognitive, the relevance, and the interactive revolution [Robertson & Hancock-Beaulieu 1992]. Regarding the first revolution, more realistic topics are demanded. To address this issue, the Pythia collection contains topics for visual concepts and events that are, as the surveys show, often the motivation to take a photograph. However, it does not provide dynamic topics to reflect the dynamic nature of an IN. Providing such dynamic INs would ultimately mean to model the user's current situation and cognitive processes, which is not trivial [Cole 2011]. Hence, the combination of realistic topics (at least in the scope of a personal photo collection) and persona-based relevance assessments addresses realism and the subjectivity of relevance. The provision of different relevance assessments per persona and document address the second revolution. Furthermore, it extends the realism of the relevance judgements by using a gradual relevance scale. While multiple ground truths model the variation of relevance between user groups, non-binary relevance assessments can express different levels of relevance within a group.

The last revolution that deals with the importance of the user interaction throughout the IR process is addressed by the provision of means to construct user simulations (see above). Although user simulations are not part of the collection, they can be designed on the basis of the demographic data complementing the images, the persona-based ground truths, and an observation of the typical interactions per persona<sup>132</sup>, e.g., an exploratory followed by a directed search. Nevertheless, these user simulations cannot replace a simulated work task situation as advised by Borlund's IIR evaluation framework [Borlund 2003]. Though, the utilization of user simulations allows more control over the experiment and guarantees repeatability. Additionally, user simulations are not affected by usability issues, which might affect the user experience and therefore the evaluation outcome when a simulated work task is used.

With regard to the evaluation of adaptive IR systems, the Pythia collection can be considered as a hybrid between the Cranfield paradigm, Borlund's framework for IIR evaluation [Borlund 2003], and the recommendations of Voorhees [2008]. The collection offers static topics such as a Cranfield-based collection but aims at providing a more

---

<sup>131</sup>To simplify matters, more advanced interaction strategies of the user simulation used in this thesis are not covered at this stage. This aspect will be discussed in Section 8.5.1 along with the evaluation results.

<sup>132</sup>An actual study of MIR usage behavior and the construction of user simulations based on these observations falls out of the scope of this dissertation.

## 8 Evaluation of the Retrieval Effectiveness

realistic scope and a gradual relevance scale. In order to evaluate the adaptivity in terms of relevance feedback of an IIR system, user simulations can be constructed on the basis of the collection's ground truth. In principle, user simulations can also provide different interactions with the system such as the change of search strategy. As a result, the collection addresses part 1 and 3 of Borlund's IIR evaluation components (see Section 7.2) while staying feasible for the automated evaluation of an IIR system. This point becomes clear if one inspects the number of possible queries and the resulting number of RF iterations that are listed in Table 8.5. The number is caused by the experimental design sketched in Section 8.2.5, which uses each relevant document per topic as a QBE document in order to make an assertion about an IIR system's effectiveness based on different query documents.

In contrast to Borlund's suggestion, the collection does not offer pre-defined simulated work task situations, although it hints at the construction of such. Borlund's second component argues for the utilization of such work tasks. Given the scope of the Pythia collection, i.e., a personal photo collection, and the conducted survey, it becomes possible to define typical work tasks, e.g., the retrieval of photos illustrating a certain event. As argued in Chapter 7, a sample work task is used for the performance comparison of different GUIs (see Chapter 9), whereas the next sections discuss the effectiveness of CQQL in a user simulation-driven Cranfield-inspired setting.

Table 8.5: Characteristics of the Pythia test sets

Collection & Collection Size	No. of QBE Docs.	Mean QBE Docs. per Topic	Standard Deviation QBE Docs. per Topic	No. of Topics	Disjoint Topics <sup>133</sup>	Collection Type
Visual concept 5,555	13,602	425.38	493.76	32	X	Personal photos
Event 5,555	43	2.87	0.35	15	X	Personal event photos
Total:	5,555	13,645		47		

### 8.4 Retrieval Effectiveness of the CQQL Approach

In this part of the dissertation, the retrieval effectiveness (see Section 7.1) of CQQL as an implementation of the principle of polyrepresentation (PoP) is investigated.

As described in Section 3.2, the principle's core hypothesis is that a document is defined by different representations such as low-level features, textual annotations, user ratings, etc. These representations can be combined to form a cognitive overlap (CO), in which highly relevant documents are likely to be contained. This introduces an "intentional redundancy" [Ingwersen 1996] amongst the representations that is often tried to be avoided in CBIR systems [Eidenberger 2003; Deselaers et al. 2008].

The main idea behind the avoidance of redundancy is to only use representations that differ significantly during retrieval in order to pick the most appropriate and minimal

<sup>133</sup>Disjoint topics states if documents are only associated with one topic.



set describing a certain information need. This implies the assumption that only the usage of *dissimilar* representations can improve the retrieval performance. Thus, studies have been conducted that examine the correlation between such representations [Eidenberger 2003; Deselaers et al. 2008]. The latter focusses explicitly on the avoidance of redundancy, while the former incorporates an investigation of the retrieval performance of a set of representations or, in other words, low-level features.

The principle of polyrepresentation is contrary to this point of view. Interestingly, studies strengthen the hypothesis of the PoP in the field of IR [Skov et al. 2004; Larsen et al. 2006, 2009]. Additionally, the utility of the redundancy avoidance is also doubted by Oviatt in the domain of multimodal interaction who refers to linguistic research stating that there is no *true* redundancy between different modalities. Instead, they “contribute different and complementary semantic information” [Oviatt 1999, p. 79].

In consequence, the retrieval effectiveness of CQQL and the PoP has to be compared in detail to other techniques used in MIR to reveal whether the results presented in IR can be transferred to MIR. First results have been presented briefly by Zellhöfer [2012b]. Section 8.4.1 extends the aforementioned work by defining a relevance feedback-supportive experimental setup and retrieval metrics of the conducted experiments.

In order to establish a baseline, the retrieval effectiveness of single representations is discussed in Section 8.4.2.

To begin with the main experiments, Section 8.4.3 compares polyrepresentative approaches against other techniques fusing multiple representations to retrieve relevant documents. Section 8.5 extends the experiment’s scope to a relevance feedback-assisted retrieval scenario based on PrefCQQL. To complete the experimental description, Section 8.5.1 focusses on the reproducibility of the results because of the dependence of the utilized weight learning algorithm (see Section 5.4.5) on random numbers.

### 8.4.1 Experimental Setup for Retrieval Effectiveness Evaluation

In order to assess the effectiveness of a retrieval technique realistically, Voorhees argues that “a candidate retrieval technique [should] be consistently better than the alternatives on a variety of test collections before concluding the effect is real.” [Voorhees 2008, p. 1884]. Thus, all conducted experiments described below are carried out with 6 different collections which are listed in Table 8.6<sup>134</sup>. All collections have been described in Section 8.2 and 8.3. The general objective of the experiments reflecting a typical query-by-example (QBE) scenario can be roughly sketched as follows:

*How effectively will the retrieval system return relevant documents  $\mathbf{R}_{\mathbf{q}_i}^+$  from a given document collection  $\mathbf{D}_i$  using a pre-defined matching function  $\mathbf{eval}$  and a set of QBE documents  $\mathbf{Q}_i$  in average?*

<sup>134</sup>Please note that this list does not include the event-based ground truth available for the Pythia collection (see Section 8.3.3). In accordance to Voorhees’ argument, we have not included this collection because none of the others feature a similar ground truth that would allow the deduction of valid conclusions.

## 8 Evaluation of the Retrieval Effectiveness

To be more precise, each collection  $D_i$  listed in Table 8.6 defines an total number of QBE documents  $|Q_i|$  that are associated with a number of topics  $|T_i|$ . The QBE documents are directly taken from the ground truth for each topic. In other words, each document that has been judged relevant for a given topic is used as a QBE document for this topic in order to evaluate a matching function's effectiveness for this topic with varying QBE documents. Table 8.6 also shows the average number of QBE documents available for each topic and their standard deviation.

For instance, there are 9,197 documents in the Caltech 101 collection ( $|D_i|$ ), where each one is relevant to only one of the (disjoint) 101 topics ( $|T_i|$ ). Consequently, there are also 9,197 QBE documents ( $|Q_i|$ ) that are used for querying the collection.

In total, the examination of a matching function requires  $|D_i| * |Q_i|$  matching operations, e.g., a CQQL-based matching between each QBE document  $q_n$  and all documents  $d_j \in D_i$ . The computational costs of the experiments are described in more detail below.

The retrieval effectiveness of a given matching function for a QBE document and a topic can then be calculated. As with most IR evaluations [Croft et al. 2009, cf. p. 319], macro-averaging is used to calculate the average effectiveness over all QBE-topic combinations using the nDCG metric (see Definition 8.8) at various cut-off levels. The nDCG metric is used because of its support of gradual relevance assessments that are supported by the MSRA-MM and Pythia collections. Furthermore, the PrefCQQL approach is expected to produce re-orderings of already relevant documents in order to present highly relevant documents before less relevant ones based on its poset-based relevance feedback. Such a re-ordering cannot be measured by other metrics, although this behavior is desirable from a user's perspective [Järvelin & Kekäläinen 2002]. The usage of cut-off levels is motivated by the evaluation approach of CQQL, which calculates a similarity score for each document in a given collection (see Section 4.2) and returns all documents in the resulting total order.

Table 8.6: Overview over the examined test collections

Collection & Collection Size ( $ D_i $ )	No. of QBE Docs. ( $ Q_i $ )	Mean QBE Docs. per Topic	Standard Deviation QBE Docs. per Topic	No. of Topics ( $ T_i $ )	Disjoint Topics <sup>135</sup>	Collection Type
Caltech 101	9,197	90.17	125.35	101	✓	Object recognition
Caltech 256	30,607	119.14	85.89	256	✓	Object recognition
MSRA-MM	65,443	100	0	68	✓	Web Image Sample
UCID	904	2.45	2.48	262	✓	Personal photos
Wang	1,000	105.70	16.75	10	✗	Stock photography
Pythia	5,555	425.38	493.76	32	✗	Personal photos
<i>Total:</i>	<i>112,706</i>	<i>58,767</i>		<i>744</i>		

Table 8.7 shows an excerpt of the examined matching functions, including their origin. Generally speaking, the matching functions using more than one representation

<sup>135</sup>Disjoint topics states if documents are only associated with one topic (see Section 8.2.5).

can be subdivided into three groups:

1. standard aggregation functions, such as the arithmetic average or the maximum;
2. polyrepresentation-motivated ones that form a cognitive overlap of various representations; and
3. matching functions that are based on recommendations from the literature.

All examined matching functions can be found in an overview, which also describes the used notation, in Appendix B.1.

In total, 17 matching functions relying on multiple representations are evaluated in terms of their retrieval effectiveness in Section 8.4.3 (main experiment I). In order to establish a baseline, 15 separate representations are examined on their own in Section 8.4.2. The weighted variants of these matching functions and their performance during PrefCQQL-based relevance feedback are discussed in Section 8.5 (main experiment II). The needed tools to reproduce the presented results are available as a supplement to this text (see Appendix E).

To facilitate reading, all of the following sections have the same structure, starting with a presentation of the experimental results, which is followed by a short discussion of the results. Section 10.1 relates the findings to the results obtained from the user study described in Chapter 9.

Table 8.7: Examined matching functions (excerpt from Appendix B.1)

Matching Function	Arithmetic/Logical Structure	Origin
Arithmetic mean	$\frac{1}{n} \sum_{i=1}^n R_i$ ; $n = 15$	Standard aggregation function
Geometric mean	$\sqrt[n]{\prod_{i=1}^n R_i}$ ; $n = 15$	Standard aggregation function
Minimum	$\min(R_1, \dots, R_n)$ ; $n = 15$	Conjunction variant in fuzzy logic, standard aggregation
Maximum	$\max(R_1, \dots, R_n)$ ; $n = 15$	Disjunction variant in fuzzy logic, standard aggregation
Conjunction	$\bigwedge_{i=1, \theta_i}^n R_i$ ; $n = 15$	Principle of polyrepresentation, cognitive overlap in CQQL
Disjunction	$\bigvee_{i=1, \theta_i}^n R_i$ ; $n = 15$	Complementary operator to the cognitive overlap in CQQL
Semantic Group	$\bigwedge_{\theta_i} \langle (\bigvee_{\theta_j} (R_3, R_9)), (\bigvee_{\theta_l} (R_2, R_3, R_{10}, R_{12}, R_6, R_1, R_5, R_4, R_7)), (\bigvee_{\theta_k} (R_8, R_{11}, R_{15}), \bigvee_{\theta_l} (R_{13}, R_{14})) \rangle$	Deselaers et al. [2008]
Eidenberger conjunction, var. 1	$\bigwedge_{\theta_i} (R_5, R_7, R_8, R_{11})$	Eidenberger [2003]
Eidenberger disjunction, var. 1	$\bigvee_{\theta_i} (R_5, R_7, R_8, R_{11})$	Eidenberger [2003]

$R_i$  denotes a representation as referenced in Table 2.1,  $|R|$  is the total number of representations. The variable  $\theta$  indicates a weight. Initially, all representations are equally weighted.

## Computational Costs of the Experiments

Table 8.8 shows the total processing duration for all matching functions including 5 relevance feedback iterations (in average) measured on an Apple MacPro (version 4,1) with 2 Quad-Core Intel Xeon CPUs with 2.26 GHz and 8 GB RAM running Mac OS X

## 8 Evaluation of the Retrieval Effectiveness

10.6.8 on the internal hard disk. The long runtime is mainly due to the missing indexing functionality in the used prototypical Pythia MIR system and the evaluation strategy of CQQL, which relies on the calculation of a similarity score for each document  $d \in D_i$  with respect to a QBE document  $q \in Q_i$ . That is, a total of 1,321,312,784 matching operations ( $MOPS_{total}$ ) have to be carried out to assess the effectiveness of one matching function over all collections as shown in the following calculation (refer to Table 8.8 for the variable values):

$$MOPS_i = |D_i| * |Q_i| \quad (8.10a)$$

$$MOPS_{total} = \sum(MOPS_i) \quad (8.10b)$$

In the worst case, each matching operation results in 19 similarity calculations (one for each representation, see Table 2.1) between each QBE document's representations and each document's representations in the collection. For instance, if the cognitive overlap of all available representations  $|R|$  is used as matching function, a total of 25,104,942,896 similarity calculations ( $simCalc_{total}$ ) occur:

$$simCalc_{total} = MOPS_{total} * |R| \quad (8.11a)$$

$$simCalc_{total} = 1,321,312,784 * 19 = 25,104,942,896. \quad (8.11b)$$

The proportion between processing duration and MOPS differs for each collection because of the varying content, the resulting XML output, its parsing costs, and swapping operations due to the memory usage of the computations.

As described in Section 2.3.2, these similarity calculations are very expensive in terms of computational costs. After each document's similarity to the query has been calculated, the results have to be sorted. Afterwards, an optional relevance feedback step triggers the weight learning, causing a re-evaluation of all documents (see Section 5.4.5). Instead of costly similarity calculations, this re-evaluation only requires a sorting of the results.

Table 8.8: Processing duration of the analyzed collections

Collection	No. of Docs. ( $ D_i $ )	No. of QBE Docs. ( $ Q_i $ )	MOPS ( $ D_i  *  Q_i $ )	No. of Topics	Processing Duration (approximately)
<b>Caltech 101</b>	9,197	9,197	84,584,809	101	1 week, 20 hours
<b>Caltech 256</b>	30,607	30,607	936,788,449	256	1.9 months
<b>MSRA-MM</b>	65,443	3,400	222,506,200	68	24 days
<b>Pythia</b>	5,555	13,602	75,559,110	47	9 days, 21 hours
<b>UCID v2</b>	904	904	817,216	262	5 hours, 40 minutes
<b>Wang</b>	1,000	1,057	1,057,000	10	13 hours
<i>Total:</i>	<i>112,706</i>	<i>58,767</i>	<i>1,321,312,784</i>	<i>744</i>	<i>3.4 months</i>

Although the Pythia MIR system prototype uses various optimizations in form of a read cache, i.e., serialized document representations are kept in memory once they are read from the hard disk, multi-threaded processing of the weight learning, memory sharing whenever possible, and compiler optimizations, there is certainly room for improvements that will be discussed in Section 10.3.

### 8.4.2 Evaluation of Single Representations

The PoP is characterized by its core hypothesis, which states that the effectiveness of an IR system can be increased when multiple representations of a document are combined (see Section 3.2). In order to examine whether this hypothesis holds true in the field of MIR, it is first compared to the results obtained by using the available representations on their own.

Consequently, this section serves *only* to establish a baseline for the experiments involving multiple representations described in Section 8.4.3. Furthermore, it compares the relative performance of the representations shown in Table 2.1 using the nDCG measure at the cut-off levels 10, 20, 30, and 100. A full discussion on the performance of different single representations and possible improvements at feature extraction or matching level falls out of the scope of this dissertation.

### Experimental Results

Figures 8.5-8.7 visualize the results of the QBE experiments described before in form of stacked bar charts for each collection listed in Table 8.8. The figures contain the retrieval metrics for all extractable representations<sup>136</sup>, whose number varies over the collections, because Exif and IPTC-based representations could only be extracted if they were present in the image documents. The representations are described in more detail in Table 2.1. To avoid a distortion of the results due to the higher level Exif/IPTC-based representations available only in some collections, these representations are neglected in the following sections, if not stated otherwise.

Hence, the experiments examine global low-level features only. As said in Section 2.3.1, local low-level features are expected to often provide a better retrieval effectiveness than global ones but are computationally more complex. In any case, to verify the PoP hypothesis, the types of the used features are only marginally important because the PoP does not make assertions about a representation's nature besides its functional or cognitive difference with respect to other representations (see Section 3.2).

The  $x$ -axis of the plots shows the matching function's ID, which is properly labeled in the accompanying legend. The "It. 0" in the legend indicates the 0th RF iteration, i.e., no relevance feedback has been given. The  $y$ -axis shows the metrics' respective values on an automatically fitted linear scale to visualize the relation of the metrics amongst their cut-off levels and between the different representations. All retrieval metrics have been calculated with *trec\_eval*<sup>137</sup>. The complete results in numerical form are available on the attached storage medium (see Appendix E).

<sup>136</sup>The region-based color histogram [Balko & Schmitt 2012] appears twice in each plot because of its implementation that calculates separate similarities for the center and border region of an image document, resulting in different retrieval effectiveness scores for each region.

<sup>137</sup>Version 9.0 as obtained from [http://trec.nist.gov/trec\\_eval/](http://trec.nist.gov/trec_eval/) at October 6th, 2012 and compiled with GCC 4.2.1 (build 5658) under Mac OS X 10.6.8.

## 8 Evaluation of the Retrieval Effectiveness

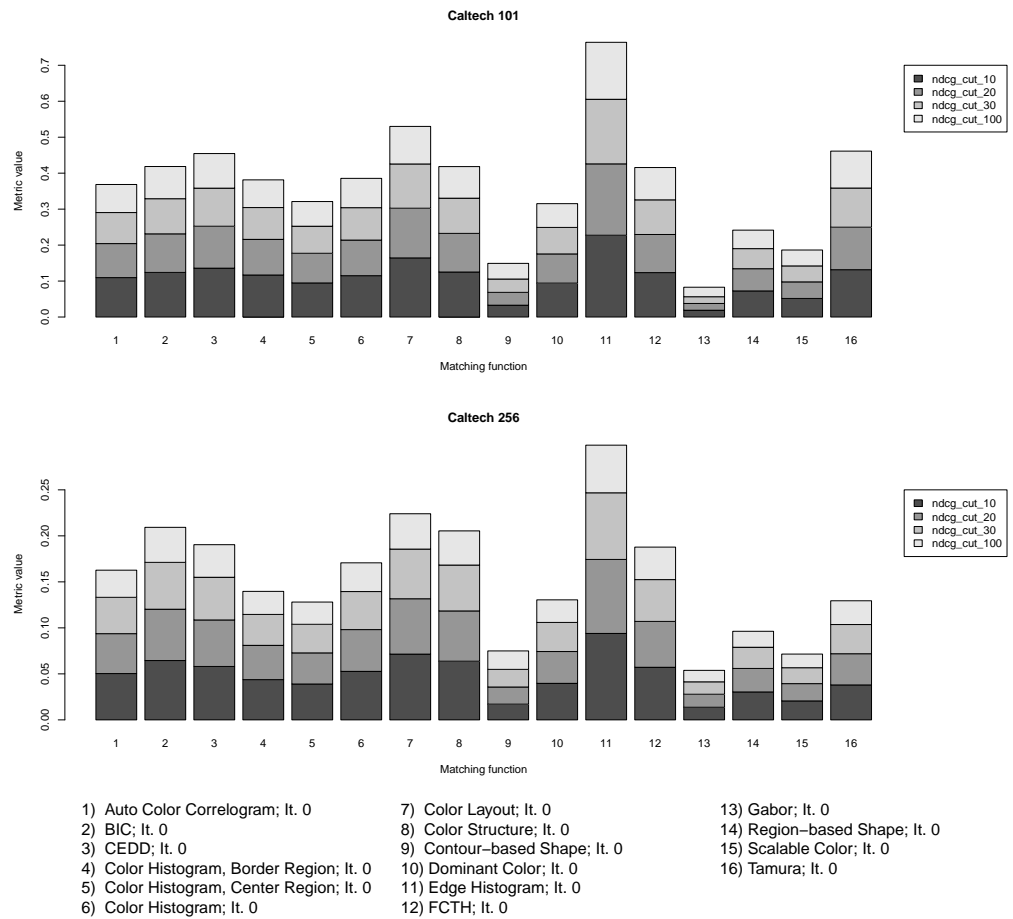


Figure 8.5: Performance comparison of single representations, part 1

## 8.4 Retrieval Effectiveness of the CQQL Approach

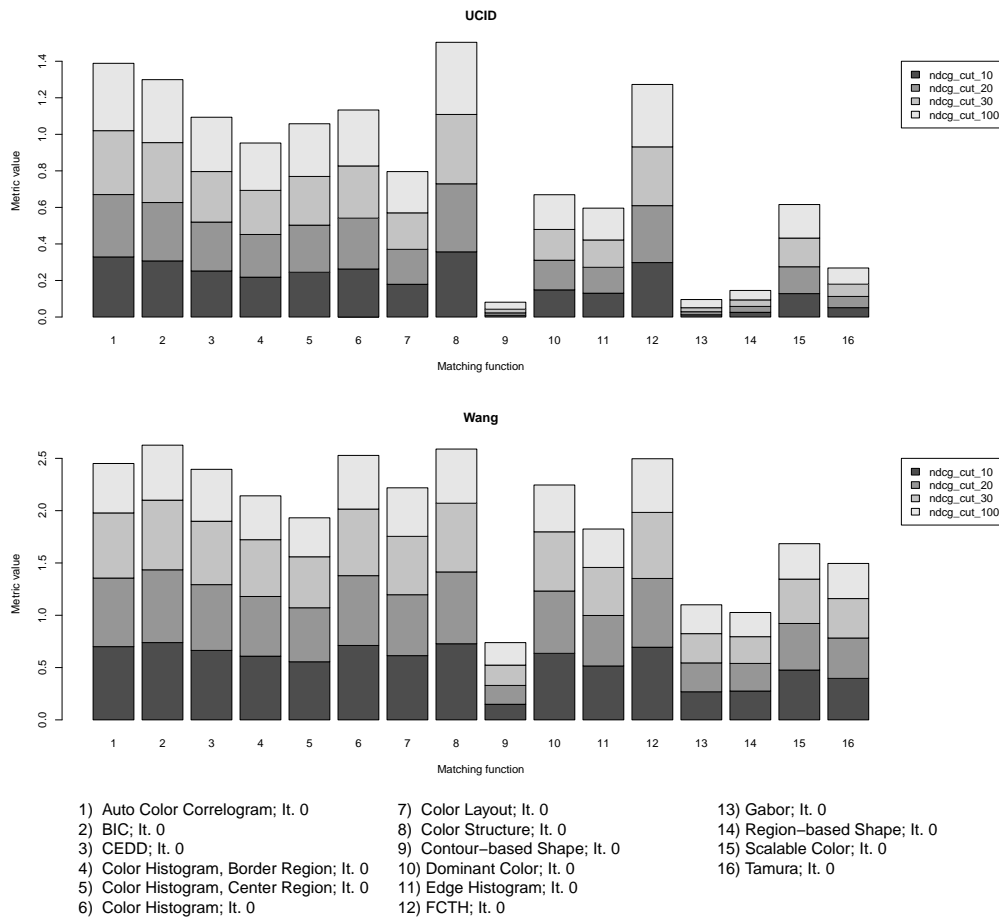
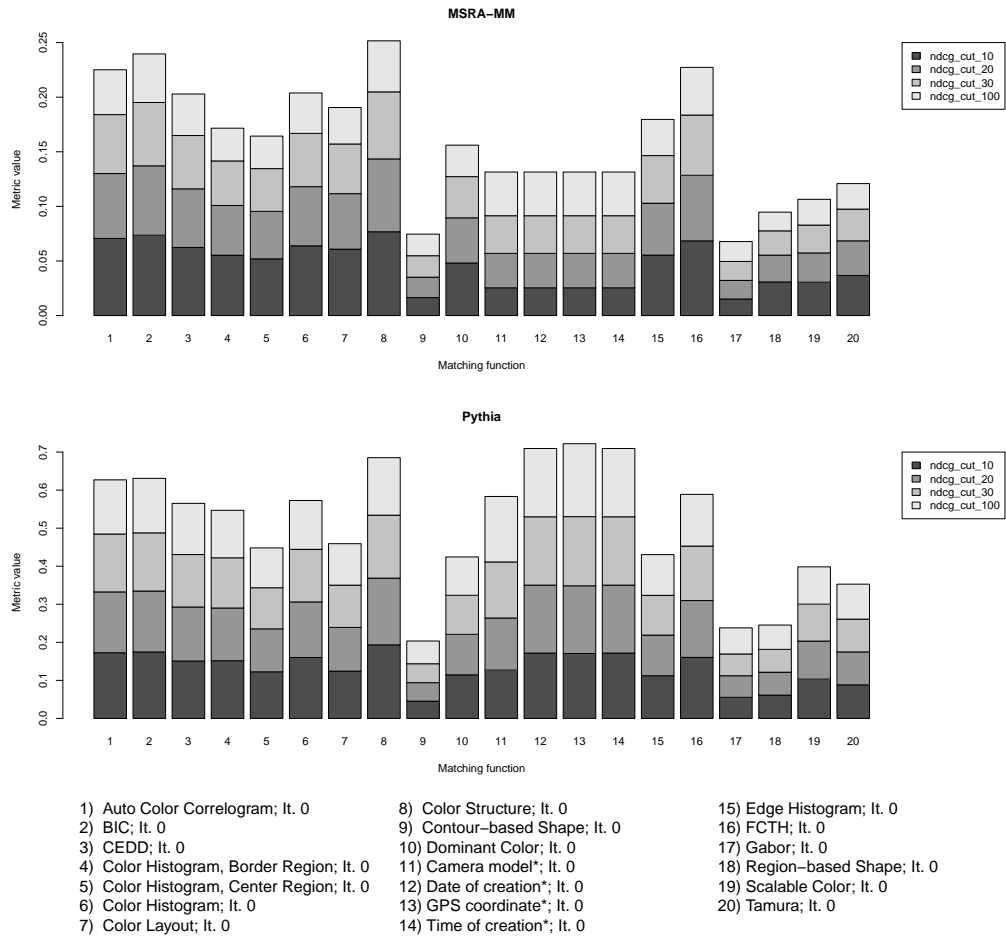


Figure 8.6: Performance comparison of single representations, part 2

## 8 Evaluation of the Retrieval Effectiveness



\* indicates Exif-based high-level features.

Figure 8.7: Performance comparison of single representations, part 3



## Discussion

The performance of the color histogram (RGB, 512 bins; see Figures 8.5-8.7 [7]) is particularly interesting because it is generally regarded as a “reasonably good baseline for general color photographs” [Deselaers et al. 2008, p. 15] in the CBIR/MIR literature. Hence, it is also used in this dissertation as the general baseline, against which the performance of other representations is measured.

Although the color histogram can be considered the “the most traditional way of describing low-level colour properties of images” [Del Bimbo 1999, p. 24] and is generally regarded as a good baseline, its effectiveness as a representation of the visual content of an image is only mediocre in the case of the examined collections (see Table 8.9). The color histogram clearly belongs to the best 50% of the tested representations, with a slight trend towards the best 25%. For instance, as Figure 8.6 shows, it gets under the best 19% representations with the Wang collection. Alas, the nDCG values for this collection are all relatively high in comparison to the other collections. This indicates that retrieval tasks can be solved easily for this collection even if only one representation is available.

Table 8.9: Best performing representations and rank of color histogram

Collection	Best	2nd best	3rd best	4th best	Rank of Color Histogram
<b>Caltech 101</b>	Edge Histogram	Color Layout	Tamura	CEDD	8 of 16
<b>Caltech 256</b>	Edge Histogram	Color Layout	BIC	Color Structure	9 of 16
<b>MSRA-MM</b>	Color Structure	BIC	FCTH	Auto Color Correlogram	7 of 20
<b>Pythia</b>	GPS coordinate	Date of creation	Time of creation	Color Structure	9 of 20
<b>UCID v2</b>	Color Structure	Auto Color Correlogram	BIC	FCTH	5 of 16
<b>Wang</b>	BIC	Color Structure	Color Histogram	FCTH	3 of 16

The total number of ranks is determined by the available representations.

Regarding their general effectiveness, there is no clear winner amongst the representations. Table 8.9 shows the four best performing representations in conjunction with the rank of the color histogram baseline for orientation. In case of MSRA-MM, UCID v2, and Wang, it can be observed that the best representations all rely on color (see Table 2.1 for the types of the representations). A overall excellent performance of the edge histogram as reported by Eidenberger [2003] could not be reproduced.

Instead, the edge histogram performs worse than the baseline if one ignores both Caltech collections. The relatively weak performance of the color-based representations with these collections is most likely due to the artifacts present in the images described in Section 8.2. For instance, space that occurs during the translation or scaling of image contents is commonly filled monochromatic (see Figure 8.1; first two images at the top). These large regions in solid color, which do not contribute to the visually interesting part of an image, i.e., the actual image’s content, can irritate color-based representations such as the color histogram. In consequence, the edge histogram clearly outperforms

## 8 Evaluation of the Retrieval Effectiveness

the other representations when used with the Caltech collections as it is robust against such artifacts, which are not present in the other collections.

Comparably, the collections used by Eidenberger [2003] are most likely biased towards the edge histogram. For instance, the author uses a collection of coats-of-arms images with 426 documents. Typically, coats-of-arms are characterized by sharp edges and contours. Additionally, Eidenberger uses an unspecified subset of the Corel collection with 260 documents and a data set called “Brodatz” with 112 documents. Unfortunately, the data sets are no longer available and therefore could not be included in this dissertation. Whether the subset of the Corel collection overlaps with the Wang collection [Wang et al. 2001], which has taken from the same collection, cannot be ascertained.

The Pythia collection takes a special place because three high-level representations are ranked best. This is mainly due to the contents of the collection (see Section 8.3) and their origin. These findings indicate that the contributing photographers indeed took photographs at certain events, which can be best described by spatial and temporal proximity.

Table 8.10: Worst performing representations

Collection	Worst	2nd worst	3rd worst	4th worst
<b>Caltech 101</b>	Gabor	Contour-based Shape	Scalable Color	Region-based Shape
<b>Caltech 256</b>	Gabor	Scalable Color	Contour-based Shape	Region-based Shape
<b>MSRA-MM</b>	Gabor	Contour-based Shape	Region-based Shape	Scalable Color
<b>Pythia</b>	Contour-based Shape	Gabor	Region-based Shape	Tamura
<b>UCID v2</b>	Contour-based Shape	Gabor	Region-based Shape	Tamura
<b>Wang</b>	Contour-based Shape	Region-based Shape	Gabor	Tamura

In contrast to the most effective representations, the situation at the lower end of the effectiveness ranking of the representations is much clearer as Table 8.10 shows. The weak performance of the contour- and region-based shape representation is not surprising because they both rely on an automatic segmentation of general images, which is still an unsolved problem [Deselaers et al. 2008]. Interestingly, this conflicts with Eidenberger’s findings, who claims that region-based shape “is a good descriptor in any situation that can be applied to any type of media content” [Eidenberger 2003, p. 486], which is most likely due to the test collections that have been examined by the author (see above).

The Gabor representation is also performing poorly over all collections. This finding is surprising because Gabor features are widely used as an effective descriptor of texture features in both CBIR and pattern recognition, as well as the neighboring research fields. Unfortunately, the actual implementation of low-level features lies outside the scope of this dissertation. Thus, this effect is not investigated further here and remains an issue for further research. For the remainder of this thesis, this observation is ac-

cepted as a phenomenon occurring with the examined collections and the used Pythia MIR system implementation.

As observed before, the artifacts in the Caltech collections seem to negatively affect the performance of the scalable color representation. For these collections, it even undercuts the weak performance of the region-based shape representation.

In order to assess the retrieval effectiveness of a combination of multiple representations, it is reasonable to compare the results against more than the traditionally used color histogram. As the conducted experiments show, the color structure representation provides a complementary viable baseline because it is placed amongst the first four performing representations in five out of six collections (see Table 8.9). In fact, this representation would obtain a good average effectiveness rank of 2.5 over all collections when Pythia's high-level representations are ignored. As said before, these representation distort the results and complicate a direct comparison of the collections. To complete the picture, the varying best performing representation for each collection is used.

### 8.4.3 Evaluation of Combined Representations – Main Experiment I

After the last section has examined the performance of single representations, it is time for the evaluation of the retrieval effectiveness of combinations (or fusions) of multiple representations as the exploitation of different representations to improve the retrieval effectiveness forms the core of the principle of polyrepresentation's hypothesis (see Section 3.2).

In order to draw valid conclusions from the statistics and retrieval metrics presented further in this text, some basic prerequisites need further attention. After these prerequisites and auxiliary conditions have been clarified, the retrieval effectiveness of different representation fusion methods is examined to establish a common baseline for the main experiments.

In the main experiments, matching functions that follow the PoP are compared to other representation fusion techniques to examine the validity of the PoP hypothesis, when implemented with CQQL.

As before, all experiments follow the design presented in Section 8.4.1. The nDCG values at various cut-off levels have been calculated with *trec\_eval*<sup>138</sup>.

#### Prerequisites – Treatment of Unavailable Representations

Whenever multiple representations have to be fused with the help of a matching function, the treatment of unavailable representations deserves special attention.

For instance, consider a matching function similar to Matching Function 13 that calculates the arithmetic mean of the similarities between a document in the collection and a QBE document using the color histogram, the texture, and the edge histogram

<sup>138</sup>Version 9.0 using standard settings as obtained from [http://trec.nist.gov/trec\\_eval/](http://trec.nist.gov/trec_eval/) at October 6th, 2012 and compiled with GCC 4.2.1 (build 5658) under Mac OS X 10.6.8.

## 8 Evaluation of the Retrieval Effectiveness

to estimate the probability of relevance (POR) of a document. Given that the collection contains 80 % colored and 20 % grayscale images, one might be confronted with the situation where color histograms might not be extractable for the latter because of the implementation of the feature extractor. As a result, the similarity regarding the color histogram representation is unavailable for 20 % of the documents, i.e., one cannot ascertain whether a document is similar or dissimilar to a given query regarding this representation. In other words, the document's similarity is unknown<sup>139</sup>.

As described in Section 2.2.2, the relational model in databases deals with such situations by introducing a 3-valued logic that extends traditional true or false judgments with "unknown" (or *NULL* in SQL).

Although CQQL has a strong DB background, it does not address the handling of such *NULL* values. As presented in Section 4.4, the evaluation of an atomic condition in CQQL is expected to yield a value out of  $[0, 1]$  and not *NULL*. However, not only CQQL is affected by this assumption. In fact, every approach that relies on an arithmetic evaluation to determine the POR is affected because *NULL* values cannot be processed *directly* during an arithmetic evaluation relying on values out of  $[0, 1]$ . Comparable problems are known during data fusion in data warehousing [Bleiholder & Naumann 2009; Bleiholder et al. 2011].

As a result, one has to develop a strategy to deal with *NULL* values. First, it might be appropriate to ignore *NULL* values. Alternatively, it might be feasible to set *NULL* values to 0. In any case, these different strategies affect the ranking of the result documents. Reconsider the example given above.

Table 8.11: Sample rank, sorted by arithmetic average

Document	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	Sum	Arithmetic Average	Arithmetic Average, <i>NULL</i> ignored
<i>d</i> <sub>1</sub>	0.25	0.60	0.25	1.10	0.37	0.37
<i>d</i> <sub>2</sub>	0.25	0.60	<i>NULL</i>	0.85	0.28	0.43

Table 8.11 shows the ranking of two documents *d*<sub>1</sub> and *d*<sub>2</sub> with the three representations *R*<sub>*i*</sub>, interpreting *NULL* values as 0. The documents are ordered by the value of their arithmetic average<sup>140</sup>, i.e.,

$$\frac{1}{n} \sum_{i=1}^n R_i \quad ; \quad n = 3. \quad (8.12)$$

If *NULL* values are ignored, the ranking changes as Table 8.12 illustrates. This is due to the lower value of  $n = 2$  for *d*<sub>2</sub>.

This example shows the effects of two basic *NULL* value handling strategies, i.e., *NULL* ignorance and the *pessimistic mapping*. The latter is called pessimistic because the mapping strategy assumes maximal dissimilarity, expressed by setting the similarity

<sup>139</sup>This example can be easily extended to the MIR domain, e.g., a query might contain the abstract of a text; however, only textual documents in the multimedia collection contain abstracts, while images do only contain a description.

<sup>140</sup>The formula uses the notation presented in Appendix B.1.

Table 8.12: Sample rank, sorted by arithmetic average with NULL ignorance

Document	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	Sum	Arithmetic Average	Arithmetic Average, NULL ignored
$d_2$	0.25	0.60	NULL	0.85	0.28	0.43
$d_1$	0.25	0.60	0.25	1.10	0.37	0.37

score to 0 when an unknown value is observed. Its logical counter-part is the *optimistic mapping* that sets unknown values to 1, declaring full similarity to the query. There are many more strategies to choose from, e.g., one might choose a *compromise* by mapping NULL values to 0.5, or one can map unknown values to the average similarity of the documents in the collection to the query.

Using weighted CQQL, it also becomes possible to exclude unknown atomic condition by appropriately setting the associated weighting variable (see Section 4.3). Alternatively, one can map the NULL value to the arithmetic neutral element depending on the evaluation of the used CQQL condition. For instance, 1 in the case of a CQQL conjunction (see Definition 4.13). Figures 8.8 and 8.9 show how this approach affects the retrieval effectiveness of CQQL. In the figures, NULL values are replaced with 0 and 1 respectively if stated explicitly. For the sake of completeness, Table 8.13 lists the amount of missing representations in the examined collections. As before, Exif- and IPTC-based are neglected because they are available only in some documents and collections.

To conclude, all following experiments use the *pessimistic mapping strategy* although it corresponds to the worst-case performance of the CQQL-based matching functions. Nevertheless, the pessimistic strategy allows a consistent mapping of unknown values to a similarity score no matter which matching function is used.

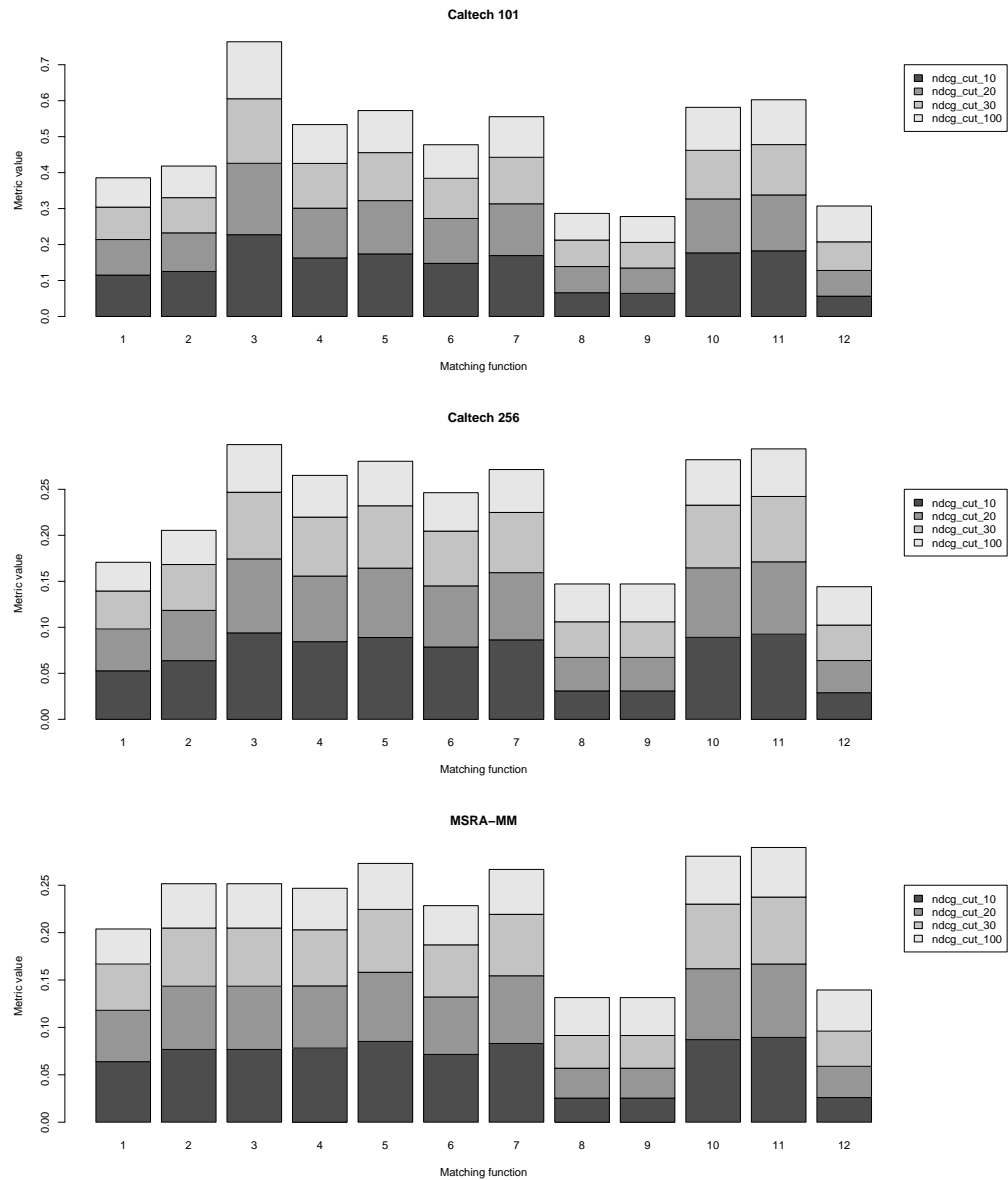
Table 8.13: Missing representations per collection

Collection	Contour-based Shape	Region-based Shape
Caltech 101	4,102 (44.60%)	1 (< 0.01%)
Caltech 256	14,850 (48.52%)	14 (< 0.01%)
MSRA-MM	44,169 (67.49%)	69 ( $\approx$ 0.1%)
Pythia	1,398 (25.17%)	7 ( $\approx$ 0.01%)
UCID v2	292 (32.30%)	0 (0.00%)
Wang	345 (34.50%)	1 (0.01%)

### Prerequisites – Distribution of the Obtained Retrieval Metric Values

The necessity of significance tests has been justified in Section 8.1.4. In IR, a typically used significance test is the Student's t-test that assumes normal distribution of the compared samples and has been shown to be very robust [Smucker et al. 2007; Croft et al. 2009]. However, the distribution of the nDCG metric's value is not normally distributed for every matching function examined in this dissertation. For instance, Table 8.14 clearly shows that the null hypothesis (i.e., the sample has a normal distribution) can be rejected regardless of the normality test used considering the nDCG results of the arithmetic mean (best features; see Matching Function 14) on the Caltech 101 collection.

## 8 Evaluation of the Retrieval Effectiveness

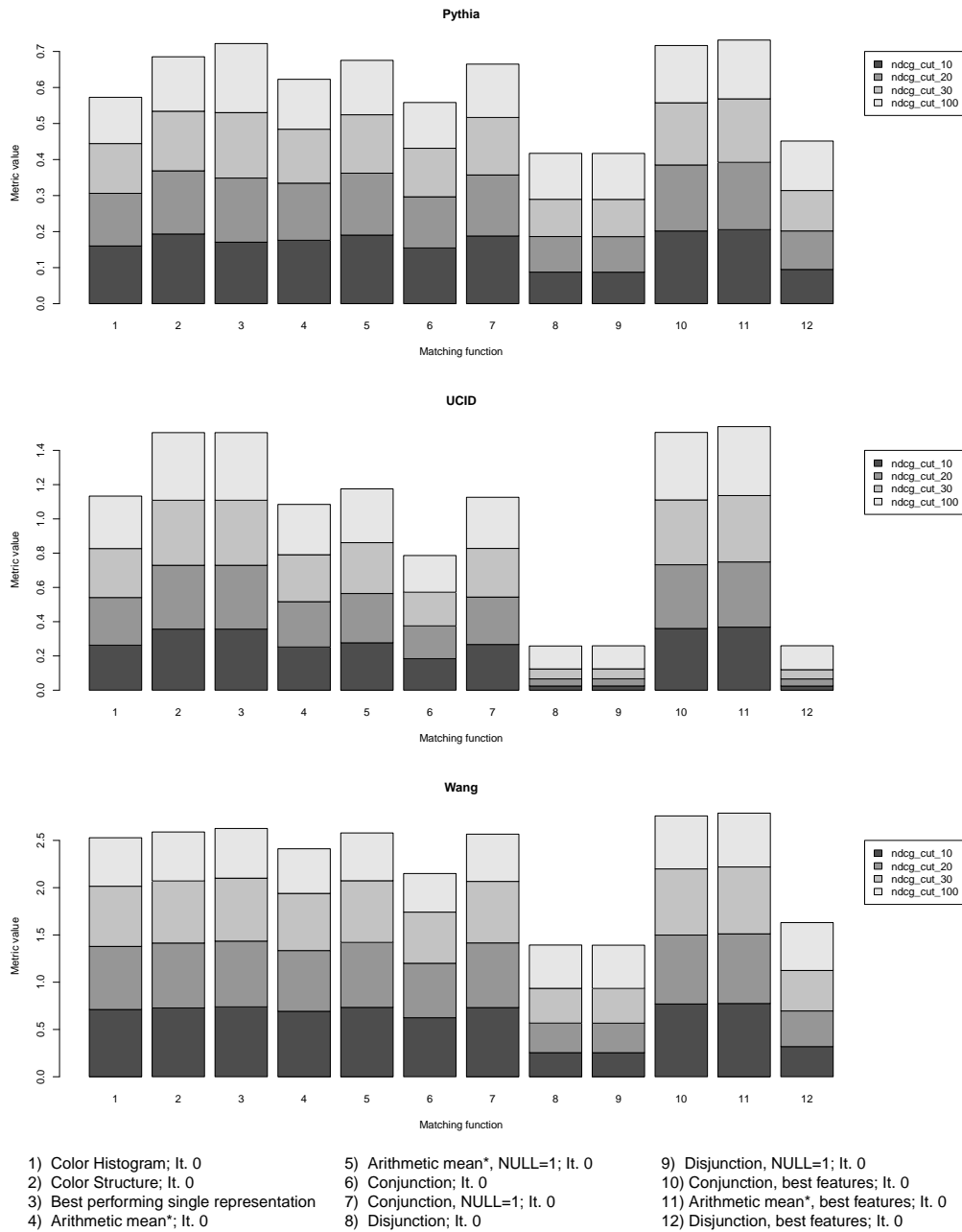


- |  |                                    |  |
|--|------------------------------------|--|
| 1) Color Histogram; It. 0                | 5) Arithmetic mean*, NULL=1; It. 0 | 9) Disjunction, NULL=1; It. 0              |
| 2) Color Structure; It. 0                | 6) Conjunction; It. 0              | 10) Conjunction, best features; It. 0      |
| 3) Best performing single representation | 7) Conjunction, NULL=1; It. 0      | 11) Arithmetic mean*, best features; It. 0 |
| 4) Arithmetic mean*; It. 0               | 8) Disjunction; It. 0              | 12) Disjunction, best features; It. 0      |

The geometric mean is omitted because it creates the same total order as the conjunction (see Matching Funct. 16).  
\* indicates standard aggregations.

Figure 8.8: Performance comparison of representation combinations using different NULL value policies, part 1

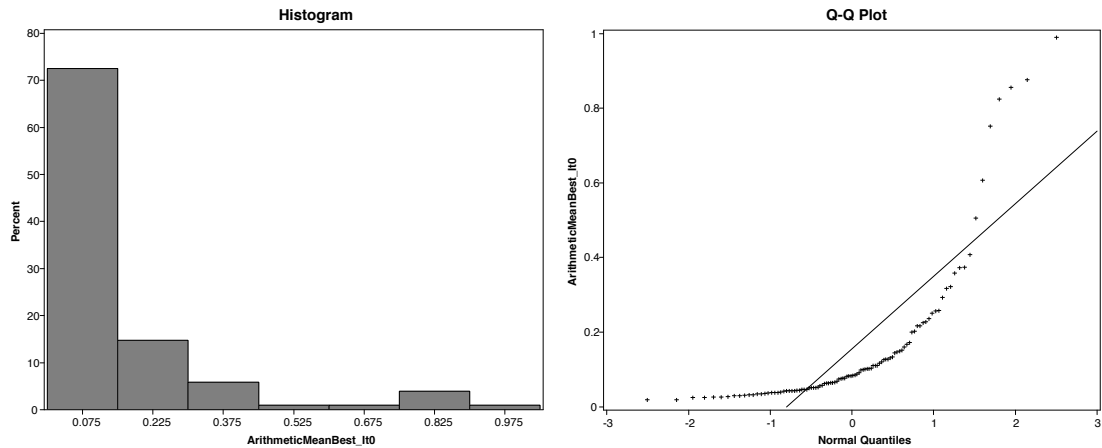
## 8.4 Retrieval Effectiveness of the CQQL Approach



The geometric mean is omitted because it creates the same total order as the conjunction (see Matching Funct. 16).  
 \* indicates standard aggregations.

Figure 8.9: Performance comparison of representation combinations using different NULL value policies, part 2

## 8 Evaluation of the Retrieval Effectiveness



Plotted by SAS 9.3 (PROC UNIVARIATE) [SAS Publishing 2011]

Figure 8.10: Sample distribution analysis of nDCG@20 values over all runs (arithmetic mean, best features; Caltech 101)

Table 8.14: Results of test for normality (arithmetic mean, best features; Caltech 101)

Test	Statistic	<i>p</i> -value
Shapiro-Wilk	0,642092	<0.0001
Kolmogorov-Smirnov	0,242963	<0.0100
Cramér-von Mises	2,193467	<0.0050
Anderson-Darling	12,03686	<0.0050

Computed by SAS 9.3 (PROC UNIVARIATE) [SAS Publishing 2011]

Not surprisingly, the quantile-quantile plot<sup>141</sup> against the normal distribution's quantiles (see Figure 8.10; right) and the corresponding histogram shows the same results.

Hence, the usage of the Student's t-test for significance testing in this thesis is not appropriate. As a consequence, all significance testing in the remainder of this text is carried out by the Wilcoxon signed rank test [Wilcoxon 1945], which does not require normally distributed samples.

### Establishment of a Baseline

Section 8.4.2 concluded that the color histogram and color structure are viable baseline representations against which representation fusions can be compared. The objective of this subsection is to find a matching function that fuses different representations to estimate the POR of a document in order to serve as a fusion baseline to complete the set of single representation baselines. Hence, this subsection only focusses on the detection of the most effective fusing matching function. The interpretation of the results

<sup>141</sup>The line depicts the expected quantiles for a normal distribution, while the dots indicate the actual distribution of the examined matching function.



of all fusing matching functions and their semantics are discussed in the following subsections.

Figure 8.11 shows a box plot<sup>142</sup> of the ranks obtained by various fusing matching functions at the aforementioned collections (see Section 8.4.1). The horizontal bar within the boxes indicates the median. In other words, 50 % of the data lies above this line. The region limited by the upper respectively the lower boundaries of the boxes and the whiskers show the upper and lower quartile, i.e., 25 % of the observed values are higher (and lower respectively) than the boundaries of the boxes. The horizontal lines that confine the whiskers denote the maximum (top) and minimum (bottom). Outliers are depicted as circles. In addition, the arithmetic mean is shown by a filled black circle.

The box plot includes typical aggregation functions, e.g., the arithmetic mean of the calculated similarities per representation (see Matching Function 13) or the minimum (see Matching Function 20). For the sake of clarity, the sum is omitted because it creates the same total order as the arithmetic mean (see Matching Function 13) using the pessimistic NULL value mapping strategy (see above). Appendix B.1.2 lists the arithmetic expressions of the different aggregation functions.

In order to facilitate the comparison with the single-representation matching functions, the ranks of the baselines (color histogram, color structure) and the respective best performing single representation are depicted as well.

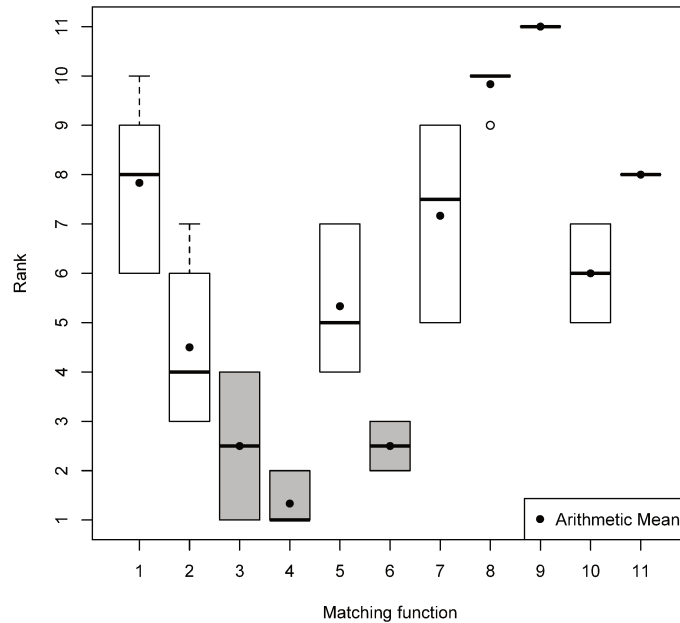
Each standard aggregation function is contained as a normal and a *best features* variant. The best features variant uses all available representations but the three worst performing ones (see Table 8.10), i.e., Gabor, contour-based, and region-based shape. This naming convention applies to all discussed matching functions in the remainder of this text. The removal of hardly effective representations is due to Larsen [2004], who recommends to remove representations that yield only ranks with little or no relevant documents [Larsen 2004, cf. pp. 36f.]. As a consequence, the best features variants contain only the most promising representations as advised by Larsen et al. [2006] (see Section 3.2.1). The best features variant outperform their normal counter-part in every tested collection.

For better visibility, the best three matching functions are shaded in Figure 8.11. The best features variant of the arithmetic mean (4) is clearly ranked first. It is followed by the respective best performing single representation (3) and the geometric mean (best features, 6) at the same place. Please note that the latter has a lower variance regarding the obtained rank. That is, it can be considered as providing a more stable retrieval effectiveness than number 3.

Figures 8.12 and 8.13 illustrate the retrieval effectiveness of the examined matching functions as before in form of stacked bar charts. Obviously, the nDCG values obtained by 4 and 6 are very close to each other in every collection. Thus, it is necessary to examine whether the effectiveness difference is significant between both matching functions with the help of the Wilcoxon signed-rank test. Table 8.15 clearly shows that the geometric mean's (best features) nDCG@20 values differ significantly from 4 for all collections but Wang. The effect with the Wang collection is not very surprising as Section

<sup>142</sup>The box plot uses the standard settings of the R [R Core Team 2012] based on Tukey [1977].

## 8 Evaluation of the Retrieval Effectiveness



- |   |  |                                |
|---|--|--------------------------------|
| 1) Color Histogram; lt. 0                 | 5) Arithmetic mean*; lt. 0               | 9) Max*; lt. 0                 |
| 2) Color Structure; lt. 0                 | 6) Geometric mean*, best features; lt. 0 | 10) Min*, best features; lt. 0 |
| 3) Best performing single representation  | 7) Geometric mean*; lt. 0                | 11) Min*; lt. 0                |
| 4) Arithmetic mean*, best features; lt. 0 | 8) Max*, best features; lt. 0            |                                |

Horizontal bands indicate the median, solid circles the mean. Empty circles indicate outliers.  
 The three best performing matching functions are shaded.  
 \* indicates standard aggregations.

Figure 8.11: Box plot of ranks obtained by single representations and combining matching functions over all collections

8.4.2 already discussed the high values for the nDCG metrics in this easily retrievable collection.

In the case of the respective best performing single representation (3) the situation is not that clear. The performance of 3 differs significantly from 4 only for Caltech 101, MSRA-MM, and UCID. With the other collections, the effectiveness of 3 is not better or worse than that of 4.

To conclude, because of its general good performance, the arithmetic average (best features variant) can serve as a baseline for matching functions relying on multiple representations at once. That is, a matching function that relies on the PoP is expected to outperform this baseline. However, the other standard aggregation functions are continuously listed with the more complex fusing matching functions to provide points of reference for the assessment of their utility.

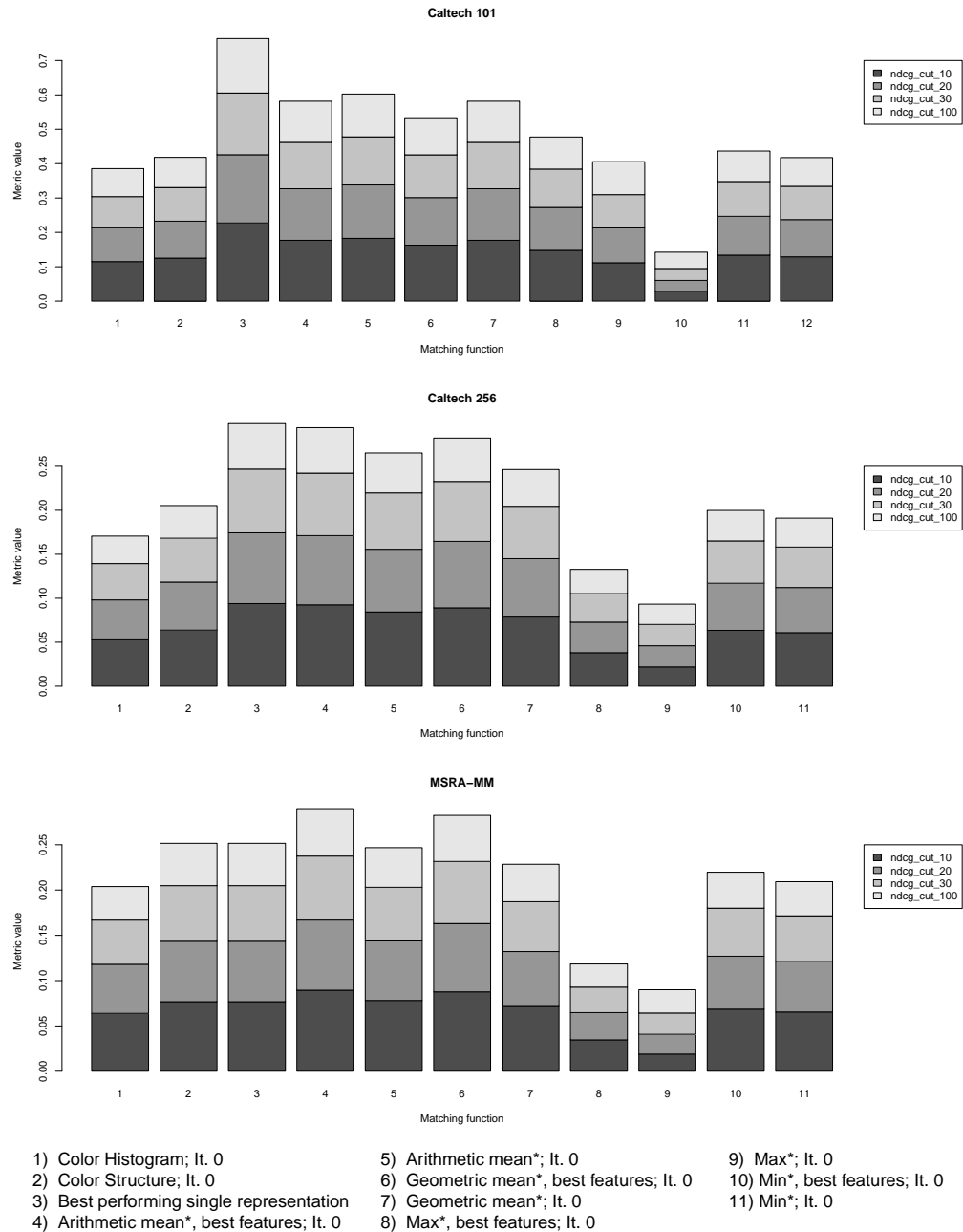
## 8.4 Retrieval Effectiveness of the CQQL Approach

Table 8.15: Results of the Wilcoxon signed-rank test for baseline matching functions for nDCG@20; insignificant differences are shaded

<i>p</i> -value of “Arithm. mean, best features” (#4) vs.	Caltech 101	Caltech 256	MSRA-MM	Pythia	UCID	Wang
Best performing single representation (#3)	< 0.0001	0.8420	< 0.0001	0.2017	0.2564	0.0371
Geometric mean, best features (#6)	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.0039

Computed by SAS 9.3 (PROC UNIVARIATE) [SAS Publishing 2011]

## 8 Evaluation of the Retrieval Effectiveness



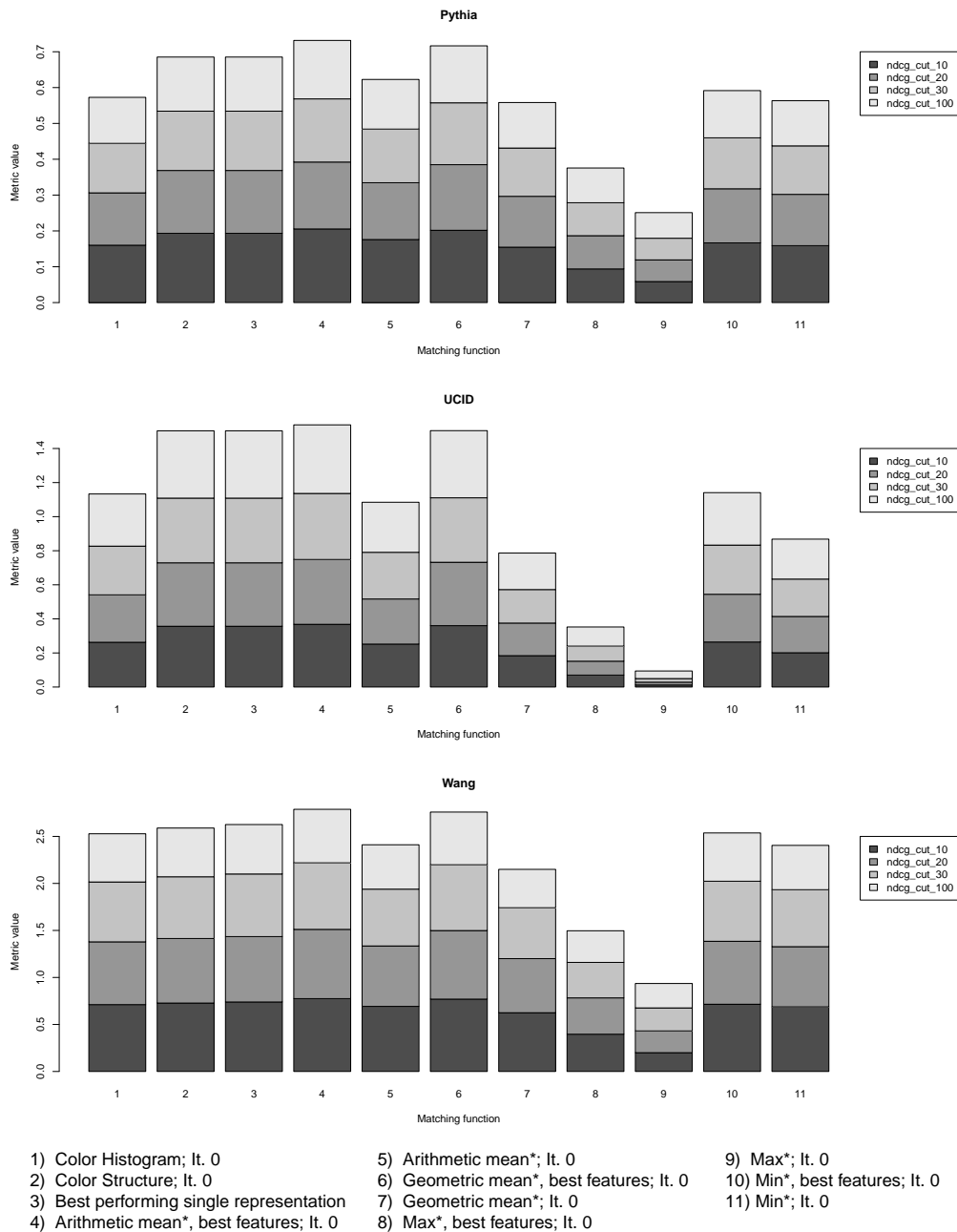
The sum is omitted because it creates the same total order as the arithmetic mean (see Matching Function 13).

\* indicates standard aggregations.

Caltech 101/256, #3: Edge Histogram; MSRA-MM, #3: Color Structure

Figure 8.12: Performance comparison of basic combinations of representations, part 1

## 8.4 Retrieval Effectiveness of the CQQL Approach



The sum is omitted because it creates the same total order as the arithmetic mean (see Matching Function 13).

\* indicates standard aggregations.

Pythia, #3: Color Structure; UCID v2, #3: Color Structure; Wang, #3: BIC

Figure 8.13: Performance comparison of basic combinations of representations, part 2

### Experimental Results – Main Experiment I

Figures 8.14 and 8.15 visualize the resulting nDCG metrics of the QBE experiments, which rely on multiple representations at once in form of stacked bar charts separated by the tested collection. The plots' interpretation follows the same pattern as discussed in Section 8.4.2. A detailed description of the examined matching functions is available in Appendix B.1, while the full results in numerical form can be found on the attached storage medium (see Appendix E). Because most of the presented matching functions include weighting variables, it is important to point out that all weight values are set to 1, i.e., each representation is evaluated as in the unweighted case (see Section 4.3).

To facilitate the comparison with the baseline single representations, the results for the color histogram, the color structure, and the best performing representation for each collection (see Table 8.9) are displayed as the first three bars (from left to right). The fourth and fifth bar show the effectiveness of the arithmetic mean variants as a baseline for the fusion of multiple representations.

As stated in Section 8.4.1, the other investigated matching functions can be subdivided into three groups: standard aggregations, polyrepresentation-motivated CQQL-based, and other literature-motivated matching functions.

**Standard aggregation matching functions** The standard aggregation matching functions consist of numbers 4, 5, 14, 15, 16, and 17<sup>143</sup>.

The maximum and the minimum (14-17) take on a special position because they also represent the disjunction (max) and conjunction (min), respectively, in the original publication of fuzzy logic [Zadeh 1988].

The geometric mean variants are no longer included because they produce the same rank as the conjunction variants (see Matching Function 16).

**Polyrepresentation-motivated CQQL-based matching functions** The central hypothesis of the PoP of the information space (or documents) is that a document is defined by different representations that can be combined to form a *cognitive overlap* (CO), in which highly relevant documents are most likely to be contained (see Section 3.2). From a logical point of view, an overlap equals the conjunction of different representations. Hence, a weighted CQQL conjunction (6/7) is juxtaposed with its logical counter-part, a weighted CQQL disjunction (8/9) of all available representations.

**Other matching functions motivated by the literature** To complete the picture, the aforementioned matching functions are compared against functions that are motivated by other publications. For instance, Matching functions 10-13 are motivated by Eidenberger [2003], who suggests to combine only color layout, dominant color, and a texture representation. This choice is based on results of a retrieval effectiveness study relying only on MPEG-7 features that revealed that the other available representations only contribute redundant information and can therefore be ignored [Eidenberger 2003]. As

<sup>143</sup>The given numbers refer to the indices used in Figures 8.14 and 8.15.

## 8.4 Retrieval Effectiveness of the CQQL Approach

Eidenberger [2003] does not give information about the combination of these representations, two conjunctive and disjunctive variants are examined.

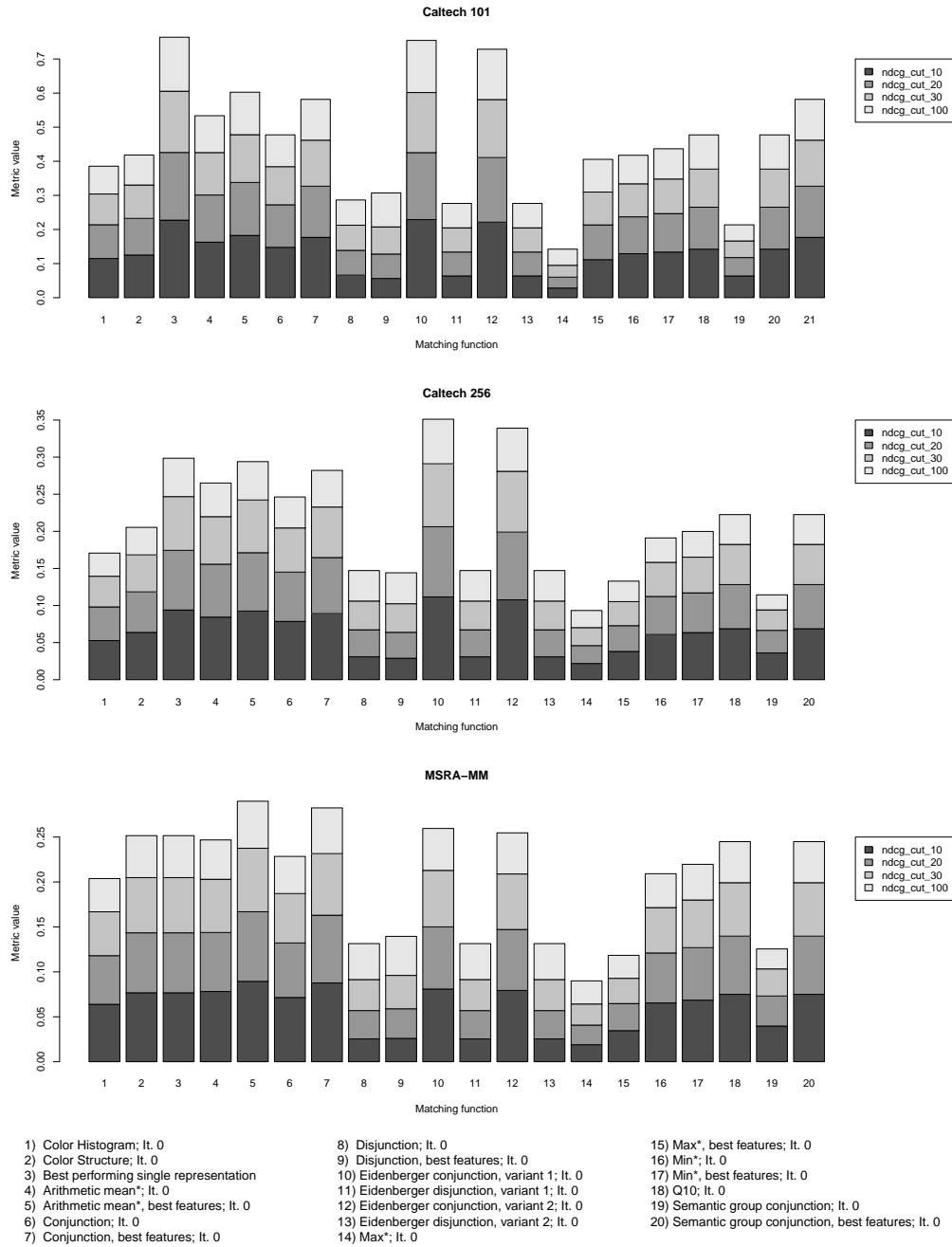
Matching functions 18-20 group representations into so-called “semantic groups”, which either fuse combined texture and color properties, solely color-related ones, or such that model edge and texture properties. Following the terminology of the PoP, functionally similar representations are grouped. The representations in each group are connected with a disjunction, modeling the assumption that they correlate, which is stated in other studies [Eidenberger 2003; Deselaers et al. 2008], and thus model the same aspects of an image. Hence, it would be enough, if one of the examined representations shares a high degree of similarity with the query in order to assess a document as highly relevant. These groups are then embedded into a conjunction forming a CO in order to reward documents that are good in all of the representation groups, while assuming redundancy of the representation sub-groups.

The 18th matching function Q10 is suggested by Zellhöfer & Schmitt [2011a], who report of a pre-study that revealed the generally good retrieval effectiveness of the CEDD [Chatzichristofis & Boutalis 2008a], FCTH [Chatzichristofis & Boutalis 2008b], color layout [Cieplinski et al. 2001], and Tamura [Tamura et al. 1978] representations, using the implementations provided by LIRE<sup>144</sup> [Lux & Chatzichristofis 2008]. Moreover, Q10 reflects the finding from Deselaers et al. [2008] that a combination of texture and edge detectors can improve the retrieval quality [Zellhöfer & Schmitt 2011a].

---

<sup>144</sup>Lucene Image REtrieval

## 8 Evaluation of the Retrieval Effectiveness



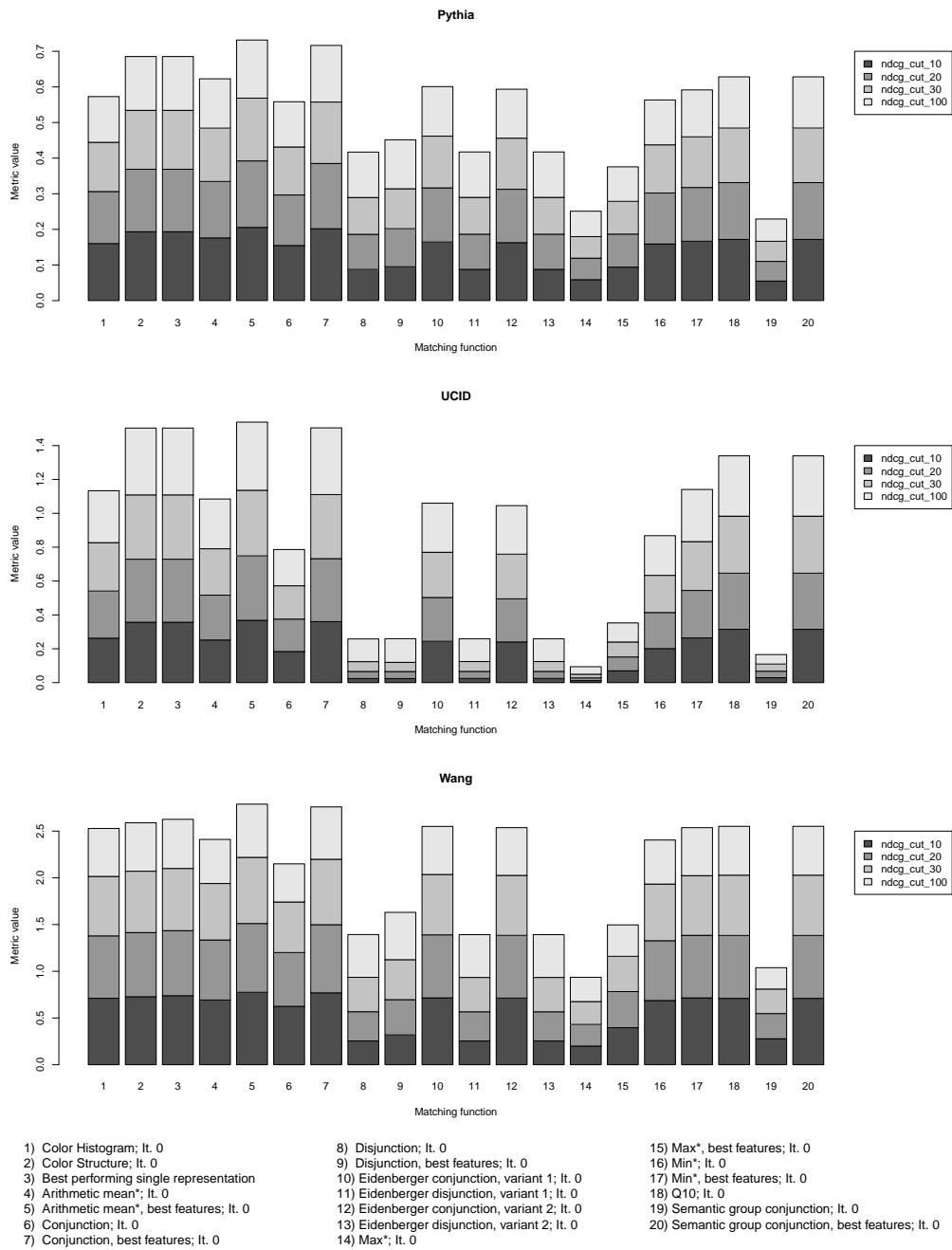
\* indicates standard aggregations.

Caltech 101/256, #3: Edge Histogram; MSRA-MM, #3: Color Structure

Figure 8.14: Performance comparison of representation combinations and standard aggregations, part 1



## 8.4 Retrieval Effectiveness of the CQQL Approach



\* indicates standard aggregations.

Pythia, #3: Color Structure; UCID v2, #3: Color Structure; Wang, #3: BIC

Figure 8.15: Performance comparison of representation combinations and standard aggregations, part 2

### Discussion – Main Experiment I

Table 8.16 shows all matching functions, including their obtained mean effectiveness rank<sup>145</sup>, in comparison to the other matching functions and its standard deviation when observed over all collections. Single representations are not considered because they have been discussed before. Additionally, the table points out whether conjunctive or disjunctive characteristics dominate in a given logic-based matching function.

Table 8.16: Matching functions and their characteristics, ordered by mean rank

ID	Matching Function	Mean Rank	Std. Deviation	Conjunctive	Disjunctive
2	<b>Arithmetic mean, best features</b>	<b>1.67</b>	1.03		
4	<b>Conjunction, best features</b>	<b>2.67</b>	1.03	✓	
7	<b>Eidenberger conjunction, variant 1</b>	<b>3.83</b>	2.56	✓	
9	<b>Eidenberger conjunction, variant 2</b>	<b>4.83</b>	2.56	✓	
17	Semantic group conjunction, best features	5.17	2.04	✓	
15	Q10	5.50	2.17		
1	Arithmetic mean	5.67	1.21		
14	Min, best features	7.83	1.60	✓	
3	Conjunction	8.33	1.97	✓	
13	Min	9.50	0.55	✓	
6	Disjunction, best features	11.83	1.17		✓
5	Disjunction	13.00	1.41		✓
8	Eidenberger disjunction, variant 1	13.00	0.89		✓
12	Max, best features	13.33	2.25		✓
10	Eidenberger disjunction, variant 2	14.00	0.89		✓
16	Semantic group conjunction	16.00	0.63	✓	
11	Max	16.83	0.41		✓

Unstable matching functions with a greater standard deviation than the mean of 1.43 are shaded.

Mean ranks  $\leq 5$  are bold.

The box plot of the same data in Figure 8.16 clearly shows a differing variance of the ranks obtained by the examined matching functions. While some matching functions are relatively *stable* regarding their mean rank (e.g., #6), others show a higher level of variance (e.g., #7).

**Definition 8.9 Effectiveness stability:** *A matching function is effectiveness-stable as long as it has a smaller standard deviation of the obtained mean rank than the average standard deviation of all matching functions it is compared to.* ◇

Matching functions that are not effectiveness-stable are shaded in Table 8.16 and Figure 8.16. From the examined matching functions, ca. 58% are stable. Effectiveness stability can be interpreted as a performance figure indicating whether the rough performance of a matching function can be predicted. Although further experiments are necessary to draw generally valid statistical conclusions, effectiveness stability is used as a quality feature of a matching function from now on.

The examination of the total order of the matching functions can only give insights into their comparative performance. In order to assess the effectiveness, it is also necessary to consider the results of the nDCG measure.

<sup>145</sup>That is, a matching function with a low rank on average performs better than one with a high rank.

## 8.4 Retrieval Effectiveness of the CQQL Approach

First of all, it is important to examine whether the effectiveness difference between the two first placed matching functions, the arithmetic mean and the conjunction in their best features variants, is significant. Figures 8.14 and 8.15 illustrate that these matching functions perform relatively equal for most collections. Nevertheless, as Table 8.17 shows, their effectiveness differs significantly for all collections but Wang. This is not surprising as the conjunction produces the same rank as the geometric mean discussed before with the same results during the establishment of the baseline.

Table 8.17: Results of the Wilcoxon signed-rank test between arithmetic mean and conjunction (best features variants) for nDCG@20

	<b>Caltech 101</b>	<b>Caltech 256</b>	<b>MSRA-MM</b>	<b>Pythia</b>	<b>UCID</b>	<b>Wang</b>
<i>p</i> -value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.0039

Computed by SAS 9.3 (PROC UNIVARIATE) [SAS Publishing 2011]

Generally speaking, the best features variants outperform their normal counterpart. The only exception is the disjunction in the Caltech 256 collection (see Figure 8.14; 8/9), although the difference is relatively small. The relative effectiveness difference between a best features and a normal variant differs from collection to collection with the semantic group conjunction variants showing the most significant difference.

Matching function Q10 (see Figures 8.14 and 8.15; 18) is particularly interesting because of its high effectiveness and its “economic” usage of representations. Q10 (see Matching Function 9) relies only on five representation, whereas the two first placed matching functions rely on twelve representations. Furthermore, Q10 combines conjunctions and disjunction, rendering it a representative of structurally more complex matching functions. Unfortunately, Q10 is not effectiveness-stable. Whether this is due to the structure of the matching function or the choice of representations remains a question for further research.

The Eidenberger variants use a comparably low number of representations (i.e., three to four, see Matching Functions 5-8). The effectiveness difference between the Eidenberger variants 1 and 2 is only marginal. In general, the Eidenberger conjunctions perform worse than the arithmetic mean and the conjunction, and are placed third and fourth (see Table 8.16). Their major drawback is that they are unstable as Q10. This is due to their effectiveness with Caltech 101 and 256, where the Eidenberger conjunctions clearly outperform the arithmetic mean and the conjunction. We assume that these effects are mainly due to the color artifacts in these collections (see Section 8.4.2).

In addition, the Eidenberger variants illustrate a general phenomenon present in all matching functions. The conjunctive variant of a matching function always performs better than its disjunctive counterpart. For instance, consider the minimum (see Figures 8.14 and 8.15; 16/17) often used as the conjunction in fuzzy logic [Zadeh 1988] and the maximum (14/15) used as the disjunction. The figures clearly show that the best features minimum outperforms the best features maximum. The same holds true for the normal variants of the minimum and maximum.

To conclude, the conjunctive variants are in general more effective than disjunctive

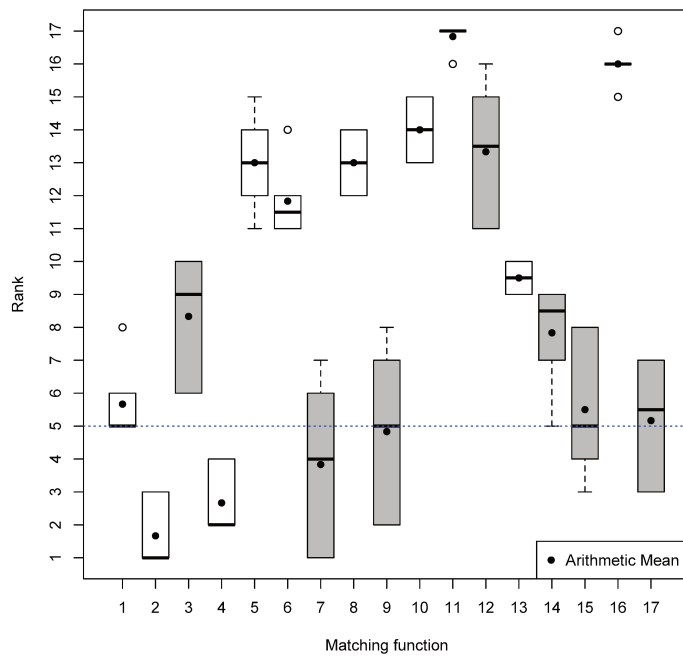
## 8 Evaluation of the Retrieval Effectiveness

matching functions (see Figures 8.14 and 8.15; 6/7 vs. 8/9; 10/12 vs. 11/13; 16/17 vs. 14/15) but not necessary effectiveness-stable as Table 8.16 shows.

Unfortunately, the fusion of multiple representation is no guarantee to improve the retrieval effectiveness – in particular if poorly performing single representations are part of the matching function. Figures 8.14 and 8.15 clearly show that there are matching functions, which rely on various representations performing worse than the color structure (2) or the collection-dependent best performing single representation (3). This effect comes in particular visible with the minimum and maximum. Although the minimum and maximum are in principle fusing matching functions, they suffer from the *dominance problem* (see Section 4.5.3) that makes them behave more like a single representation.

To sum up, although the best features variants of the arithmetic mean and the conjunction are not necessarily outperforming the best performing single representation in every collection, it is reasonable to assume that the utilization of multiple representations will become handy during relevance feedback as it allows more degrees of freedom for the learning algorithm used by PrefCQQL. Whether this assumption holds is investigated in Section 8.5. Furthermore, these matching functions are effectiveness-stable, which makes them a viable baseline for the following experiments.

## 8.4 Retrieval Effectiveness of the CQQL Approach



Horizontal bands indicate the median, solid circles the mean. Empty circles indicate outliers.  
Unstable matching functions are shaded.

Figure 8.16: Box plot of the ranks obtained by different matching functions over all collections; the matching function IDs refer to Table 8.16

## 8.5 Retrieval Effectiveness of PrefCQQL – The Impact of Preference Elicitation and Adaption on Ranks

Table 8.16 shows the stable retrieval effectiveness of the best features variants of the arithmetic mean and conjunction matching functions in the unweighted case.

The objective of this section is to examine the effectiveness of these and other matching functions during relevance feedback using the PrefCQQL approach (see Section 5.4). To achieve this, the following experiments extend the method presented in Section 8.4.1 by including relevance feedback, which is input into the weight learning algorithm of PrefCQQL (see Section 5.4.4). In other words, this section investigates the effectiveness of the principle of polyrepresentation’s hypothesis in interactive MIR.

### 8.5.1 Experimental Setup for User Simulation and Relevance Feedback

Obviously, manual relevance feedback is no longer possible for the vast number of QBE submissions used in the experiments described before (see Table 8.8). Thus, all RF-based experiments in this thesis rely on user simulations. In order to recapitulate various arguments for user simulations presented in Sections 7.2 and 8.3.4, they are generally regarded as a cost-effective experimental method that does not suffer from user learning effects and at the same time guarantees reproducibility [Baskaya et al. 2011, 2012]. Another advantage of user simulations in the context of this research is the fact that they are not affected by potential usability issues of the GUI, which is evaluated separately in Chapter 9.

As illustrated in Figure 8.4, user simulations are meant to represent the user’s feedback in interactive IR experiments. To provide RF to the system, a “lazy” simulated user (simuser) that interacts with the system following a simple strategy is designed.

#### Interaction strategy of the “lazy” simulated user

1. *The simuser submits a query consisting of one relevant QBE document to the system, which uses one of the pre-defined matching functions over the full interaction time.*
2. *After the presentation of the results, the simuser inspects the 20 top-most documents. For each of the 20 documents:*
  - a) *Relying on the the ground truth provided with the collection, the simuser checks if an irrelevant document  $\mathbf{d}_{\text{irrelevant}}$  directly precedes a relevant document  $\mathbf{d}_{\text{relevant}}$ , i.e.,  $\mathbf{d}_{\text{relevant}} < \mathbf{d}_{\text{irrelevant}}$ .*
  - b) *If so, the simuser inverts the preference into  $\mathbf{d}_{\text{relevant}} > \mathbf{d}_{\text{irrelevant}}$ .*
3. *The simuser resubmits the new preferences to the system.*
4. *In case of a query incompatibility, the simuser cancels the interaction, and the last valid weights are used. Otherwise, the simuser continues with 2. until the maximum number of RF iterations is reached.*

Without doubt, there are more sophisticated interaction strategies. Nevertheless, the motivation behind this simuser is to model an impatient and potentially lazy user in order to assess the effectiveness of PrefCQQL in a sub-optimal<sup>146</sup> setting. The value of 20 documents has been chosen accordingly because 20 reasonably-sized image thumbnails can be displayed simultaneously at common screen resolutions and remain recognizable without additional scrolling at the same time. For an example, see Figure 9.5. Section 8.5.5 discusses how many preferences are set in a typical usage scenario.

Although the PrefCQQL RF approach (see Section 5.4) allows finer grades of relevance input, other strategies have not been tested. This is mainly due to the binary ground truth of the collections Caltech 101/256, UCID, and Wang, which serve as the basis of the simuser’s relevance decisions. Finer graded relevance feedback would only be possible for the MSRA-MM and Pythia collections.

Furthermore, the feedback given by the simuser is not necessarily realistic because the simuser is mostly providing perfect input, i.e., little or no erroneous input is given, although such input would be observed in a real user study. This drawback could only be compensated to a certain extent by using the ground truth of the Pythia collection (see Section 8.3.4), which also contains potential erroneous user input regarding the relevance assessments.

In principle, the experiments could be extended to feature different personas as provided with the Pythia collection (see Section 8.3.4). For the sake of comparability, the experiments only use the Pythia average persona, i.e., a ground truth that represents the average relevance judgements of all Pythia assessors, because all other collections do not feature multiple ground truths.

For a discussion on the impact of different personas on the retrieval effectiveness, see a reports on the results from the ImageCLEF campaign 2013 [Zellhöfer 2013].

Preference conflicts (see Section 5.4.7) cannot occur with this interaction strategy, because it is not possible to state cyclic preference graphs. Query incompatibilities remain possible and are handled as described in Section 6.2.5.

To conclude, all other side conditions stated in Section 8.4.3 apply for the user simulation-driven experiments.

### A Short Note on the Reproducibility of the Results

In principle, the Nelder-Mead algorithm (see Section Section 5.4.5) limits the reproducibility of the experiments described in this thesis. This is due to the reliance of the algorithm on random numbers in order to solve the given optimization problem. As a consequence, each run of the algorithm might find different solutions resulting in different weights and thus result ranks.

All experiments are reproducible using the code provided with this dissertation (see Appendix E) because fixed random seeds were used in conjunction to the random number generator *grand* provided by Qt 4.7. The random number generator produces a

<sup>146</sup>That is, a potentially sub-optimal setting for the learning algorithm, which is nevertheless realistic because users typically avoid extensive input [Shneiderman & Plaisant 2005].

## 8 Evaluation of the Retrieval Effectiveness

roughly uniform distribution of random values as shown in Figure B.96 in Appendix B.5.1.

The main termination parameters for the learning algorithm are  $kNMax = 10,000$ , a fault tolerance of 0.01, and 100 simplex starts (see Section 5.4.5). The simplex runs are parallelized. The choice of the parameters was determined in pre-studies on the basis of their generally good performance. As described in Section 5.4.5 and Listing 5.2 in particular, the learning algorithm uses the SumMin strategy. For further details on the implementation used during the experiments, refer to the source code<sup>147</sup> in the supplement.

With regard to the user simulation, all user interactions and relevance decisions are reproducible if the code accompanying this dissertation is used (see Appendix E).

### 8.5.2 Typical Relevance Feedback Performance – Main Experiment II

To start with, this section presents the retrieval effectiveness development during PrefCQQL-based relevance feedback iterations of various matching functions. All experiments are based on the aforementioned “lazy” simulated user, who interacts only with the top-20 documents and aborts the interaction with the system after a maximum of five RF iterations.

The subsequent sections discuss special cases of relevance feedback. Section 8.5.3 presents the results of a simuser inspecting only the top-20 documents, but who is willing to provide relevance feedback for up to 15 iterations using two exemplary matching functions. Section 8.5.4 discusses the internal weight development of the weights learnt by PrefCQQL’s weight learning algorithm in order to explore correlations between the change of weights and other figures such as the retrieval effectiveness.

### Experimental Results – Main Experiment II

The plots shown in this section follow the same pattern as used before; however, they have been extended to feature multiple relevance feedback iterations. The experiment uses the same matching functions as Section 8.4.3. Hence, the same categorization applies. The minimum and maximum matching function variants are omitted in this experiment because they do not support weights as required by PrefCQQL.

The RF iterations are denoted as follows. For instance, “It. 0” indicates the 0th RF iteration (no relevance feedback has been given), i.e., the incrementing index after “It.” shows the RF iteration. To give another example, “It. 2” means that the simuser has initially retrieved a rank of documents and has provided relevance feedback twice. In other words, the simuser has both modified the preference graph and submitted it to the learning algorithm twice.

Initially (at RF iteration 0), all  $\theta_i = 1$  (the unweighted case, see Section 4.3). Later, the learning algorithm alters the values of the weighting variables  $\theta_i$ . As said before, the impact of the learning algorithm on the weight values is discussed separately in Section 8.5.4.

<sup>147</sup>See namespace `dbis::weightlearning`, and `dbis::weightlearning::CQQLWeightLearning` in particular.



Because of the high number of RF iterations and matching functions, this section only describes some characteristic results. Further results are available in Appendix B.2 and, also in numerical form, in the supplement to this thesis (see Appendix E).

To give a rough idea, Figures 8.17 and 8.18 illustrate the retrieval effectiveness development during the first two RF iterations for some typical matching functions. The weighted arithmetic mean<sup>148</sup> as the weighted counter-part of the arithmetic mean, which has been shown to be a feasible fusion function (see Table 8.16), serves as a baseline to assess the effectiveness of the other matching functions.

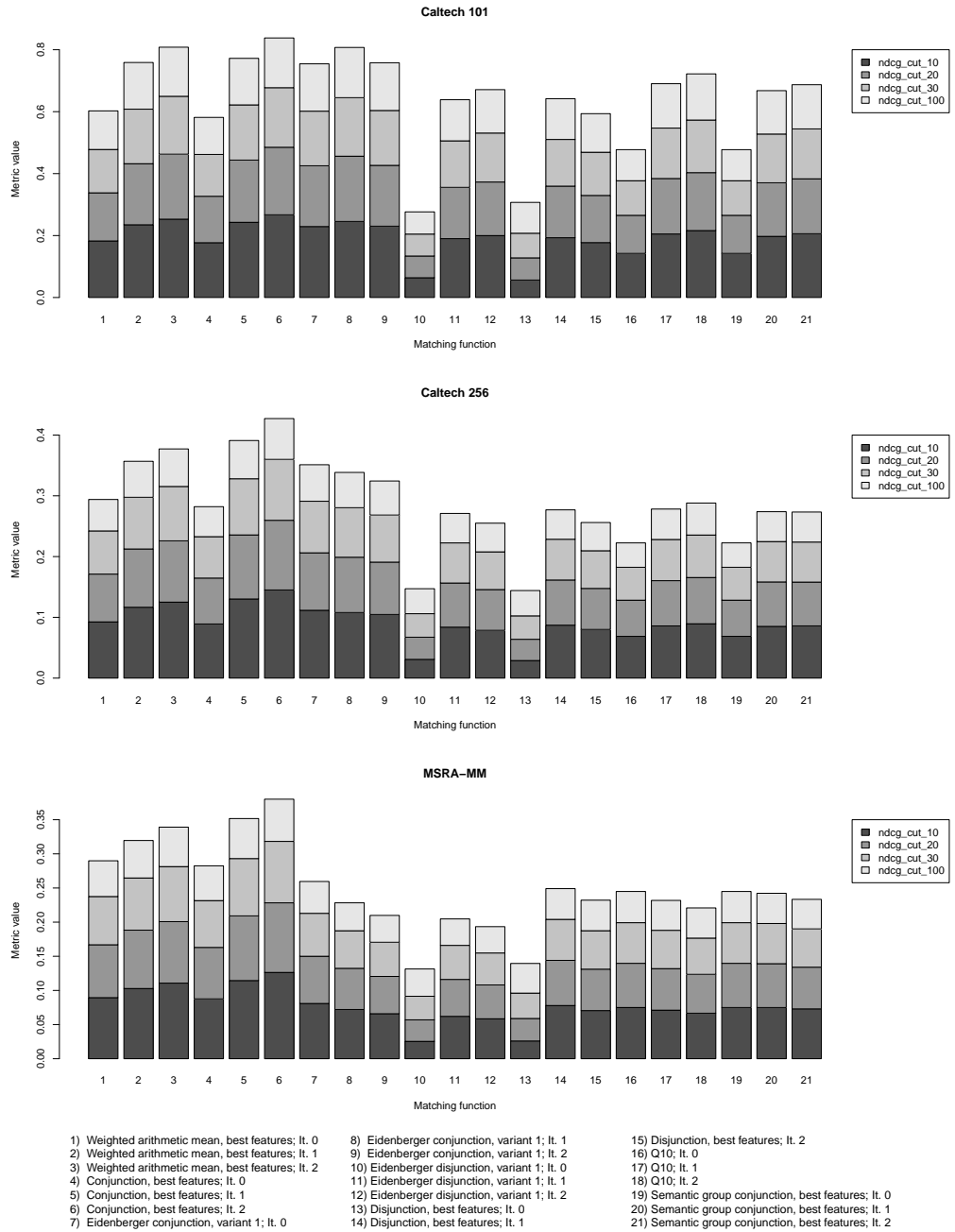
To illustrate one possible outcome of five RF iterations, Figures 8.19 and 8.20 compare the conjunction (best features variant) with the Eidenberger conjunction (variant 1). For the sake of comparability, the figures also include the baselines.

Appendix B.2 contains stacked bar charts for the other matching functions of which some will be addressed in the subsequent discussion of the experimental results.

---

<sup>148</sup>Please note that RF iteration 0 of the weighted arithmetic mean corresponds to the unweighted arithmetic mean because all  $\theta_i$  are set equal.

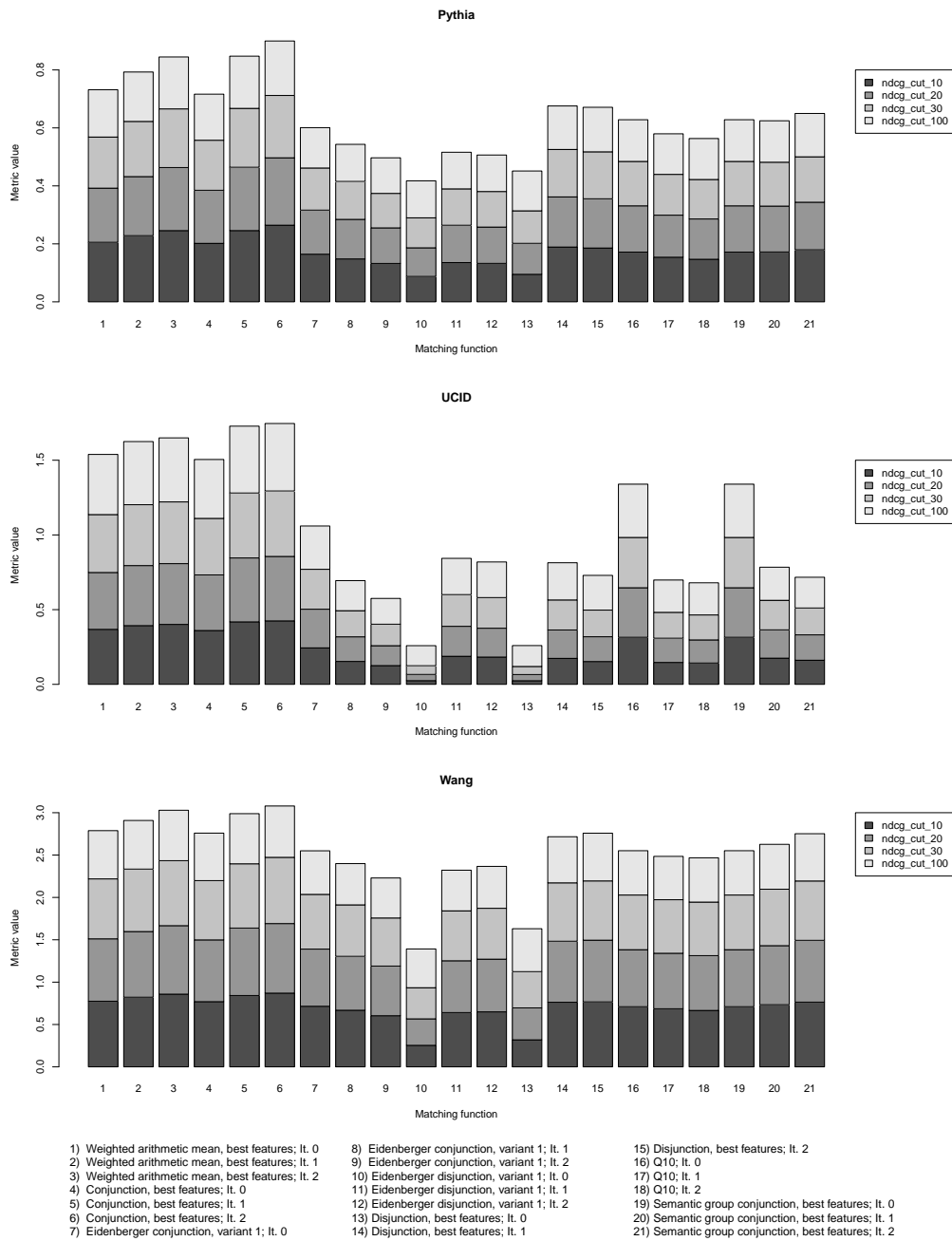
## 8 Evaluation of the Retrieval Effectiveness



\* indicates standard aggregations.

Figure 8.17: RF performance comparison of characteristic matching functions, part 1

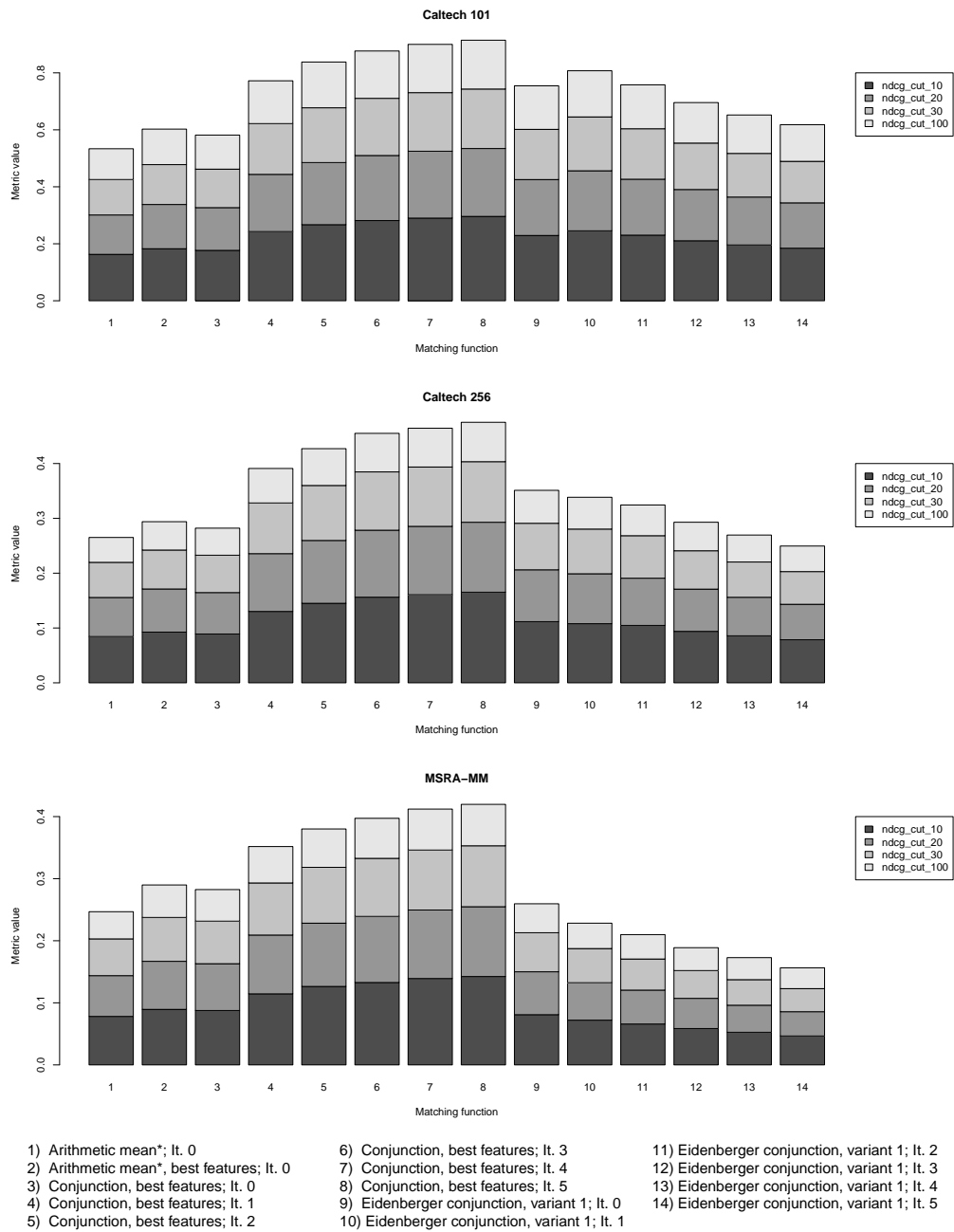
## 8.5 Retrieval Effectiveness of PrefCQQL



\* indicates standard aggregations.

Figure 8.18: RF performance comparison of characteristic matching functions, part 2

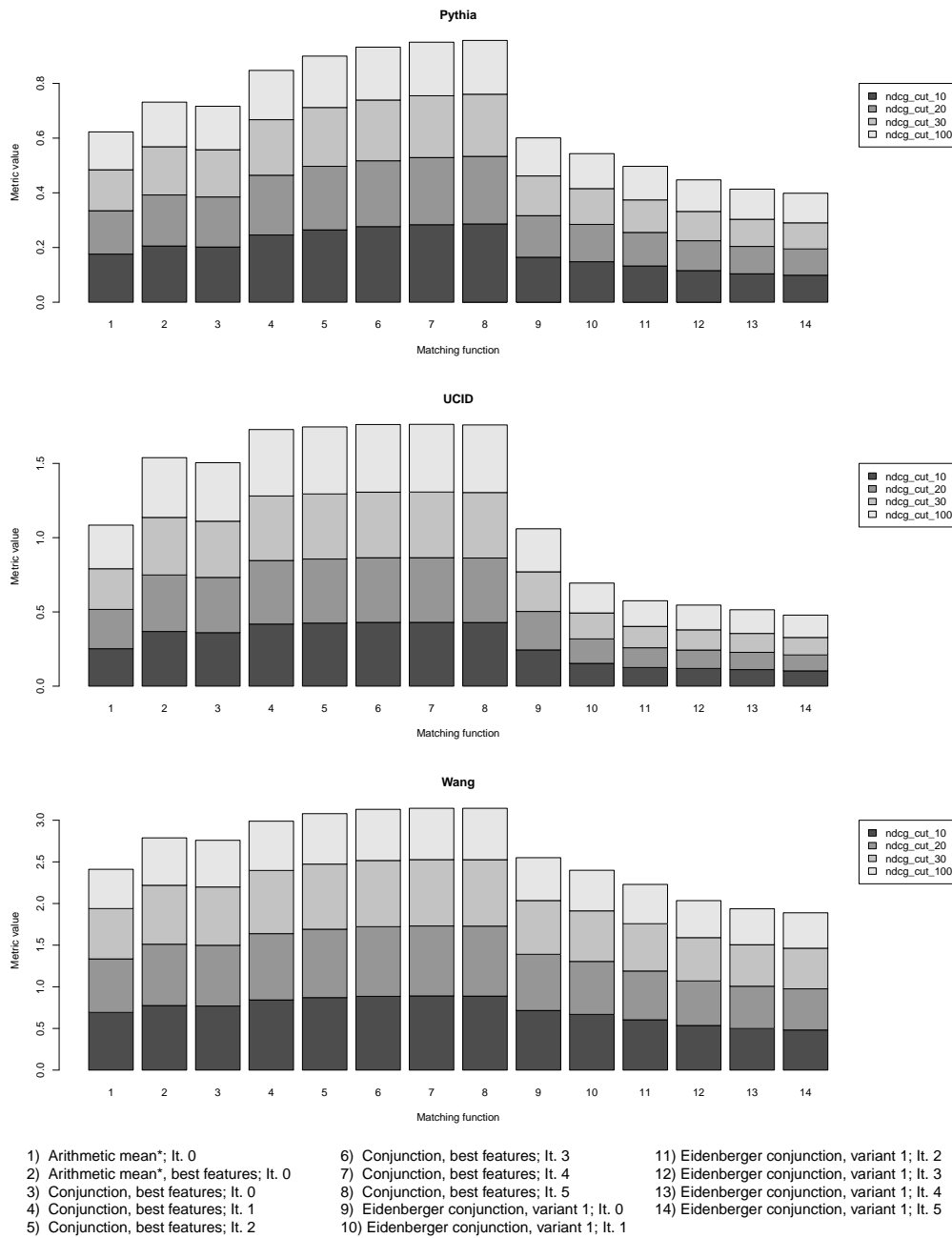
## 8 Evaluation of the Retrieval Effectiveness



\* indicates standard aggregations.

Figure 8.19: Performance comparison of RF-enabled representation combinations and standard aggregations, part 1

## 8.5 Retrieval Effectiveness of PrefCQQL



\* indicates standard aggregations.

Figure 8.20: Performance comparison of RF-enabled representation combinations and standard aggregations, part 2

### Discussion – Main Experiment II

The graphs clearly show that the retrieval effectiveness of the baseline matching function, the weighted arithmetic mean, gradually increases during the relevance feedback iterations (see Figures 8.17 and 8.18; 1-3). The second best matching function from Section 8.4.3, the conjunction (best features variant), behaves similarly, although it reaches a higher effectiveness than the weighted arithmetic mean (see Figures 8.17 and 8.18; 4-6) from RF iteration 1 onward.

Albeit the Eidenberger conjunction (7-9) also features conjunctive characteristics, its effectiveness does not increase during the RF iterations as the effectiveness of the conjunction. Instead, its performance falls during the RF iterations for all examined collections but Caltech 101. Here, the effectiveness climbs at the first RF iteration and decreases immediately afterwards.

The disjunctive matching functions, i.e., the Eidenberger disjunction (10-12) and the best features disjunction (13-15), follow a similar pattern. At the first RF iteration, the effectiveness significantly raises and falls directly thereafter.

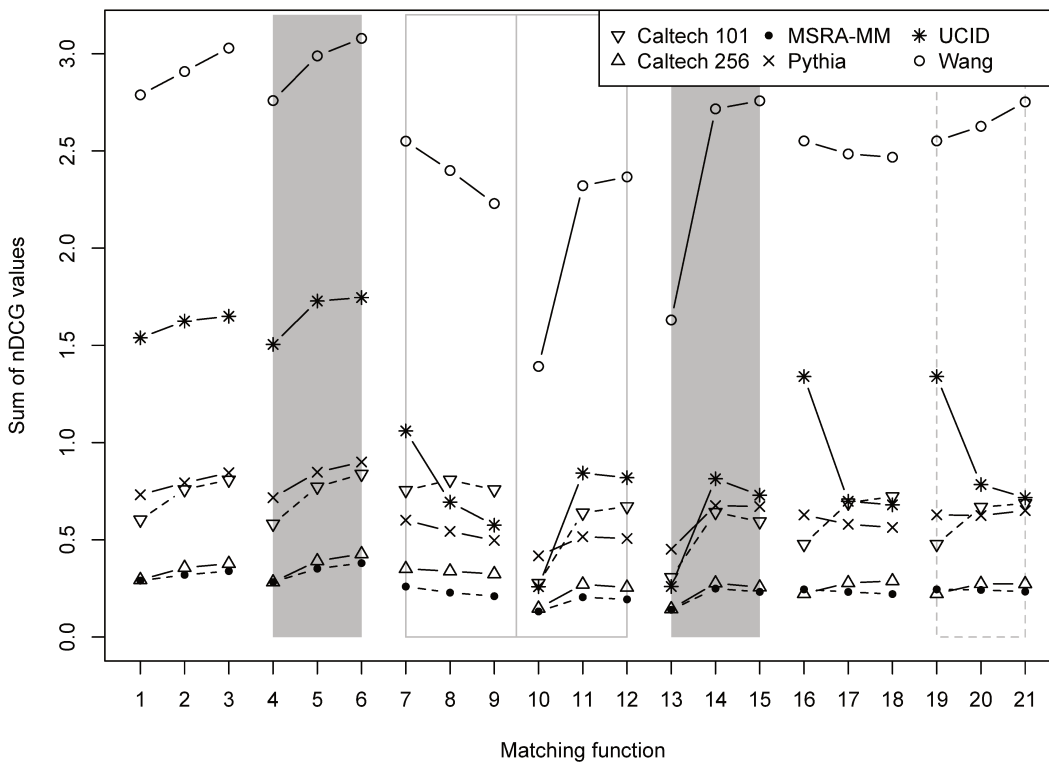
The two last matching functions, i.e., Q10 (16-18) and the semantic group conjunction (19-21), act rather unpredictably over all collections. Both matching functions feature conjunctions and disjunctions, while Q10 also relies on a small amount of representations (five in total). Although the use of Q10 might lead to a relatively small increase of the retrieval effectiveness during RF, e.g., for the Caltech collections, there is a trend of decreasing effectiveness. As the results from the UCID collection illustrate (see Figure 8.18; 16-18), this trend can become very significant. Roughly speaking, the semantic group conjunction shows an RF performance pattern, which is similar to Q10 and manifests in increases as well as decreases of the performance during RF.

These discussed trends are characteristic for all matching functions over all collections as Figure 8.21 summarizes. To facilitate the direct comparison of the characteristic performance development patterns of the different matching functions, each point indicating the effectiveness of one matching function at a given RF iteration is connected with a line.

Figure 8.21 in particular illustrates that the weighted arithmetic mean and the conjunction have the same performance development pattern over all collections. That is, the performance plot follows an ascending slope, which is steeper in the case of the conjunction (see Figure 8.21; 4-6). This means that the conjunction adjusts faster to the user's IN during RF in comparison to the weighted arithmetic mean, which surpasses the conjunction in terms of effectiveness when no RF is used. This fast reaction on RF was postulated in Section 5.4.4 and demanded in User Story 1; it can obviously be guaranteed with PrefCQQL.

Figures 8.19 and 8.20 show that the aforementioned trends are also present when five RF iterations with two exemplary matching functions, the conjunction (best features) and the Eidenberger conjunction (variant 1), are examined. For the sake of comparability, the graphs also contain the baselines.

Further plots compare the performance development during RF for up to five RF iterations of various matching functions against the best performing matching function



- |   |   |  |
|---|---|--|
| 1) Weighted arithmetic mean, best features; It. 0 | 8) Eidenberger conjunction, variant 1; It. 1  | 15) Disjunction, best features; It. 2                |
| 2) Weighted arithmetic mean, best features; It. 1 | 9) Eidenberger conjunction, variant 1; It. 2  | 16) Q10; It. 0                                       |
| 3) Weighted arithmetic mean, best features; It. 2 | 10) Eidenberger disjunction, variant 1; It. 0 | 17) Q10; It. 1                                       |
| 4) Conjunction, best features; It. 0              | 11) Eidenberger disjunction, variant 1; It. 1 | 18) Q10; It. 2                                       |
| 5) Conjunction, best features; It. 1              | 12) Eidenberger disjunction, variant 1; It. 2 | 19) Semantic group conjunction, best features; It. 0 |
| 6) Conjunction, best features; It. 2              | 13) Disjunction, best features; It. 0         | 20) Semantic group conjunction, best features; It. 1 |
| 7) Eidenberger conjunction, variant 1; It. 0      | 14) Disjunction, best features; It. 1         | 21) Semantic group conjunction, best features; It. 2 |

Related RF iterations are grouped visually for the sake of clarity.

Figure 8.21: Relevance feedback effectiveness development trends of characteristic matching functions

## 8 Evaluation of the Retrieval Effectiveness

for RF: the conjunction (best features) (see Appendix B.2). For instance, Figures B.2 and B.3 show that the decrease of retrieval effectiveness of the disjunction further continues after RF iteration 2. Moreover, the figures juxtapose the effectiveness of the conjunction/disjunction during RF and the standard aggregations introduced in Section 8.4.3. In a similar manner, Figures B.6 and B.7 compare the retrieval effectiveness decrease during RF of the Q10 matching function and the disjunction variants.

Figures B.8 and B.9 show similar results to the ones displayed in Figure 8.21 for the remaining Eidenberger matching functions. Comparable effects can be observed for the semantic group conjunction variants in Figures B.10 and B.11.

Table 8.18: Matching functions characteristics during relevance feedback

Matching Function	Number of Representations	Effectiveness stability RF It. 0	Conjunctive	Disjunctive	Consistent RF Perform.
Arithmetic mean	↗	✓			✓
Arithmetic mean, best features	↗	✓			✓
Q10	↘				
Conjunction	↗		✓		✓
Conjunction, best ft.	↗	✓	✓		✓
Eidenberger conjunction, var. 1	↘		✓		
Eidenberger conjunction, var. 2	↘		✓		✓
Semantic group conjunction	↗	✓	✓		
Semantic group conjunction, best ft.	↗		✓		
Disjunction	↗	✓		✓	
Disjunction, best ft.	↗	✓		✓	✓
Eidenberger disjunction, var. 1	↘	✓		✓	
Eidenberger disjunction, var. 2	↘	✓		✓	

↗: large number of representations; ↘: small number of representations

**Consistency of the Relevance Feedback Performance** Figure 8.21 and the discussion above showed that a matching function's performance during RF is not necessarily consistent over all collections. In other words, in the extreme case the same matching function might increase its effectiveness during RF with one collection and decrease it when used with another collection. To give an example, Q10 improves the result quality during RF when used with the Caltech 101 collection, but dramatically loses effectiveness when used with UCID (see Figure 8.21; 16-18). In any case, a consistent performance of a matching function during RF is desirable because it affects the system's conformity with user expectations during the interaction with a MIR system (see Section 9.3), required by User Story 2.

This part of the discussion of main experiment II examines some variables that might have an impact on the *consistency of the RF performance*: the number of representations



used in a matching function, the dominant logic characteristics of a matching function in accordance with Table 8.16, and the effectiveness stability (see Definition 8.9) of a matching function. Table 8.18 lists these characteristics for the investigated matching functions and states whether a matching function can be considered consistent in terms of its RF performance over all collections. Please note that RF performance consistency does not mean an actual increase of retrieval effectiveness during RF. Instead, it is a measure of the predictability of a matching function's effectiveness development during RF. In the best case, this development is positive, i.e., the effectiveness gradually increases. For instance, the conjunction (see Figure 8.21; 4-6) is an example for a generally positive development.

Table 8.18 reveals a correlation between the effectiveness stability and the RF performance consistency. However, this correlation must not be mistaken with causality as a closer examination relying on conditional probability theory (see Appendix A.2.2) shows:

$$P(\text{'consistent RF performance'} | \text{'stability'}) = 0.5$$

That is, the probability of obtaining a consistent RF performance, given that the function is effectiveness-stable, is 50% based on the observations made in main experiment II.

A higher correlation can be observed between a high number of representations ( $\nearrow$ ) and the consistency of the RF performance. Moreover, there is evidence for a slight causal relation between the number of representations in a matching function and its RF consistency:

$$P(\text{'consistent RF performance'} | \text{'\nearrow'}) = 0.63$$

As before, this only means that the performance of the matching function is predictable; it does not indicate whether it is also effective.

The hypothesis that the dominant logical characteristic (i.e., whether it is conjunctive or disjunctive) has an impact on the RF performance consistency suggests itself.

$$P(\text{'consistent RF performance'} | \text{'conjunctive'}) = 0.5 \quad (8.13a)$$

$$P(\text{'consistent RF performance'} | \text{'disjunctive'}) = 0.25 \quad (8.13b)$$

$$P(\text{'consistent RF performance'} | \text{'other'}) = 0.66 \quad (8.13c)$$

Unfortunately, this hypothesis cannot be justified for the logical characteristics of a matching function. Although there is evidence that the dominance of disjunctive characteristics in a matching function most likely will not lead to a consistent RF performance, the probability of achieving RF consistency with a conjunctive matching function is also only 50%. If no clear predominance is present in the matching function ("other"), there is evidence that such matching functions might cause a consistent RF performance. However, the data suggests that the logical characteristics of a matching function has an impact on its retrieval effectiveness (see below).

If one examines the conditional probabilities of obtaining a consistent RF performance considering a high number of representations and the logical characteristics of

## 8 Evaluation of the Retrieval Effectiveness

the matching function at once, the results become clearer.

$$P(\text{'consistent RF performance'} \mid \text{conjunction}) = 0.5 \quad (8.14a)$$

$$P(\text{'consistent RF performance'} \mid \text{disjunction}) = 0.5 \quad (8.14b)$$

$$P(\text{'consistent RF performance'} \mid \text{other}) = 1.00 \quad (8.14c)$$

These results suggest that no dominant logical characteristic will cause a consistent RF performance. Alas, the experiments are strongly biased towards these matching functions. There are only three matching functions of this class, i.e., the two weighted arithmetic mean variants and Q10, of which the weighted arithmetic mean variants both feature a high number of representations and have been shown to be effective over all collections. Hence, these results and the previous ones in Equation (8.13) are of limited explanatory power.

This restriction has to be made for all of the aforementioned assertions. Because of the limited size of the observations, the conclusions drawn from the data cannot be considered statistically significant in a strict sense. However, they give evidence for first trends that have to be validated in future research.

**Retrieval Effectiveness of the Best Performing Matching Functions** Figure 8.21 illustrates that the weighted arithmetic mean and the conjunction (both in their best features variant) surpass all other matching functions in terms of retrieval effectiveness during RF. As discussed before, the conjunction becomes more effective than the weighted arithmetic mean from the first RF iteration onward.

In analogy to the significance tests between the arithmetic mean's and the conjunction's nDCG values described in Section 8.4.3, this part of the dissertation examines whether the effectiveness differences between the RF iterations of various matching function differ significantly. In addition, it discusses if the retrieval effectiveness between the two best matching functions differs at each RF iteration.

Table 8.19 lists the results of the Wilcoxon signed-rank test for the nDCG@20 metric between two RF iterations for three exemplary matching functions per collection. The objective of this analysis is to reveal whether the retrieval effectiveness changes significantly between two RF iterations. Again, this does not imply an increase of the effectiveness. The table shows clearly that the examined matching functions yield a change in the retrieval effectiveness when RF is given for the first time (RF It. 0 → It. 1) for all collections but Wang. A similar observation was made in Section 8.4.3 and is also most likely due to the good effectiveness values obtained with the Wang collection already without RF or fusion-based matching functions (see Section 8.4.2). Furthermore, in combination with the effectiveness measures presented above in Figure 8.21, these figures show that the matching functions adapt fast to the user's input preferences, which was requested by User Story 1.

The situation changes between the first and the second RF iteration when the same RF strategy is continuously pursued by the user simulation. In this case, the differences become insignificant for the disjunction with the Pythia and for all matching functions with the UCID collection. The same effect occurs at the transition between the second

Table 8.19: Wilcoxon signed-rank test  $p$ -values of differences between RF iterations' nDCG@20 of various matching functions

Collection	Matching Function	RF It. 0 → It. 1	RF It. 1 → It. 2	RF It. 2 → It. 3
Caltech 101	Conjunction, best features	< 0.0001	< 0.0001	< 0.0001
	Arithmetic mean, best features	< 0.0001	< 0.0001	< 0.0001
	Disjunction, best features	< 0.0001	< 0.0001	0.0003
Caltech 256	Conjunction, best features	< 0.0001	< 0.0001	< 0.0001
	Arithmetic mean, best features	< 0.0001	< 0.0001	< 0.0001
	Disjunction, best features	< 0.0001	< 0.0001	< 0.0001
MSRA-MM	Conjunction, best features	< 0.0001	< 0.0001	< 0.0001
	Arithmetic mean, best features	< 0.0001	0.0001	< 0.0001
	Disjunction, best features	< 0.0001	< 0.0001	0.0041
Pythia	Conjunction, best features	< 0.0001	< 0.0001	< 0.0001
	Arithmetic mean, best features	< 0.0001	< 0.0001	< 0.0001
	Disjunction, best features	< 0.0001	0.0965	< 0.0001
UCID	Conjunction, best features	< 0.0001	0.0450	0.0020
	Arithmetic mean, best features	< 0.0001	0.0637	0.1003
	Disjunction, best features	< 0.0001	0.0338	0.3615
Wang	Conjunction, best features	0.0020	0.0078	0.0039
	Arithmetic mean, best features	0.0098	0.0039	0.0039
	Disjunction, best features	0.0195	0.5566	0.3594

Computed by SAS 9.3 (PROC UNIVARIATE) [SAS Publishing 2011]

Insignificant differences are shaded.

and third RF iteration. The significant difference between the effectiveness of RF iteration 2 and 3 of the disjunction with the Pythia collection is also visible as a decrease in Figure B.5 (7-8). For MSRA-MM, the disjunction's effectiveness stagnates between the second and third RF iteration.

The results provide sufficient evidence that the PrefCQQL approach is effective when used with the described user simulation. However, its utility can be lowered during proceeding RF iterations. This effect is most likely due to the tactic used by the user simulation that only inspects the first 20 documents and inverts preferences when appropriate (see Section 8.5.1). In case a fair amount of relevant document is already present in the top-20 documents retrieved in RF iteration 1, the user simulation adds only few or no preferences that alter the weights of the matching function insufficiently in order to retrieve a rank with a significantly different nDCG@20 value. Interestingly, this effects seems to occur more likely with matching functions that are little effective such as the disjunction.

Table 8.20 compares the retrieval effectiveness values expressed by the nDCG@20 metric of the conjunction and the weighted arithmetic mean between the same RF iteration. The objective of this analysis is to unveil if the retrieval effectiveness differs significantly between these two matching functions at a given RF iteration. That is, whether it matters which of the two best matching functions is used during RF.

The table clearly shows that the differences are significant, with the exception of the first RF iteration for the Caltech 101 collection. Again, the differences with Wang are insignificant, which is not surprising given the arguments presented before. Despite

## 8 Evaluation of the Retrieval Effectiveness

Table 8.20: Wilcoxon signed-rank test  $p$ -values of the differences between the RF iterations' nDCG@20 values for the weighted arithmetic mean and the conjunction (best features variants)

		Arithmetic mean; lt. 0	Arithmetic mean; lt. 1	Arithmetic mean; lt. 2	Arithmetic mean; lt. 3
Caltech 101	Conjunction; lt. 0	< 0.0001			
	Conjunction; lt. 1		0.0003		
	Conjunction; lt. 2			< 0.0001	
	Conjunction; lt. 3				< 0.0001
Caltech 256	Conjunction; lt. 0	< 0.0001			
	Conjunction; lt. 1		< 0.0001		
	Conjunction; lt. 2			< 0.0001	
	Conjunction; lt. 3				< 0.0001
MSRA-MM	Conjunction; lt. 0	< 0.0001			
	Conjunction; lt. 1		< 0.0001		
	Conjunction; lt. 2			< 0.0001	
	Conjunction; lt. 3				< 0.0001
Pythia	Conjunction; lt. 0	< 0.0001			
	Conjunction; lt. 1		< 0.0001		
	Conjunction; lt. 2			< 0.0001	
	Conjunction; lt. 3				< 0.0001
UCID	Conjunction; lt. 0	< 0.0001			
	Conjunction; lt. 1		< 0.0001		
	Conjunction; lt. 2			< 0.0001	
	Conjunction; lt. 3				< 0.0001
Wang	Conjunction; lt. 0	0.0039			
	Conjunction; lt. 1		0.0039		
	Conjunction; lt. 2			0.0039	
	Conjunction; lt. 3				0.0078

Computed by SAS 9.3 (PROC UNIVARIATE) [SAS Publishing 2011]

Insignificant differences are shaded.

the exception with Caltech 101, Table 8.20 (in conjunction with the analyses presented before, e.g., in Figure 8.21) provides evidence that the conjunction is superior to the weighted arithmetic mean in terms of retrieval effectiveness during RF and that the conjunction yields different results from the arithmetic mean.

**Conclusion** As said before, the conjunction (best features variant) has been shown to be the most effective of the examined matching functions. This includes its consistent RF performance over all collections. It is followed by the weighted arithmetic mean (best features variant), which is the most effective matching function of those listed in Section 8.4.3 and shows a similar RF effectiveness pattern, although at a lower level.

The outcome of the experiments suggests that the consistency of the RF performance does not depend on the number of representations or the prevalent logical characteristic of the matching function. However, there is evidence showing that the retrieval effectiveness during RF depends on the logical structure of the matching function *and* the number of used representations to model the user's IN. This is in accordance with the hypothesis of the principle of polyrepresentation (see Section 3.2), which suggests that the most relevant documents will be located in the cognitive overlap – the intersection of different result sets formed by different representations. This intersection corresponds to the conjunction of all conditions which model the probability of relevance of each representation in CQQL.

In contrast, disjunctions degrade the retrieval effectiveness during RF. In general, disjunctive matching functions improve their effectiveness in the first RF iteration but fall off in quality directly thereafter. Notwithstanding, these matching functions never reach the same level of effectiveness as their conjunctive counter-parts at RF iteration 1. This phenomenon is clearly shown in Figures B.6, B.7, B.8 or B.9. This observation was also made in Section 8.4.3 in the absence of RF. Hence, the conclusion of Section 8.4.3 that conjunctive matching functions are in general more effective than their disjunctive counterpart can be reinforced.

The discussion on disjunctive matching function makes clear that PrefCQQL-based RF does not guarantee an increase of effectiveness. Similar effects become visible with the Q10 and semantic group conjunction matching functions, which both feature disjunctive characteristics. This provides further evidence that the presence of disjunctions in a matching function forms an obstacle for a positive effectiveness development during RF – a phenomenon also observable in the absence of RF (see Table 8.16).

However, conjunctive matching functions are also not free from problems. Figures 8.19 and 8.20 juxtapose the conjunction (best features variant) and the Eidenberger conjunction (variant 1). The figures clearly show a decrease of retrieval effectiveness for the Eidenberger conjunction, which features only four representations in comparison to the conjunction with twelve representations. This development is surprising because the Eidenberger conjunction has been proven to be relatively effective before (see Table 8.16) and is expressed with the help of conjunctions (see Matching Function 5). Nevertheless, its retrieval effectiveness degrades during RF. This phenomenon is an indication of an *underfitting* of the IN in conjunction with the learning algorithm used by

## 8 Evaluation of the Retrieval Effectiveness

PrefCQQL. Given that only four weights are available to satisfy the simuser-specified preferences, the learning algorithm has not enough parameters to capture the trend or “idea” of the existent IN and can only return a rough or oversimplified weight model resulting in an actual effectiveness decrease. Underfitting is also a reasonable explanation for Q10 suffering from the same issues, although it offers one more adjustable weighting variable (see Matching Function 9).

Interestingly, the Eidenberger disjunction variants do not suffer from comparable underfitting issues in the same way the Eidenberger conjunctions do.

Furthermore, the data suggests that disjunctions are relatively robust regarding underfitting although they show the typical “*disjunctive effect*” – a decrease of retrieval effectiveness after the first RF iteration. This effect is also characteristic to other disjunctions that feature more representations (see Figure 8.21; 10-12 vs. 13-15). In fact, the disjunctions perform almost similarly, although there is a slight indication that a high number of representations also leads to a higher effectiveness at RF iteration 1.

Given that the disjunctive effect is also present in the Eidenberger disjunctions, which feature only up to four representations (see Matching Functions 7 and 8), and that the effect occurs relatively consistent over all collections, there is no clear evidence that the effect can be explained as being caused by overfitting in case of the disjunction based on twelve representations. Instead, there is evidence that the behavior and effectiveness development are due to the disjunctive characteristics of matching functions.

This claim is supported by the examination of the semantic group conjunctions (see Matching Functions 10f.), which feature both a high number of weighting variables and representations as well as conjunctions and disjunctions. Figures B.10 and B.11 show a relatively unstable RF effectiveness development consisting of increases and decreases. As an exception, the performance of the semantic group conjunction (best features variant) constantly decreases until it starts to increase at the fifth RF iteration. We attribute this unsteady effectiveness to gradually emerging overfitting effects, which lower the robustness of the matching function. In other words, the learning algorithm is adapting towards “noise”, which is not needed to model the current IN, and therefore cannot provide a good prediction of the gradually developing IN and adjust the weighting variables’ values accordingly. Comparable effects cannot be observed with the disjunctive matching functions that suffer from the disjunctive effect but are relatively robust.

To sum up, the conjunction (best features variant) and the weighted arithmetic mean (best features variant) do not tend to overfit, i.e., they can be considered robust – at least for five RF iterations. As Figures 8.19 and 8.20 illustrate, the retrieval effectiveness of the conjunction gradually increases for the first five RF iterations. Whether this trend continues, the effectiveness stagnates, or overfitting effects occur after an intense training, i.e., a high number of RF iterations, is discussed in Section 8.5.3.

### 8.5.3 Long-Enduring Relevance Feedback Performance

Although there is no evidence in the literature that users are willing to give extensive relevance feedback, the closer examination of such – theoretically possible – extreme cases is compelling. The observation of the RF performance development over long-

enduring RF cycles allows insight into the sensitivity of PrefCQQL’s learning algorithm to overfitting. This investigation is particularly interesting because both matching functions examined in this section, the conjunction and disjunction (best features variants), have been shown to be consistent in terms of their RF performance. To conclude their effectiveness discussion, this section examines if the aforementioned matching functions are subject to overfitting in case of 15 RF iterations when used with the exemplary Pythia collection.

## Experimental Results

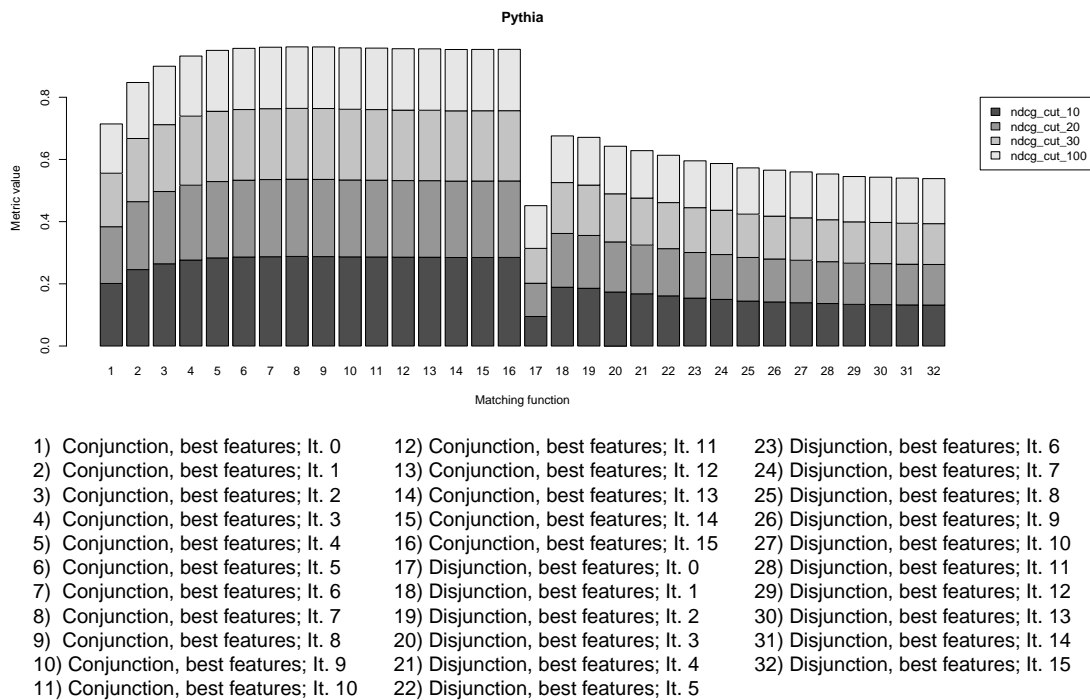


Figure 8.22: Effects of long-enduring relevance feedback

## Discussion

Figure 8.22 clearly shows that the characteristic RF effectiveness development patterns that have been discussed before start to fade from the fifth RF iteration onward. At this stage, the retrieval effectiveness of both matching functions stabilizes, i.e., it neither improves nor decreases significantly.

This assertion is confirmed – at least for the conjunction – by Table 8.21, which lists the results of the Wilcoxon signed-rank test of the obtained nDCG@20 values between the subsequent RF iterations of both matching functions. From the fifth RF iteration onward, the effectiveness of the conjunction does not change significantly when the RF

## 8 Evaluation of the Retrieval Effectiveness

tactic of the described user simulation is used (see Section 8.5.1). This effect is consistent with the explanation given in the discussion part of Section 8.5.2.

Although the retrieval effectiveness of the disjunction also tends to stabilize, Table 8.21 shows that it is subject to statistically significant change. From the ninth RF iteration onward, no significant change of the nDCG@20 values can be observed. This phenomenon gives further evidence that the decrease of the retrieval effectiveness might not be due to overfitting. Instead, it is most likely due to the disjunction itself which supports the claim made in the last section.

Since the effectiveness of the conjunction does not decrease, even at extreme cases such as 15 RF iterations, there is also no evidence for overfitting problems with this matching function. As a consequence, PrefCQQL can be considered robust with regard to the examined user simulation and matching functions. On the other hand, the learning algorithm of PrefCQQL tends to stabilize when no more informative preferences are input.

Table 8.21: Wilcoxon signed-rank test  $p$ -value results between long-enduring RF iterations based on their nDCG@20 values

RF Iteration	Conjunction	Disjunction
lt. 0 → lt. 1	< 0.0001	< 0.0001
lt. 1 → lt. 2	< 0.0001	0.0965
lt. 2 → lt. 3	< 0.0001	< 0.0001
lt. 3 → lt. 4	< 0.0001	0.0375
lt. 4 → lt. 5	0.0339	0.0023
lt. 5 → lt. 6	0.7700	< 0.0001
lt. 6 → lt. 7	0.8983	0.0010
lt. 7 → lt. 8	0.6084	< 0.0001
lt. 8 → lt. 9	0.0291	0.0007
lt. 9 → lt. 10	0.6876	0.0090
lt. 10 → lt. 11	0.1089	0.0010
lt. 11 → lt. 12	0.4083	0.0883
lt. 12 → lt. 13	0.0965	0.1185
lt. 13 → lt. 14	0.8646	0.0131
lt. 14 → lt. 15	0.9470	0.3972

Computed by SAS 9.3 (PROC UNIVARIATE) [SAS Publishing 2011]

Insignificant differences are shaded.

### 8.5.4 Development of the Weighting Variables

The last sections addressed the retrieval effectiveness of the PrefCQQL approach in a typical QBE setting. This section focusses on a particular detail of PrefCQQL: the development of the weight values during RF, i.e., how the change of the weight values caused by PrefCQQL's learning algorithm correlates with the change of the retrieval effectiveness. The objective of this section is to reveal whether statistics of the weight values can be used to draw conclusions about the system's or user's current state in order to assist during the IIR process as suggested by Ingwersen [1996].

As said before, each matching function contains a number of weighting variables



$\theta_i \in [0, 1]$ . The syntax of the matching functions and the indices of the weighting variables, as well as their implementation, is given in Appendix B.1. Constrained by the given matching function, the PrefCQQL learning algorithm (see Section 5.4.4) tries to learn a weighting scheme  $\omega$  (see Definition 5.21), which assigns each weighting variable a value, based on the user's preferences. In other words, each RF iteration  $\tau$  corresponds to a particular  $\omega_\tau$  yielding a specific rank of result documents. In the following, we interpret  $\omega_\tau$  as a vector whose  $i$ -th component corresponds to the weight value of  $\theta_i$ . To give an example, the weighting scheme vector  $\omega_\tau = (1, 1, 1, 1)$  assigns the weight value 1 to each weighting variable of the Eidenberger conjunction (variant 1):  $\bigwedge_{\theta_i}(R_5, R_7, R_8, R_{11})$ .

### Descriptive Statistics of the Weight Values

To start with, Table 8.22 lists the descriptive summary statistics of the weight values of the weighting scheme vector  $\omega$  averaged over all examined RF iterations and collections for different matching functions. For a better comparability of the location parameters of the weight values, the data is also available as a boxplot in Figure 8.24.

Table 8.22: Summary statistics of the weights values over all RF iterations and collections for different matching functions

	Matching Function (see Appendix B.1, #)	Mean	Median	Min	Max	Gini
1	Conjunction (1)	0.4116	0.4353	0.1988	0.6605	0.1690
2	Conjunction, best features (2)	0.3569	0.3327	0.1888	0.5979	0.1946
3	Disjunction (3)	0.3638	0.3588	0.1808	0.5567	0.1765
4	Disjunction, best features (4)	0.3790	0.3737	0.2172	0.5700	0.1754
5	Eidenberger conjunction, variant 1 (5)	0.3780	0.3816	0.2796	0.4691	0.1208
6	Eidenberger conjunction, variant 2 (6)	0.3926	0.4738	0.1817	0.5223	0.1928
7	Eidenberger disjunction, variant 1 (7)	0.5625	0.5618	0.4649	0.6615	0.0733
8	Eidenberger disjunction, variant 2 (8)	0.7635	0.8463	0.5465	0.8976	0.1022
9	Q10 (9)	0.4439	0.4517	0.2391	0.6532	0.1995
10	Semantic group conjunction (10)	0.4859	0.4811	0.1953	0.7686	0.2146
11	Semantic group conjunction, best feat. (11)	0.4545	0.4673	0.2439	0.7825	0.1596
12	Weighted arithmetic mean, best feat. (15)	0.3480	0.4431	0.1048	0.5553	0.2536

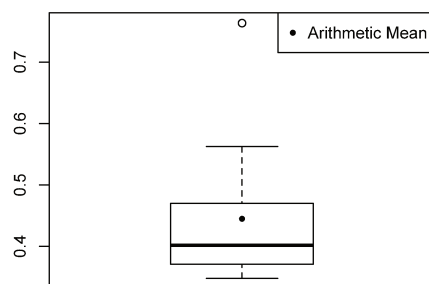
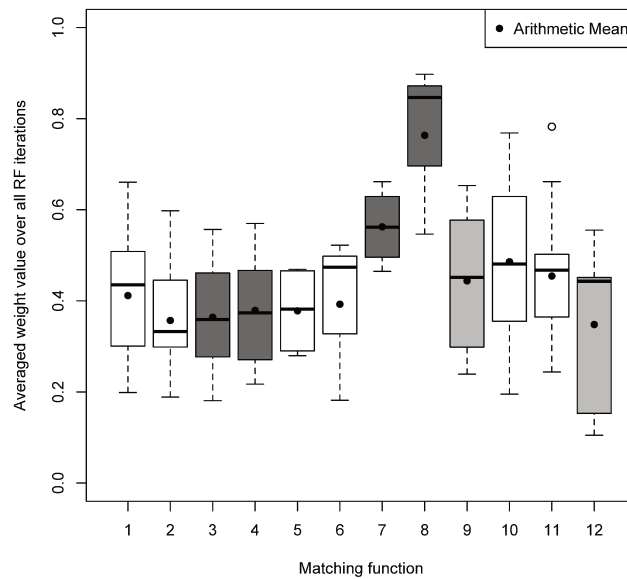


Figure 8.23: Boxplot of the mean weights values over all matching functions

Figure 8.23 illustrates that the mean of the components of  $\omega$ , i.e., the weight values of the different matching functions, lies in a relatively small interval with one outlier:

## 8 Evaluation of the Retrieval Effectiveness



- |                               |                                       |   |
|-------------------------------|---------------------------------------|---|
| 1) Conjunction                | 5) Eidenberger conjunction, variant 1 | 9) Q10  |
| 2) Conjunction, best features | 6) Eidenberger conjunction, variant 2 | 10) Semantic group conjunction                |
| 3) Disjunction                | 7) Eidenberger disjunction, variant 1 | 11) Semantic group conjunction, best features |
| 4) Disjunction, best features | 8) Eidenberger disjunction, variant 2 | 12) Weighted arithmetic mean, best features   |

Dark gray indicates disjunctive and gray indicates matching functions w/o a dominant logical characteristic.

Figure 8.24: Boxplot of the standard deviation of the weights values for different matching functions

the Eidenberger disjunction (variant 2). Furthermore, the weight values never reach the extreme points of 0 and 1 on average. As Appendices B.3.1 and B.3.2 show, this phenomenon is not due to averaging effects – at least not for the first five RF iterations. To inspect the weight value development in more detail, refer to Appendix B.3.1 that lists weight value development graphs per matching function and collection. Furthermore, the graphs illustrate the weight value development per weighting variable averaged over all RF iterations. To complete the examination, Appendix B.3.2 displays the weight value development per RF iteration separated by matching functions and collections on the right of each figure.

Figures B.12 and B.16 exemplify that the relation between the weighting variables' values is not constant between the collections. That is, the weight values do not have to develop identically over all collections. In other words, each collection might have a distinct averaged weighting scheme that is learned by PrefCQQL.

If the collections are examined separately, it becomes clear that the relation between the weighting variables' values stays relatively similar throughout the RF iterations as Appendix B.3.2 shows. That is, the weight values follow mostly a characteristic pattern

towards the values converge. For an example, see Figure B.18 (right).

Interestingly, as Figure 8.24 illustrates, the statistical dispersion of the averaged weight values varies differently per matching function. Unfortunately, this effect can neither be attributed to the logical structure of the matching function nor its number of used representations on the basis of the available data.

To conclude, the inequality of the weight values within  $\omega_\tau$  is investigated in order to find out if all weighting variables (and therefore representations) contribute equally to the calculation of a document's probability of relevance. One common coefficient to measure the inequality of values in a distribution is the *Gini coefficient* [Gini 1997]. It was originally proposed to measure the income inequality in different countries but can also be used to determine the inequality of the weight value (the "income") distribution amongst the weighting variables. The Gini coefficient yields 0 if the "income" is equally distributed and 1 if only one entity (or weighting variable in our case) perceives all the "income" available in the system, i.e., maximum inequality. Table 8.22 illustrates that the weight values are relatively equally distributed amongst the weighting variables. That is, many weighting variables obtain a weight value of similar magnitude.

### Weighting Scheme Distances and Inequality of the Learned Weight Values

In order to evaluate the effectiveness of PrefCQQL's learning algorithm, this part of the dissertation discusses the distance between two weighting scheme vectors  $\omega_\tau$  and  $\omega_{\tau'}$ , which are derived from a set of user preferences during two subsequent RF iterations  $\tau$  and  $\tau'$ . This evaluation is motivated by the hypothesis that the learning algorithm has an enduring strong impact on the weighting scheme, i.e., each learning step changes the weight values significantly.

At first sight, one way to calculate the difference between two weighting schemes is the Kullback-Leibler divergence [Kullback & Leibler 1951] that determines the difference between two probability distributions. Unfortunately,  $\omega_\tau$  itself is not a probability distribution and cannot be interpreted as such because the weight values in CQQL do not necessarily sum up to 1 (see Section 4.3). Furthermore, the Kullback-Leibler divergence would require all weight values to be greater than  $0^{149}$ , which cannot be guaranteed in general, although this requirement is satisfied by the presented experiments.

As a consequence, the class of Minkowski distances (see Section 2.3.2) is a viable means to calculate the difference between two weighting scheme vectors. The Manhattan distance (see Definition 2.24,  $p = 1$ ) is particularly feasible because we are only interested in the "way" between  $\omega_\tau$  and  $\omega_{\tau'}$ . That is, "situations where for example a difference of 1 in the first variable, and of 3 in the second variable is the same as a difference of 2 in the first variable and of 2 in the second" [Kaufman & Rousseeuw 2009, p. 12].

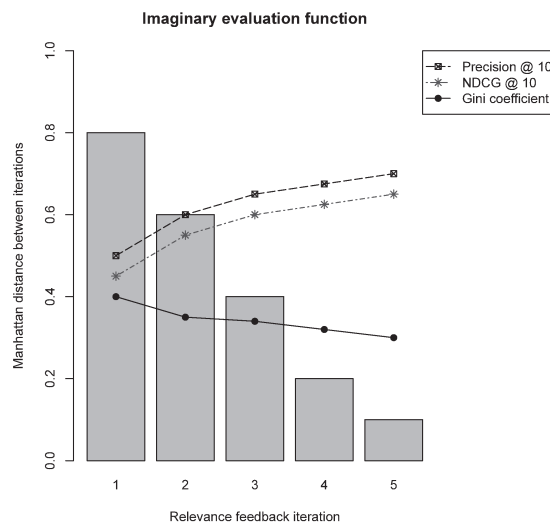
As mentioned above, another interesting research question regarding PrefCQQL is to reveal whether there is a relation between the distance of the weighting schemes between two subsequent RF iterations, their Gini coefficient and the resulting retrieval

<sup>149</sup>Otherwise a division by zero would occur.

## 8 Evaluation of the Retrieval Effectiveness

effectiveness. Figure 8.25 presents a composite plot of various features of an imaginary matching function. The plot is meant to juxtapose effectiveness measures, weighting scheme distances, and the Gini coefficients of  $\omega_\tau$  at RF iteration  $\tau$ . The  $x$ -axis of the plot denotes the RF iteration. RF iteration 0 is omitted because at this iteration all components of  $\omega_\tau$  are set to 1, the Gini coefficient is 0, and the distance cannot be calculated because of the missing reference point<sup>150</sup>. Each bar displays the Manhattan distance between  $\omega_\tau$  at RF iteration  $\tau$  indicated by the tick at the  $x$ -axis and  $\omega_{\tau-1}$  of the preceding RF iteration. That is, the bar at tick 2 shows the distance between  $\omega_1$  and  $\omega_2$ .

The  $y$ -axis indicates the value of the Manhattan distance as well as the value of all other measures at the given RF iteration.



All weighting variables are initially set to 1.

Figure 8.25: Hypothetic weight development analysis with respect to retrieval effectiveness

The imaginary matching function's features visualized in Figure 8.25 illustrate that the distance between  $\omega_0$  and  $\omega_1$  is the greatest and gradually decreases thereafter. This behavior is motivated by the observations made in Section 8.5.2, which show that the retrieval effectiveness typically changes most at RF iteration 1. We attribute this observation to a significant change of weight values.

Additionally, the matching function is effective. This can be concluded from the increasing retrieval effectiveness measures precision @ 10 and nDCG @ 10. Moreover, the effectiveness metric curves are inversely proportional to the Manhattan distance bars. That is, the effectiveness raises slower as the distance falls. In case of an ineffective matching function, these curves would fall proportionally to the decrease of the dis-

<sup>150</sup>There is no RF iteration -1 and hence no  $\omega_{-1}$  of which one could calculate the distance to  $\omega_0$ .

tance values. This is due to the aforementioned hypothesis, which assumes that the weighting scheme distance between two RF iterations correlates with effectiveness.

Furthermore, Figure 8.25 displays the Gini coefficient of  $\omega_t$  at RF iteration  $t$  that stabilizes throughout the RF iterations. This effect is based on the assumption that the inequality of the weight values correlates with the weighting scheme distance. Hence, the Gini coefficient is supposed to stabilize (expressing that an optimal distribution is approached), while the distance changes in smaller steps.

The complete plots for all examined matching functions and collections are available in Appendix B.3.2. Unfortunately, these plots do not justify the presented hypothesis. For instance, while Figure B.18 is in full accordance with the made assumptions, Figure B.33 falsifies the hypothesis because of the varying weighting scheme distances that have no significant effect on the Gini coefficient.

For the sake of brevity, we have omitted here a discussion of all 142 plots presented in Appendix B.3.2. Instead, the next section focusses on an analysis of the correlation between the Gini coefficient, the weight distance, and the retrieval effectiveness exemplarily represented by the nDCG@10 metric.

### Correlation of Weighting Scheme Distances and Retrieval Effectiveness

Figure 8.26 shows three scatter plots that display the correlation of different statistics aggregated over all examined collections. Separate scatter plots per collection are available in Appendix B.3.3. The upper left plot shows the correlation of the Gini coefficient and nDCG@10, while the upper right contrasts the Gini coefficient with the weighting scheme distance (denoted as Manhattan weight distance). The bottom plot illustrates the correlation of the weighting scheme distance and the retrieval effectiveness, which is expressed using nDCG@10. In addition, all plots contain the corresponding fitted linear model of the data that is drawn with a dashed line<sup>151</sup>.

Furthermore, the bottom line of each plot displays the value of the *Pearson product-moment correlation coefficient* [Mendenhall et al. 1999, Sec. 12.8]<sup>152</sup>. The Pearson product-moment correlation coefficient lies in the interval  $[-1; 1]$  and is a common measure of the linear correlation between the two variables given at the  $x$ - and  $y$ -axis of each respective plot. A value of 1 indicates total positive correlation of the two variables, -1 shows total negative correlation, and 0 means that there is no correlation.

In the following, the scatter plots are discussed in clockwise order starting from the upper left plot. The first scatter plot displays a tiny negative correlation between the Gini coefficient and nDCG@10. Given the small value of the Pearson product-moment correlation coefficient of -0.0667, it is not appropriate to acknowledge a resilient correlation between these variables. That is, it is not reasonable to assume that a change of the weight value distribution amongst the weighting variables will result in a change of the retrieval effectiveness during RF. Interestingly, the correlation coefficient's value is mainly due to the observed correlation with the Pythia and Wang collections as the

<sup>151</sup>The linear fitted model is calculated with R 2.15 standard settings [R Core Team 2012].

<sup>152</sup>The correlation coefficient is computed by the implementation provided by R 2.15 [R Core Team 2012].

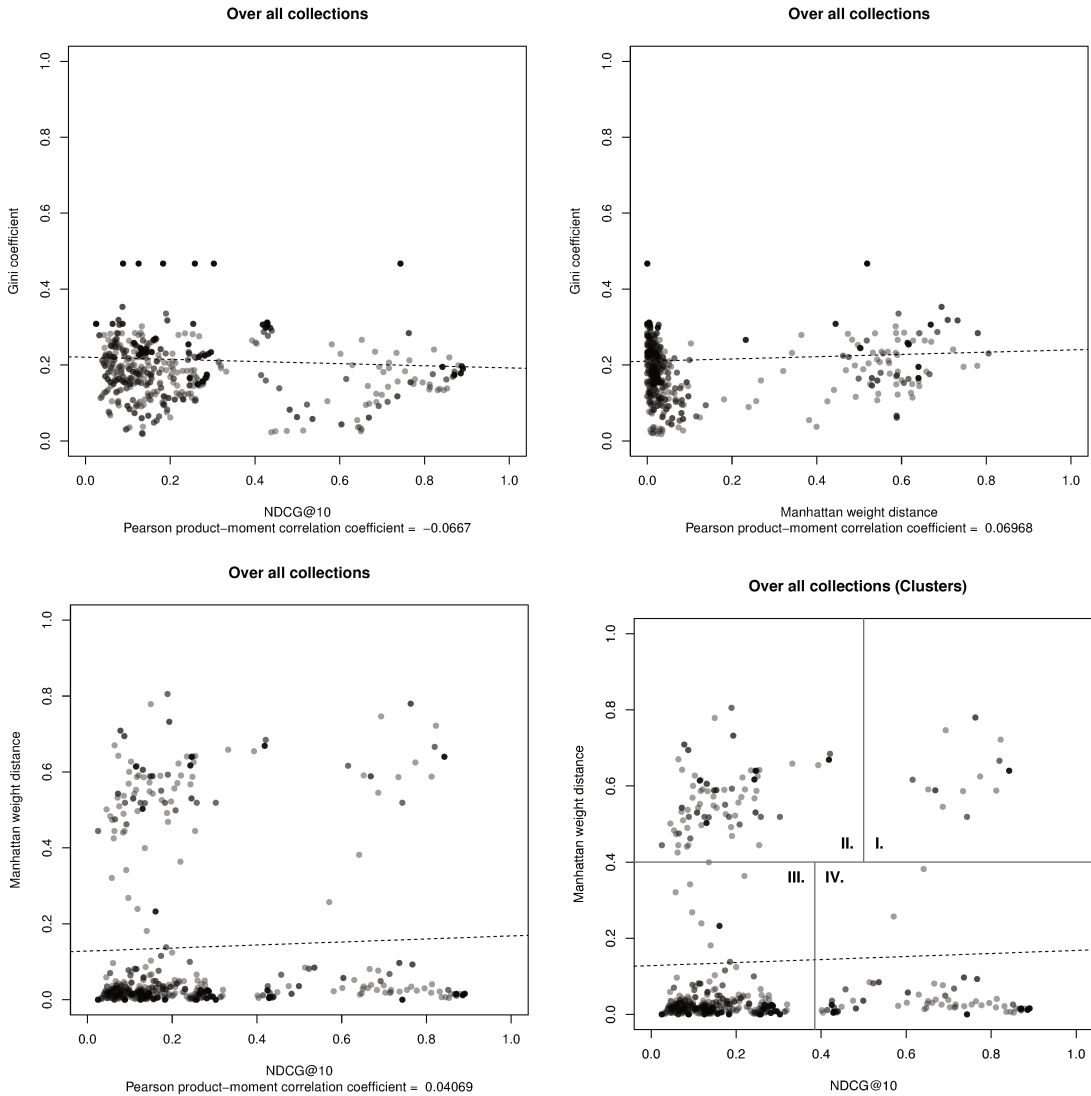
## 8 Evaluation of the Retrieval Effectiveness

other collections indicate a correlation ranging from slightly to considerably positive (see Figures B.90 and B.91).

A similar statement to the one regarding the first scatter plot can be made for the second that indicates a comparably small but positive correlation between the Gini coefficient and the weighting scheme distance. Unlike the correlation between the Gini coefficient and nDCG@10, the separated examination of the correlation between the Gini coefficient and the weighting scheme distance per collection does not yield further insights as Figures B.92 and B.93 illustrate.

Although the last scatter plot does not show a correlation between the weighting scheme distance and the retrieval effectiveness, it is interesting because it brings out four more or less distinct clusters that are present in the experimental data. It is noteworthy that cluster I. is sparsely populated, i.e., there are a few matching functions that feature both a large distance between the weighting schemes between two RF iterations and a high effectiveness. The second cluster contains the observations with a relatively strong change of the weight values but with a relatively low effectiveness. Cluster III. contains all observations with a small change of the weighting schemes and are relatively weak retrieval performance. This cluster is densely populated indicating that many RF iterations result into a small change of the weight values learnt by PrefCQQL's weight learning algorithm on the basis of the preferences input by the simulated user. The fourth cluster, which is slightly more populated than the second, contains all observations that are relatively effective regarding the retrieval of relevant documents but that also show a small change in the learned weight values.

The finding that there are a lot of observations with a low effectiveness is not surprising as a high number of matching functions with low effectiveness has been presented in Section 8.5.2. Really interesting is the fact that most observations have a small weighting scheme distance and a varying degree of effectiveness as Figure 8.26 (bottom) clearly illustrates. In other words, the utilized user simulation's preference input (or RF) yields small changes of the weight values that might result in an increase of the retrieval effectiveness. The discussion in Section 8.5.2 showed that the retrieval effectiveness of a matching function mainly depends on its logical structure and the number of used representations. This disproves the hypothesis made in the last part of this section, which states that PrefCQQL's learning algorithm will have a strong impact on the weighting scheme's values (see Page 275). Instead, the data shows that even slight modifications of the weight values have an impact on the change of the retrieval effectiveness during RF. Moreover, the data clearly illustrates that the most dramatic adaption of the weighting scheme towards the user's needs happens between RF iteration 0 and 1 (see Appendix B.3.2). Unfortunately, the question whether statistics of the weight values might be used to conclude the state of the retrieval process, posed at the beginning of Section 8.5.4, has to be answered negatively. From a statistical point of view, there is no evidence of a resilient correlation between the Gini coefficient, the weighting scheme distance, or the retrieval effectiveness.



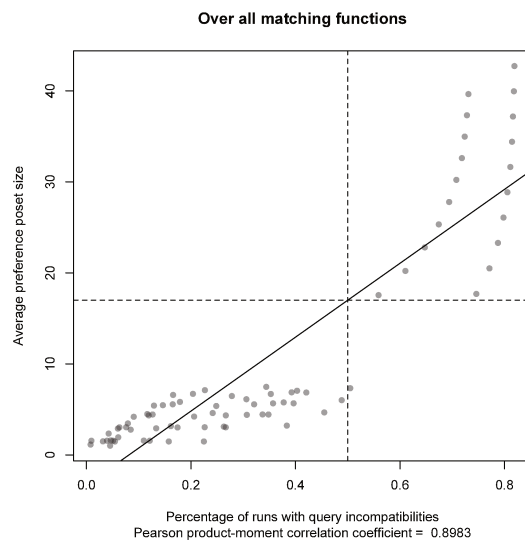
Dashed lines indicate the fitted linear model of the data. Darker points indicate higher densities.

Figure 8.26: Correlation of different measures with the weighting scheme distance between RF iterations

### 8.5.5 Query Incompatibilities and Preference Poset Size

Section 5.4.7 addressed the problem of query incompatibilities, i.e., the cases in which a user-specified preference cannot be expressed with the current CQQL condition serving as the matching function. This section examines the correlation between the number of preferences and the percentage of query incompatibilities of the submitted queries in order to assess how often query incompatibilities occur in a typical QBE setting as described above in Section 8.5.1.

Figure 8.27 illustrates the correlation between the average preference poset size and the percentage of runs (i.e., the number of queries submitted to the MIR system) causing query incompatibilities. The dashed lines show the point from which  $\geq 50\%$  of the runs have query incompatibilities, i.e., an average preference poset size of approximately 17. The solid line depicts the fitted linear model of the data computed with the aforementioned R standard parameters. As before, the plot also discloses the value of the Pearson product-moment correlation coefficient of the data.



The solid line indicates the fitted linear model of the data. Darker points indicate higher densities.

Figure 8.27: Correlation of query incompatibilities and average preference poset size

The Pearson product-moment correlation coefficient, the fitted linear model, and the scatter plot clearly show that there is a positive correlation between the average preference poset size and the percentage of runs with query incompatibilities. In other words, the chance to observe a query incompatibility rises with the number of stated preferences. Although this result is not surprising, it is reasonable to further investigate the impact of the matching function on the correlation of the average preference poset size and the percentage of runs with query incompatibilities.

The bubble plot in Figure 8.29 illustrates this relation. The  $y$ -axis lists the different matching functions at different RF iterations and the  $x$ -axis shows the average pref-



erence poset size. The areas of the circles indicate the percentage of observed query incompatibilities with respect to the number of total submitted queries. To facilitate the comparison, a 100 % reference circle is depicted in the upper right corner of the plot. For the sake of visual clarity, the contents of Figure 8.29 are also included in the supplement to this dissertation along with the detailed percentages of query incompatibilities per collection (see Appendix E).

Besides providing a general overview, Figure 8.29 allows further insights into the relation of matching functions, preference poset sizes, and the amount of query incompatibilities.

First, it becomes apparent that the number of specified preferences (average preference poset size) varies per matching function and RF iteration.

Second, the best performing matching functions differ in their ratio between the percentage of query incompatibilities and the number of specified preferences. Figure 8.28 visualizes this ratio on the  $y$ -axis, whereas the  $x$ -axis indicates the RF iteration. The conjunction (best features variant) is stable over the first two RF iterations, while the weighted arithmetic mean (best features variant) causes a higher amount of query incompatibilities. Later on, their ratios converge. Hence, the conjunction (best features variant) has a lower risk of causing a query incompatibility in relation to the number of preferences during early RF phases than the weighted arithmetic mean.

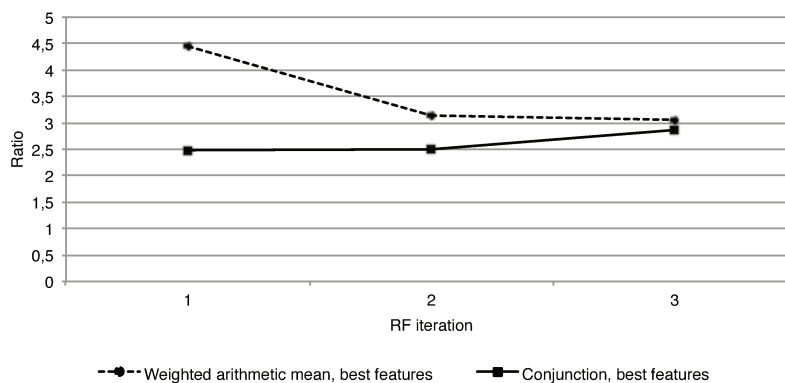


Figure 8.28: Ratio between query incompatibility percentage and number of specified preferences

In general, one can observe a deterioration in quality from RF iteration 5 onward. That is, the risk of observing query incompatibilities rises significantly between RF iteration 5 and 6. For instance, consider the conjunction (best features variant): 22.7 % of the submitted queries result in a query incompatibility at RF iteration 5, while 55.9 % of the runs at the sixth RF iteration yield query incompatibilities. Regarding the disjunction (best features variant), the gain of query incompatibilities is even more severe. The figure rises from 34.5 % to 74.6 %. Interestingly, this phenomenon corresponds almost perfectly to the moment from which the retrieval effectiveness during RF stagnates (see Section 8.5.3). Given the fact that the Pythia MIR system uses the last valid

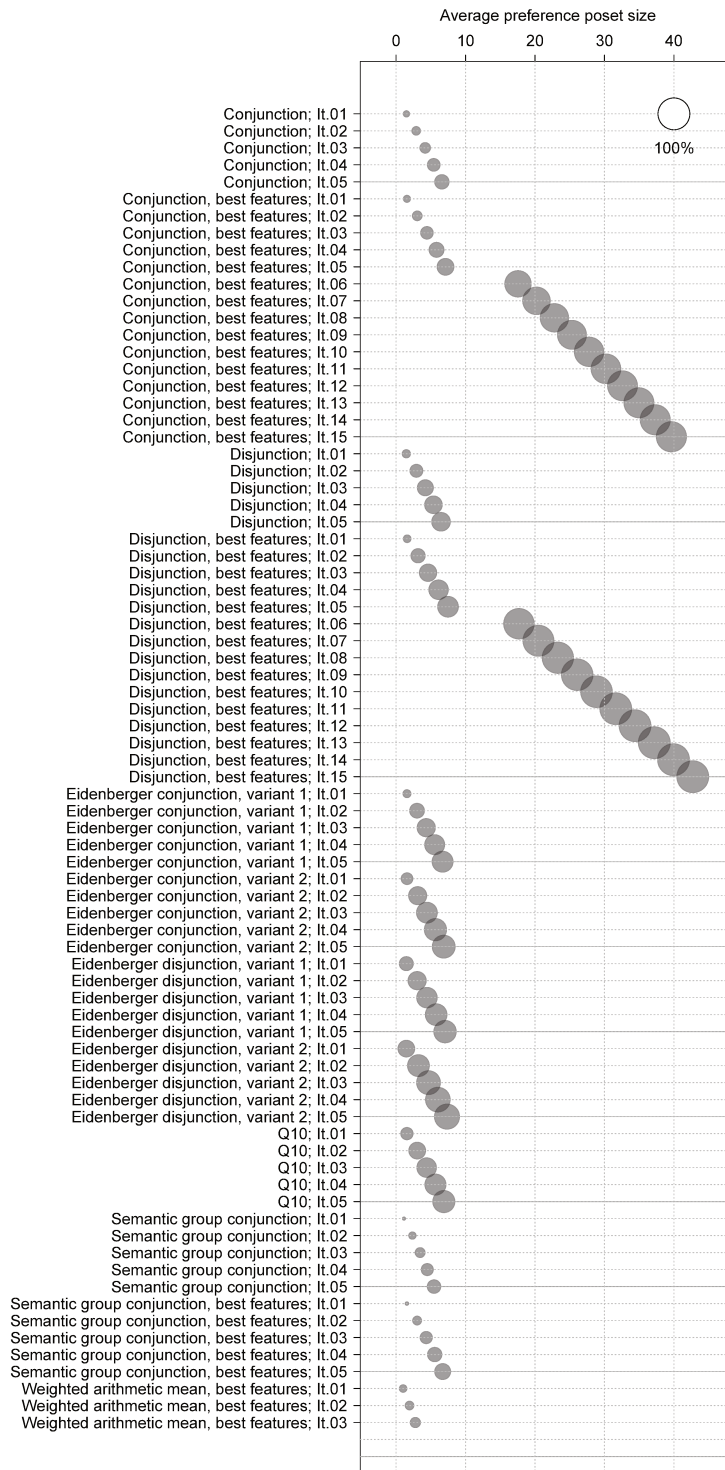
## 8 Evaluation of the Retrieval Effectiveness

weights (see Section 8.5.1) whenever a query incompatibility is detected, the explanation for the stagnation of the retrieval effectiveness during long-enduring RF has to be extended. Apparently, the high percentage of query incompatibilities has an impact on the observed retrieval effectiveness as well.

Moreover, there is evidence suggesting that matching functions with few weighting variables are subject to a quick increase of the query incompatibility percentage as the Eidenberger variants or Q10 clearly illustrate. As before, this effect can be attributed to an underfitting problem because the small amount of weighting variables offer insufficient means for PrefCQQL's learning algorithm to find a utility function for the specified preferences.

Consequently, this argument also explains the relatively small percentage of query incompatibilities for the semantic group variants. These matching functions feature the highest amount of weighting variables and therefore provide enough freedom for the learning algorithm to satisfy the given set of preferences.

## 8.5 Retrieval Effectiveness of PrefCQQL



Area of circles indicates percentage of query incompatibilities.

Figure 8.29: Bubble chart of matching function, average preference poset size per RF iteration, and percentage of query incompatibilities

## 8 Evaluation of the Retrieval Effectiveness

## 9 Evaluation of the User Experience

Chapter 7 motivated the division of the evaluation presented in this dissertation into the examination of retrieval effectiveness and usability. Following a thorough discussion on the retrieval effectiveness of the CQQL approach in MIR, this chapter focusses on the usability of the Pythia MIR system and its user experience.

Because of the complexity and cost of a full-featured user study, only a cutout of all possible usability issues of a limited group of users and work tasks are investigated. Furthermore, not all possible GUI variants or interaction designs could be tested. Thus, the results presented in this chapter must be seen as an *initial* user study. However, this does not mean that the outcome of the study is meaningless. On the contrary, the results show quantitatively and qualitatively founded trends regarding the usability of the Pythia MIR system and provide resilient insights into its user experience. The validity and resilience of the user study is covered separately in Section 9.3.1.

From a conceptual point of view, the studies discussed in this chapter complement the feedback obtained from the user workshops presented in Section 6.1 as part of the user-centered design process of the Pythia MIR system. In addition to the more or less formal reviews conducted throughout its development phase, Section 9.3 presents a formal, quantitative examination of the system's usability based on the ISO 9241/110 norms.

In order to gain subjective and qualitative insights into the user experience and preferences between the MIR system's tools and mechanisms presented in Chapter 6, an additional questionnaire is used, which is discussed in Section 9.4. The utilization of qualitative methods is highly advocated by experts in the field of usability as well as in social sciences. For instance, Cooper et al. [2007] highlight two main advantages of qualitative over quantitative research: the sensitivity towards nuances of user behavior and the high complexity of (hidden) variables that hinder the utilization of quantitative methods in order to examine human behavior.

To supplement the results obtained from the qualitative study, Section 9.5 examines screen recordings of five randomly chosen subjects to learn about specific usability problems.

The assessment of the user experience of the Pythia MIR system is based on the judgments of a group of 59 test persons whose demographics and basic behavioral variables are presented in Section 9.2. This group was confronted with a common simulated work task that had to be solved with three GUI variants based on the same MIR system. This work task forms the basis for both the quantitative and the qualitative parts of the user study. The complete experimental setup is described in the next section.

### 9.1 Experimental Setup

All described experiments were conducted on Apple MacPro workstations (version 4,1) with 2 Quad-Core Intel Xeon CPUs with 2.26 GHz and 8 GB RAM running Mac OS X 10.6.8 (German localized version) on the internal hard disk. Each computer was equipped with the same mouse and keyboard model, the desktop design was identical, and all subjects had access to the same file system hierarchy and contents. The computers shared the same screen setup, i.e., a 30" Apple Cinema HD display running a resolution of 2560×1600 pixels. Each test person had access to a separated computer located in a university laboratory.

Before the actual test was started, all subjects were given a short written and illustrated tutorial (see Appendix E) that motivated the study and gave a quick overview about the central GUI components as well as Mac OS X specific specialities (such as the position of window resizing handles etc.). The introduction was available to the participants throughout the whole study.

In order to assess the usability of the system, the test persons had to solve the following simulated work task with three GUI variants, which are presented in the next subsections. The full instructions given to the subjects can be found in Appendix C.3.2.

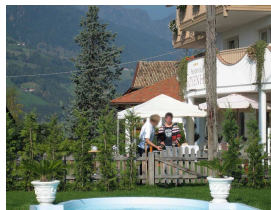


Figure 9.1: Initial QBE document of the usability test

**Simulated work task** *Imagine you have been searching similar images to the one shown above [Figure 9.1]. You did expect to retrieve more images that resemble the sample images given below [Figure 9.2]. To start with, the first query image has been pre-defined.*

- *Your objective is to get as many photographs of Alpine mountain landscapes amongst the first 20-30 results as possible.*
- *Photographs depicting persons should be avoided.*
- *The desired images focus on the landscape. Potentially visible buildings should only be a part of the motif but are not its central point.*
- *Mountain or landscape photographs that are not taken in Europe or the Alps are not relevant.*

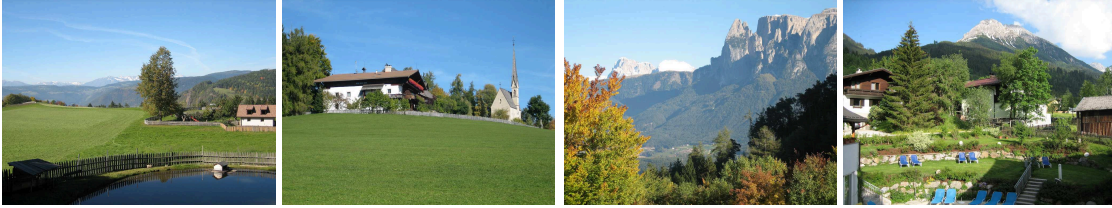


Figure 9.2: Sample target images of the usability test

To limit the impact of tiring and learning effects on the evaluation, the order of the confrontation with the GUI variants was permuted per subject. In total, the participants were given a maximum of 15 minutes to interact with each variant. After 15 minutes or anytime the users decided to abort working with a GUI variant because they were satisfied with the obtained results, the study participants were taken to a Web-based questionnaire with 35 usability related questions (see Appendix B.4) meant to assess the GUI variant's usability.

The usability questionnaire is based on the ISO 9241/10 and ISO 9241/110 norms on software ergonomics. After all GUI variants were tested, the participants of the study had to fill in an additional questionnaire with nine questions about their demographics and computer usage, and eight qualitative questions asking about their personal preference between the tested GUI variants (see Appendix C.3.1).

In total, each participant of the study had to answer 122 usability-related questions and could give an optional textual comment in form of an open question.

### 9.1.1 Restrictions of the Retrieval Engine

As explained in Chapter 7, it is hard to isolate statements about a MIR system's usability from statements about its retrieval effectiveness because both aspects of a system interact with each other and affect the overall user experience. Thus, the matching function and the collection used in the experiments was fixed to establish a common retrieval effectiveness baseline for all test persons. In other words, the subjects could adapt the utilized CQQL query's weights but could not change its logical structure. As a consequence, the fixation of the retrieval variables allows a comparison of the functionally different GUI variants that will be mainly based on the variables affecting their usability and user experience.

Throughout the experiment, the participants of the study could only access images from the Pythia collection (see Section 8.3) and had to use the Bielefeld conjunction<sup>153</sup>. This conjunctive matching function combines various visual representations, GPS data, the time and date a picture was taken, the used camera model, and a face detection-based person presence indicator, i.e., whether one or more persons is depicted in the image. Additionally, it combines different retrieval paradigms. For instance, the person presence indicator is a Boolean predicate, while the GPS data uses spatial proximity

<sup>153</sup>This CQQL-based conjunction is named after the user workshops that took place in the city of Bielefeld and where the matching function was used for the first time.

## 9 Evaluation of the User Experience

calculations and the visual representations rely on typical MIR retrieval techniques as outlined in Section 2.3.2. The form of a conjunction has been chosen to implement a cognitive overlap that is advocated by the principle of polyrepresentation and the compelling results shown with this class of matching functions during the retrieval effectiveness evaluation (see Section 8.5). For a detailed description of this matching function, see Appendix B.1.1 (Matching Function 12).

To complete the description of the experimental setup, the GUI variants used during the usability study are presented in the following sections.

### 9.1.2 Introductory Phase

After the participants of the study had read the tutorial (see Section 9.1), they had a maximum of 30 minutes to familiarize themselves with the Pythia MIR system and the GUI mechanisms of Mac OS X. During the introductory phase, the subjects were allowed to interact with the MIR system's GUI without restrictions. That is, the participants of the study had access to all GUI components presented in Section 6.3.2. A distinct task was not defined. Instead, each test person had full access to the Pythia collection (see Section 8.3) and could use the included images as queries. An exemplary user interaction during the introductory phase is available as a video supplement to this thesis (see Appendix E).

Although all participants could ask for help in case of problems with the Mac OS X interface<sup>154</sup>, e.g., if a window was minimized by accident or if they needed help with the file manager, there was no pro-active support from staff supervising the study. Assistance regarding the Pythia MIR system was only available in form of the aforementioned tutorial, which could be accessed whenever needed.

After 30 minutes or whenever a subject decided to end the introductory phase, she or he was taken to one of the GUI variants, which are described in the following subsections.

### 9.1.3 GUI Variant T2

Figure 9.3 depicts the available widgets for GUI variant T2. From a functional point of view, this GUI resembles the similarity search feature of Google's image search (see Figure 9.4; red arrow). The main differences to Google are that a similarity search in T2 is started via a drag and drop mechanism instead of a click in a menu, the support of multiple QBE documents, and the (optional) expert search functionality in form of a representation weight setup dialog.

In accordance with Bing's or Google's image search, the result images are visualized in a matrix that is sorted from the upper left side to the lower right by descending probability of relevance. In addition, the current influence of the representations on the query can be visualized using the weight inspector.

---

<sup>154</sup>This assistance was necessary because it could not be assumed that all participants were familiar with Mac OS X. In fact, only one third of the test persons had worked with this operating system before (see Section 9.2).



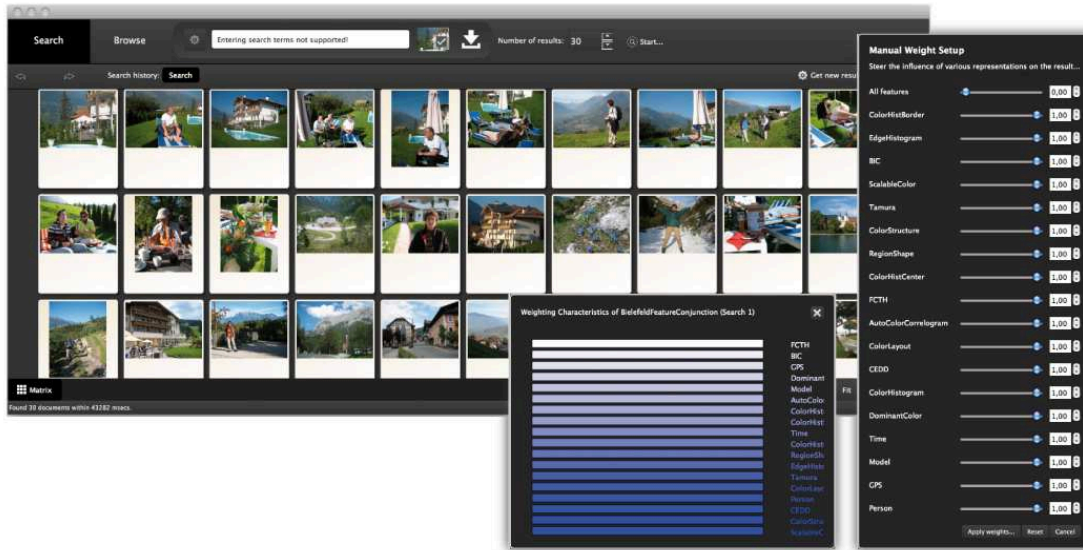


Figure 9.3: Available widgets in GUI variant T2

The browsing and textual query functionality was disabled during the test.

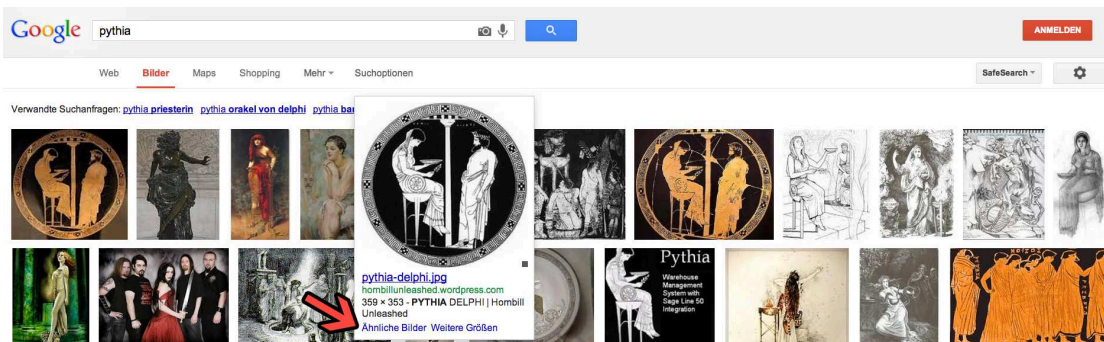


Figure 9.4: Similarity search feature in Google’s image search

Because of the functional similarity between T2 and Google, one could argue to test Google directly as a baseline system. As said before, the user experience of a MIR system is affected by its retrieval effectiveness. Hence, the direction comparison with Google would distort the results of the study because Google’s retrieval engine obviously could not be aligned with the one of Pythia MIR.

Furthermore, such a comparison would also be heavily influenced by the aesthetics of the respective GUI [Hearst 2009, Sec. 1.11]. For instance, van der Heijden [2003] has shown that the aesthetic of a website has an impact on its usability. Therefore, an internally developed GUI is used to keep the impact of this variable on the results

## 9 Evaluation of the User Experience

of the usability test low. In any case, it is possible to interpret T2 as the functional baseline, although it has been extended by the manual weighting dialog. From this perspective, the study design can also be seen as part of a “competitive usability testing” [Shneiderman & Plaisant 2005, p. 148] with GUI design counterbalancing.

### 9.1.4 GUI Variant T3

Figure 9.5 shows the functionalities provided by GUI variant T3. In contrast to T2, only a qualitative preference dialog is available instead of the manual weight setting dialog. The result list is displayed in matrix form. T3’s main purpose is to evaluate the usability of the preference approach proposed in Chapter 5.

As in T2, the browsing and textual query functionality was disabled during the test.

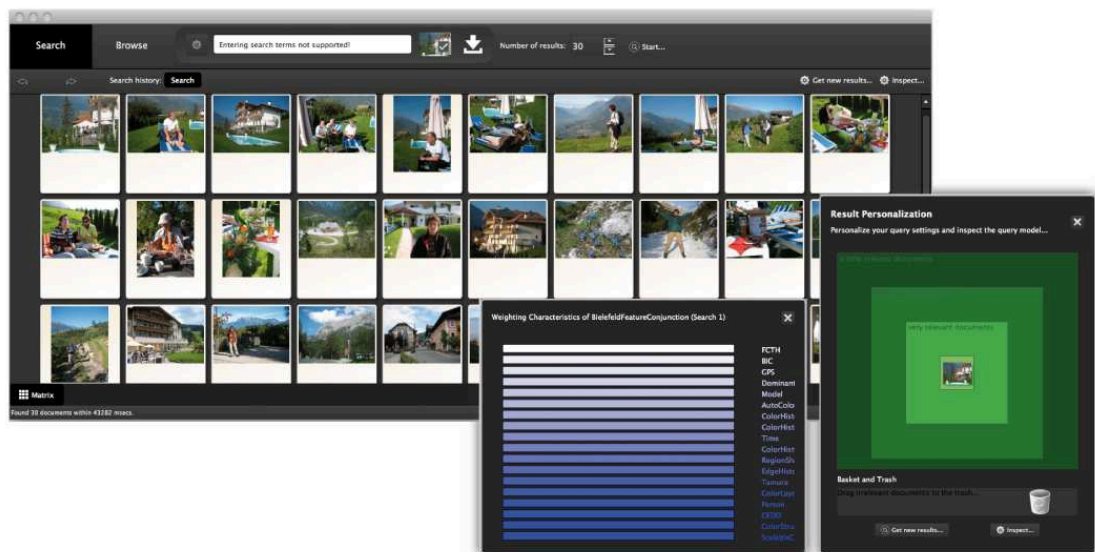


Figure 9.5: Available widgets in GUI variant T3

### 9.1.5 GUI Variant T4

T4 is the most powerful GUI variant of the three examined ones. Figure 9.6 depicts all widgets and the results in SOM visualization mode. In addition, the results can be visualized either using a matrix or a k-medoid clustering mechanism.

The GUI features the same functionalities as T2 and T3. Furthermore, it supports a number of pre-defined facets, e.g., to search for images that display persons or that share the same location as the query image(s). This GUI variant equals the Pythia MIR system presented in Chapter 6.

As before, browsing and textual queries were not available to the subjects.

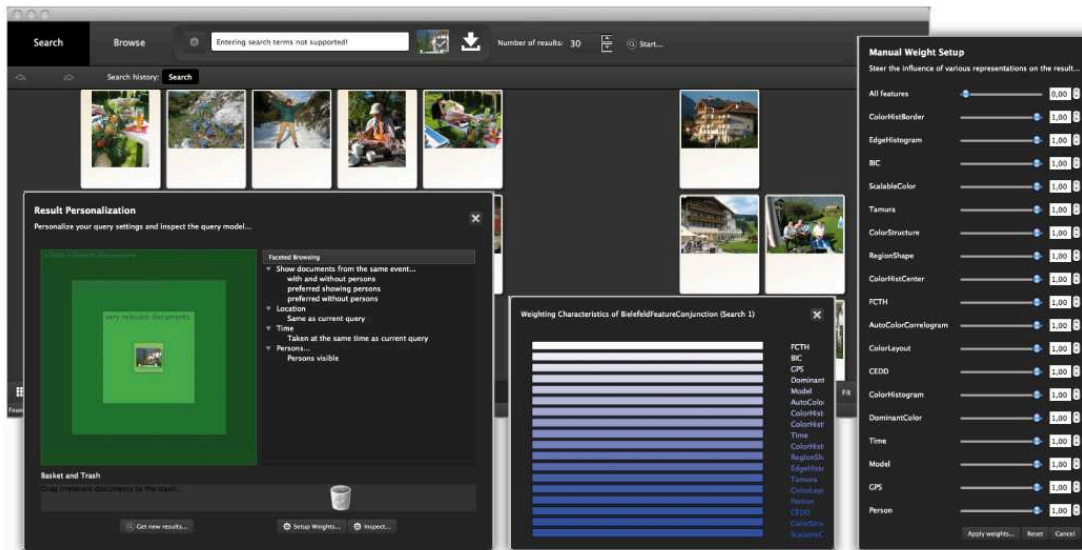


Figure 9.6: Available widgets in GUI variant T4

### 9.1.6 Summary

Generally speaking, all GUI variants share the same look and feel. That is, all central GUI elements are placed at the same positions and support the same interactive mechanisms such as drag and drop or double clicks. Because of the introductory tutorial and the initial confrontation phase of the users with the MIR system, one can assume a comparable level of familiarity with the software amongst all participants.

Additionally, the study participants had no means to alter the structure of the underlying CQQL query. All subjects worked with the same document collection and the same initial query image. Hence, one major factor that can affect the user experience – the effectiveness of the MIR system – was more or less controlled during the usability study.

Although effects of hidden variables, such as a changing emotional approach towards the study and tiring or learning effects, cannot be fully eliminated in a usability study, the study design aims at limiting the effect of such variables by following the aforementioned principles. In consequence, the major variables that have an impact on the perceived, subjective usability of each GUI variant are most likely related to the actual (interaction) design of the GUI and its functionality.

To recapitulate, Table 9.1 compares the functionality of each GUI variant to their counterparts. It lists the supported information seeking strategies (see Section 3.3) and preference approaches (see Section 5.3) in addition to the supporting search input widgets. The table also displays the available result visualizations in each GUI variant. For the sake of completeness, assistance tools such as the breadcrumb history that allows a chronological navigation through the user's search session are also registered.

Table 9.1: Functional comparison of the GUI variants

GUI Variant	Information Seeking Strategies & Preference Approach	Search Input	Result Visualization	Assistance Tools
T2	Query by Example Quant. Preferences	Query by Example Manual Weights	Matrix Inspector	Breadcrumb History Result Limiter
T3	Query by Example Qualit. Preferences	Query by Example Preferences	Matrix Inspector	Breadcrumb History Result Limiter Trash Bin
T4	Query by Example Quant. Preferences Qualit. Preferences Faceted Navigation	Query by Example Manual Weights Preferences Facets	Matrix SOM Cluster Inspector	Breadcrumb History Result Limiter Trash Bin

## 9.2 Demographics and Computer Usage of the Test Persons

In total, 59 subjects were recruited to take part in the main study, 30 of which were female. As suggested by Preece et al. [2002], 6 additional subjects took part in a pre-study that aimed at revealing ambiguous formulations in the instructions and problems in the workflow of the usability test. The data of these subjects has been removed from the data discussed in this dissertation. However, for the sake of completeness, it is contained as a supplement (see Appendix E). All test persons completed a questionnaire with 17 questions asking about their demographics, computer usage, and preferences regarding the examined systems. The first two aspects are summarized in this section, while Section 9.4 discusses the other aspects. The complete questionnaire can be found in Appendix C.3.1, while the full results are available in the supplement (see Appendix E).

Although not all users, i.e., 30.5% of the subjects, are familiar with Mac OS X, they can be considered computer-savvy with respect to their daily computer usage (see Figure 9.7; left).

As shown in Figure 9.7 (right), the majority of the study participants are students. This is also reflected by the distribution of the years of birth (see Figure 9.8; left). Only seven students took a class in IR, and only four visited MMIR. Approximately one third (33.90%) of the participants has a background related to computer science (see Table 9.2). Nevertheless, some subjects (23.73%) stated that they have little knowledge of CBIR (see Figure 9.8). Most test persons (71.19%) have no knowledge of the principles of CBIR. As Figure 9.9 illustrates, this fact does not prevent the participants from using CBIR systems, predominantly Google's image search. Only two subjects have been confronted with the examined system before. One person had only used the system once prior to the user study, while the other subject had used it approximately five times.

To conclude, the test persons are mostly computer-savvy laypersons regarding the scientific fields related to this thesis. Regarding their studies or field of work, only one third can be directly related to computer science. The largest homogeneous group forming one quarter (27.12%) has a background in the humanities. This group is closely followed by information and media technologists (18.64%).

## 9.2 Demographics and Computer Usage of the Test Persons

Table 9.2: Study and work field of the subjects; computer science-related fields are emphasized

Study Program / Field of Work	Percentage
<i>Business Administration and Engineering (Computer Science)</i>	3.39%
<i>Information and Media Technology</i>	18.64%
<i>Computer Science</i>	3.39%
<i>eBusiness</i>	8.47%
Land Use and Water Management	3.39%
Physiotherapy	1.69%
Electrical Engineering	3.39%
Mechanical Engineering	1.69%
Culture and Technology	27.12%
Baker	1.69%
Process Engineering	5.08%
Social Work	3.39%
Business Administration and Engineering (Production)	1.69%
Urban and Regional Planning	3.39%
Business Administration	1.69%
Architecture	8.47%
Not available	3.39%

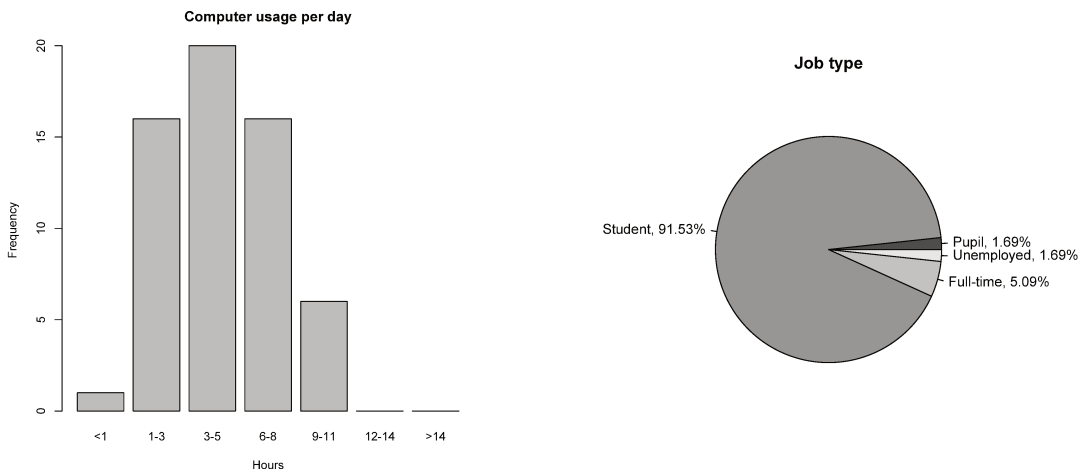


Figure 9.7: Daily computer usage and job type of the subjects

## 9 Evaluation of the User Experience

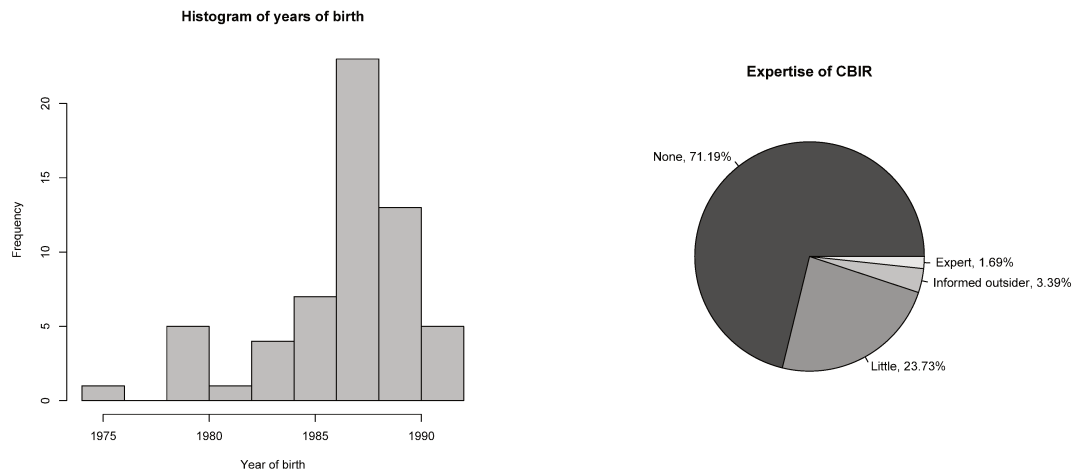


Figure 9.8: Year of birth and expertise of CBIR of the subjects

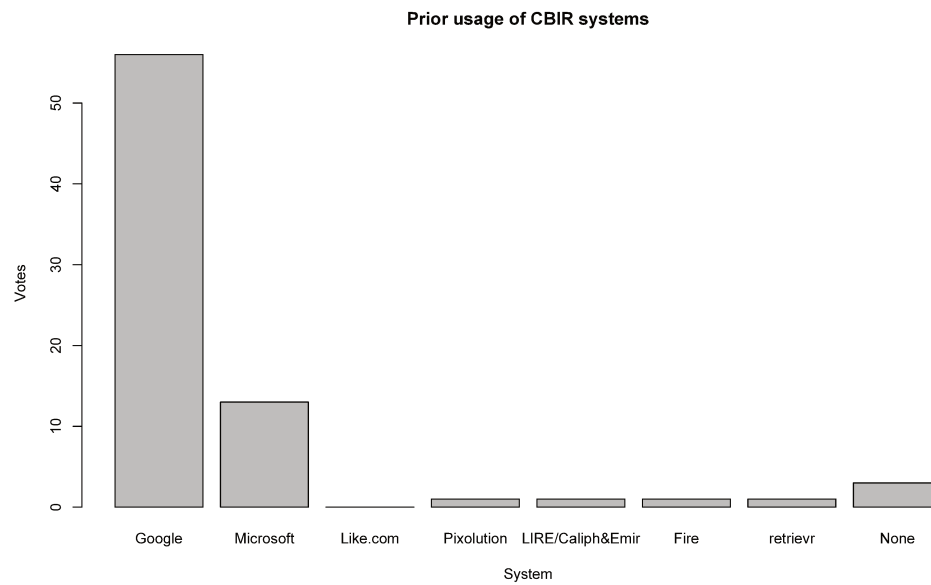


Figure 9.9: Usage of other CBIR systems

### 9.3 Results of the Quantitative Usability Study

In total, the 59 participants of the user study had to answer 35 questions addressing five different general usability goals. These usability goals – or dialogue principles – are defined in the ISO 9241/110 norms as general ergonomic principles that should apply to a usable human-computer interface. For each principle, the subjects had to answer five questions in form of 7-level Likert items. The layout and question order of the used questionnaire is based on a paper-based design by Jochen Prümper and Michael Anft and was transferred to an electronic form by the author of this text<sup>155</sup>. For a complete overview of the questions, see Appendix B.4.

The five main usability goals addressed in the questionnaire are:

1. *suitability for the task*, i.e., whether the examined tool is usable for solving a user's work task;
2. *self-descriptiveness*, i.e., whether the application's functionality and terminology can be understood without extensive extrinsic assistance;
3. *controllability*, i.e., whether the software is easy to control;
4. *conformity with user expectations*, i.e., whether the application's reactions on the user's actions are predictable;
5. *error tolerance*, i.e., the grade to which erroneous user input is tolerated and can be corrected without losing the current working progress;
6. *suitability for individualization*, i.e., how well the application can be adapted towards the user's individual way of solving a task; and
7. *suitability for learning*, i.e., how much time it takes to learn to use the software and how much memory load is caused by the learning.

Generally speaking, these principles coincide with the general usability goals suggested by Preece et al. [2002, p. 14], Shneiderman & Plaisant [2005, Ch. 2], or other specialized usability literature such as Nielsen [2009] (see Section 6.1.4). The primary objective of a user study in the described form is to obtain quantitative data of a system's usability. In particular, this study design aims at revealing specific usability issues. In accordance with other authors, Cooper et al. suggest to conduct such studies after the development of an interaction mechanism or different design variants in order to validate their effectiveness [Cooper et al. 2007, cf. p. 71]. Comparable survey designs are presented, e.g., in Preece et al. [2002, Sec. 13.3] or Shneiderman & Plaisant [2005, Sec. 4.4].

Each participant could interact with each GUI variant for 15 minutes at most in order to get an impression of its usability while solving the task described in Section 9.1.

<sup>155</sup>The original version is no longer available. An updated version that mainly differs from the used one in its layout can be obtained from <http://people.f3.htw-berlin.de/Professoren/Pruemper/instrumente.html> (October 29th, 2012)

## 9 Evaluation of the User Experience

Table 9.3 shows the actual averaged interaction duration. Roughly speaking, the test persons spent almost the same time with each GUI variant. Hence, it can be assumed that the judgments of the subjects are comparable because each participant had roughly the same confrontation time with each GUI.

Table 9.3: Interaction duration with the GUI variants

GUI Variant	Mean Duration [min]	Std. Deviation Duration [min]	Median Duration [min]
T2	9.24	4.49	9.11
T3	9.63	3.65	9.73
T4	9.62	4.05	8.68

Maximum allowed time was 15 minutes.

Table 9.4 compares the Likert scales of each GUI variant, i.e., the summation of all Likert items per variant grouped by their usability goal. All histograms display the level of agreement or disagreement with the given statement on the left or right side of the graph in form of 7-level Likert items. The statements on the right side of each histogram are desirable from a usability perspective. In other words, the more a histogram is left-skewed (i.e., it leans to the right), the better the examined system is regarding a usability principle. If the histogram has no or little skew, the subjects were undetermined about the statements.

For the sake of brevity, mostly the consolidated results of the questionnaire for each usability goal are presented in this part of the thesis. Only outstanding results are discussed separately. The evaluation of the Likert items for each GUI variant are available in Appendices B.4.1ff.

**Suitability for the task:** Regarding the suitability for the task, the test persons consider T4 to be the best system, followed by T3. The opinions about T2 are undetermined with a slight trend towards negativity.

**Self-descriptiveness:** T2 is not considered to be self-descriptive by most subjects, in addition to a large undecided group. The opinions about T3 and T4 are undetermined with a slightly positive trend, whereas T3 gets a higher amount of fully positive agreement than T4. This effect is mainly due to the fact that T3 is using more comprehensible terms and symbols than T4 (see Appendices B.4.2 and B.4.3; 2b).

**Controllability:** The opinions about the controllability of T2 and T3 are not very clear. Both show a vaguely positive trend, although the group of undetermined subjects is large. Generally speaking, T4 is considered easier to control in comparison to T2 and T3, although only a few people fully agree with the positive statement. The ease of control of T4 is attributed to the more flexible and adaptable GUI that does not force a certain interactive workflow on the user (see Appendices B.4.2 and B.4.3; 3a,b,d,e). This flexibility comes at the price of a more complex GUI (see Appendix B.4.3; 3c).



### 9.3 Results of the Quantitative Usability Study

**Conformity with user expectations:** T2 conforms with user expectations slightly less than the two other GUI variants. The histogram of T3 and T4 are very similar. In direct comparison, T3 gets a higher level of agreement to the positive statement than T4. This effect is caused by T3's better way of communicating whether an input was successful or not (see Appendices B.4.2 and B.4.3; 4b).

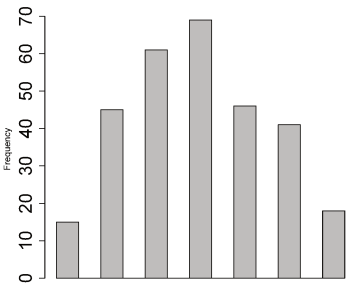
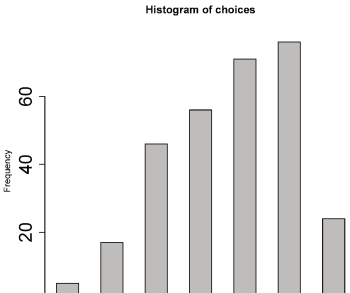
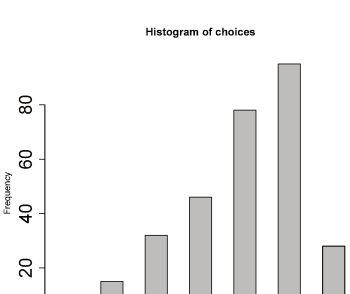
**Error tolerance:** The subjects are very undecided about the error tolerance of all GUI variants. Nevertheless, the feedback about T3 and T4 is slightly better than T2's results, while T4 is considered the most error tolerant.

**Suitability for individualization:** In general, with respect to their suitability for individualization, all GUI variants share a large group of undetermined test persons. While T2 has a negative trend to a certain extent, T3 and T4 are perceived positively.

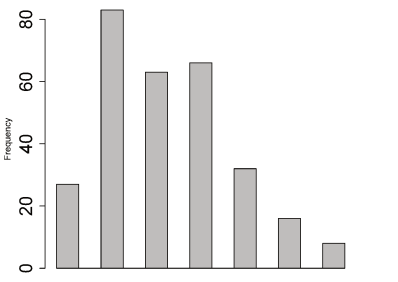
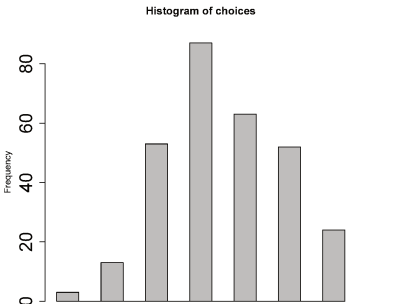
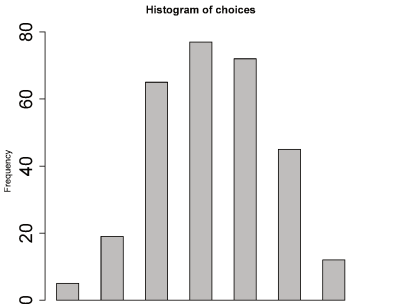
**Suitability for learning:** T4 is regarded as easy to learn. Although T4 gets a high amount of positive agreement, it is obvious that only a small group of subjects fully agrees that T4 is easy to learn. T3 roughly follows the same trend but with a less total amount of positive feedback. The situation for T2 is not very determined, although there is a slight positive trend.

## 9 Evaluation of the User Experience

Table 9.4: General usability of the GUI variants

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale <i>From top to bottom: T2, T3, T4</i>	The software... (good)
is not suitable for the task.	<p style="text-align: center;">Histogram of choices</p>  <p style="text-align: center;">Histogram of choices</p>  <p style="text-align: center;">Histogram of choices</p> 	is suitable for the task.

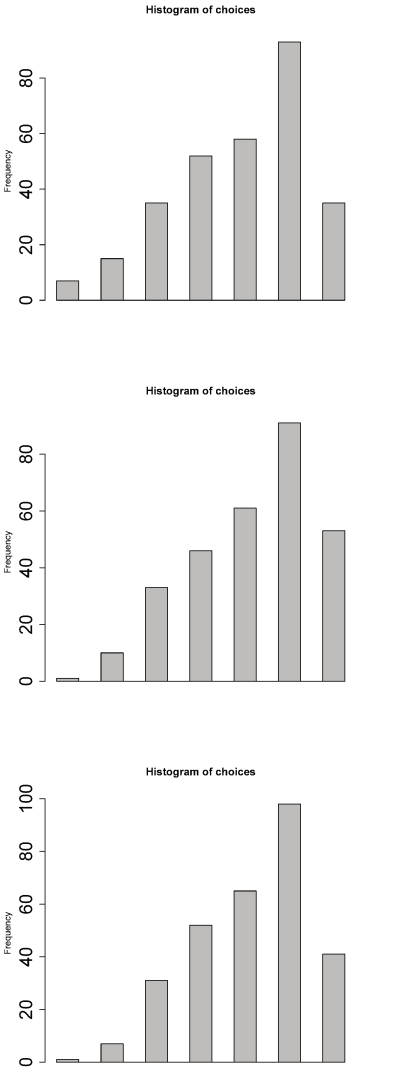
### 9.3 Results of the Quantitative Usability Study

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale <i>From top to bottom: T2, T3, T4</i>	The software... (good)																																																
is not self-descriptive.	<p style="text-align: center;">Histogram of choices</p>  <table border="1" data-bbox="518 347 917 627"> <caption>Frequency of Agreement Choices for 'is not self-descriptive'</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>28</td></tr> <tr><td>2</td><td>82</td></tr> <tr><td>3</td><td>63</td></tr> <tr><td>4</td><td>67</td></tr> <tr><td>5</td><td>32</td></tr> <tr><td>6</td><td>16</td></tr> <tr><td>7</td><td>8</td></tr> </tbody> </table> <p style="text-align: center;">Histogram of choices</p>  <table border="1" data-bbox="518 694 917 996"> <caption>Frequency of Agreement Choices for 'is self-descriptive'</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>3</td></tr> <tr><td>2</td><td>13</td></tr> <tr><td>3</td><td>53</td></tr> <tr><td>4</td><td>85</td></tr> <tr><td>5</td><td>63</td></tr> <tr><td>6</td><td>52</td></tr> <tr><td>7</td><td>24</td></tr> </tbody> </table> <p style="text-align: center;">Histogram of choices</p>  <table border="1" data-bbox="518 1075 917 1377"> <caption>Frequency of Agreement Choices for 'is self-descriptive'</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>5</td></tr> <tr><td>2</td><td>19</td></tr> <tr><td>3</td><td>65</td></tr> <tr><td>4</td><td>76</td></tr> <tr><td>5</td><td>71</td></tr> <tr><td>6</td><td>45</td></tr> <tr><td>7</td><td>12</td></tr> </tbody> </table>	Level	Frequency	1	28	2	82	3	63	4	67	5	32	6	16	7	8	Level	Frequency	1	3	2	13	3	53	4	85	5	63	6	52	7	24	Level	Frequency	1	5	2	19	3	65	4	76	5	71	6	45	7	12	is self-descriptive.
Level	Frequency																																																	
1	28																																																	
2	82																																																	
3	63																																																	
4	67																																																	
5	32																																																	
6	16																																																	
7	8																																																	
Level	Frequency																																																	
1	3																																																	
2	13																																																	
3	53																																																	
4	85																																																	
5	63																																																	
6	52																																																	
7	24																																																	
Level	Frequency																																																	
1	5																																																	
2	19																																																	
3	65																																																	
4	76																																																	
5	71																																																	
6	45																																																	
7	12																																																	

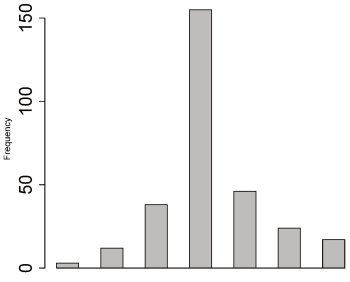
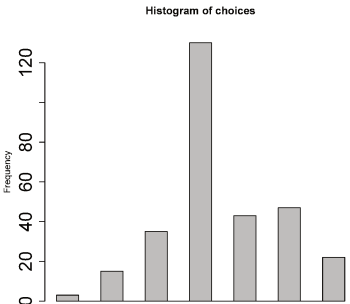
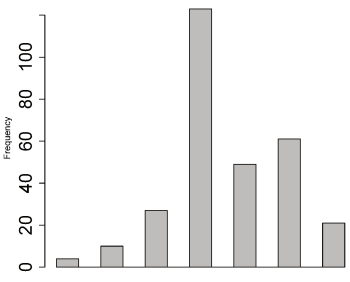
## 9 Evaluation of the User Experience

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale <i>From top to bottom: T2, T3, T4</i>	The software... (good)																
is not easy to control.	<p style="text-align: center;">Histogram of choices</p> <p>This histogram shows the frequency of agreement choices for the statement 'is not easy to control.' across seven Likert scale points. The y-axis represents frequency from 0 to 80. The distribution is roughly bell-shaped, peaking at the 4th point (frequency ~90).</p> <table border="1"> <thead> <tr> <th>Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>5</td></tr> <tr><td>2</td><td>10</td></tr> <tr><td>3</td><td>25</td></tr> <tr><td>4</td><td>90</td></tr> <tr><td>5</td><td>65</td></tr> <tr><td>6</td><td>80</td></tr> <tr><td>7</td><td>20</td></tr> </tbody> </table>	Choice	Frequency	1	5	2	10	3	25	4	90	5	65	6	80	7	20	is easy to control.
	Choice	Frequency																
	1	5																
2	10																	
3	25																	
4	90																	
5	65																	
6	80																	
7	20																	
<p style="text-align: center;">Histogram of choices</p> <p>This histogram shows the frequency of agreement choices for the statement 'is not easy to control.' across seven Likert scale points. The y-axis represents frequency from 0 to 80. The distribution peaks at the 5th point (frequency ~78).</p> <table border="1"> <thead> <tr> <th>Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>10</td></tr> <tr><td>3</td><td>30</td></tr> <tr><td>4</td><td>70</td></tr> <tr><td>5</td><td>78</td></tr> <tr><td>6</td><td>65</td></tr> <tr><td>7</td><td>45</td></tr> </tbody> </table>	Choice	Frequency	1	0	2	10	3	30	4	70	5	78	6	65	7	45		
Choice	Frequency																	
1	0																	
2	10																	
3	30																	
4	70																	
5	78																	
6	65																	
7	45																	
<p style="text-align: center;">Histogram of choices</p> <p>This histogram shows the frequency of agreement choices for the statement 'is not easy to control.' across seven Likert scale points. The y-axis represents frequency from 0 to 100. The distribution peaks at the 6th point (frequency ~100).</p> <table border="1"> <thead> <tr> <th>Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>5</td></tr> <tr><td>2</td><td>10</td></tr> <tr><td>3</td><td>20</td></tr> <tr><td>4</td><td>60</td></tr> <tr><td>5</td><td>65</td></tr> <tr><td>6</td><td>100</td></tr> <tr><td>7</td><td>35</td></tr> </tbody> </table>	Choice	Frequency	1	5	2	10	3	20	4	60	5	65	6	100	7	35		
Choice	Frequency																	
1	5																	
2	10																	
3	20																	
4	60																	
5	65																	
6	100																	
7	35																	

### 9.3 Results of the Quantitative Usability Study

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale <i>From top to bottom: T2, T3, T4</i>	The software... (good)
<p>is not conform with user expectations.</p>	 <p>The figure consists of three histograms, each titled 'Histogram of choices', representing data for time points T2, T3, and T4. The y-axis for all histograms is 'Frequency'.</p> <ul style="list-style-type: none"> <li><b>T2 (top histogram):</b> The y-axis ranges from 0 to 80. The distribution is skewed to the right, with the highest frequency (approx. 90) at choice 6, and a secondary peak at choice 7 (approx. 35).</li> <li><b>T3 (middle histogram):</b> The y-axis ranges from 0 to 80. The distribution is skewed to the right, with the highest frequency (approx. 90) at choice 6, and a secondary peak at choice 7 (approx. 50).</li> <li><b>T4 (bottom histogram):</b> The y-axis ranges from 0 to 100. The distribution is skewed to the right, with the highest frequency (approx. 95) at choice 6, and a secondary peak at choice 7 (approx. 40).</li> </ul>	<p>is conform with user expectations.</p>

## 9 Evaluation of the User Experience

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale <i>From top to bottom: T2, T3, T4</i>	The software... (good)																																																
is not error tolerant.	<p style="text-align: center;">Histogram of choices</p>  <table border="1" data-bbox="678 347 1029 627"> <caption>Frequency of Agreement Choices for 'is not error tolerant.'</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>5</td></tr> <tr><td>2</td><td>15</td></tr> <tr><td>3</td><td>40</td></tr> <tr><td>4</td><td>155</td></tr> <tr><td>5</td><td>45</td></tr> <tr><td>6</td><td>25</td></tr> <tr><td>7</td><td>20</td></tr> </tbody> </table> <p style="text-align: center;">Histogram of choices</p>  <table border="1" data-bbox="678 694 1029 996"> <caption>Frequency of Agreement Choices for 'is error tolerant.'</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>5</td></tr> <tr><td>2</td><td>15</td></tr> <tr><td>3</td><td>35</td></tr> <tr><td>4</td><td>125</td></tr> <tr><td>5</td><td>45</td></tr> <tr><td>6</td><td>45</td></tr> <tr><td>7</td><td>25</td></tr> </tbody> </table> <p style="text-align: center;">Histogram of choices</p>  <table border="1" data-bbox="678 1108 1029 1388"> <caption>Frequency of Agreement Choices for 'is error tolerant.'</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>5</td></tr> <tr><td>2</td><td>10</td></tr> <tr><td>3</td><td>25</td></tr> <tr><td>4</td><td>110</td></tr> <tr><td>5</td><td>50</td></tr> <tr><td>6</td><td>60</td></tr> <tr><td>7</td><td>20</td></tr> </tbody> </table>	Level	Frequency	1	5	2	15	3	40	4	155	5	45	6	25	7	20	Level	Frequency	1	5	2	15	3	35	4	125	5	45	6	45	7	25	Level	Frequency	1	5	2	10	3	25	4	110	5	50	6	60	7	20	is error tolerant.
Level	Frequency																																																	
1	5																																																	
2	15																																																	
3	40																																																	
4	155																																																	
5	45																																																	
6	25																																																	
7	20																																																	
Level	Frequency																																																	
1	5																																																	
2	15																																																	
3	35																																																	
4	125																																																	
5	45																																																	
6	45																																																	
7	25																																																	
Level	Frequency																																																	
1	5																																																	
2	10																																																	
3	25																																																	
4	110																																																	
5	50																																																	
6	60																																																	
7	20																																																	

### 9.3 Results of the Quantitative Usability Study

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale <i>From top to bottom: T2, T3, T4</i>	The software... (good)
cannot be individualized.	<p><b>Top Histogram (T2):</b> The y-axis ranges from 0 to 100. The distribution peaks at the 4th category with a frequency of approximately 100.</p> <p><b>Middle Histogram (T3):</b> The y-axis ranges from 0 to 80. The distribution peaks at the 4th category with a frequency of approximately 90.</p> <p><b>Bottom Histogram (T4):</b> The y-axis ranges from 0 to 80. The distribution peaks at the 4th category with a frequency of approximately 90.</p>	can be individualized.

## 9 Evaluation of the User Experience

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale <i>From top to bottom: T2, T3, T4</i>	The software... (good)																
is not easy to learn.	<p style="text-align: center;">Histogram of choices</p> <table border="1"> <caption>Frequency of Agreement Choices for 'is not easy to learn.' (T2)</caption> <thead> <tr> <th>Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>12</td></tr> <tr><td>2</td><td>34</td></tr> <tr><td>3</td><td>60</td></tr> <tr><td>4</td><td>39</td></tr> <tr><td>5</td><td>57</td></tr> <tr><td>6</td><td>57</td></tr> <tr><td>7</td><td>36</td></tr> </tbody> </table>	Choice	Frequency	1	12	2	34	3	60	4	39	5	57	6	57	7	36	is easy to learn.
	Choice	Frequency																
	1	12																
2	34																	
3	60																	
4	39																	
5	57																	
6	57																	
7	36																	
<p style="text-align: center;">Histogram of choices</p> <table border="1"> <caption>Frequency of Agreement Choices for 'is not easy to learn.' (T3)</caption> <thead> <tr> <th>Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>2</td></tr> <tr><td>2</td><td>14</td></tr> <tr><td>3</td><td>22</td></tr> <tr><td>4</td><td>42</td></tr> <tr><td>5</td><td>71</td></tr> <tr><td>6</td><td>86</td></tr> <tr><td>7</td><td>57</td></tr> </tbody> </table>	Choice	Frequency	1	2	2	14	3	22	4	42	5	71	6	86	7	57		
Choice	Frequency																	
1	2																	
2	14																	
3	22																	
4	42																	
5	71																	
6	86																	
7	57																	
<p style="text-align: center;">Histogram of choices</p> <table border="1"> <caption>Frequency of Agreement Choices for 'is not easy to learn.' (T4)</caption> <thead> <tr> <th>Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>3</td></tr> <tr><td>2</td><td>8</td></tr> <tr><td>3</td><td>24</td></tr> <tr><td>4</td><td>29</td></tr> <tr><td>5</td><td>92</td></tr> <tr><td>6</td><td>100</td></tr> <tr><td>7</td><td>41</td></tr> </tbody> </table>	Choice	Frequency	1	3	2	8	3	24	4	29	5	92	6	100	7	41		
Choice	Frequency																	
1	3																	
2	8																	
3	24																	
4	29																	
5	92																	
6	100																	
7	41																	



### 9.3.1 Validity of the Results

One intuitive presumption about the validity of a usability study is that a large number of subjects is needed. The advantages of a large test group are obvious: it induces different, subjective points of view of the examined system; it provides demographic variance; and it can potentially discover a large amount of *different* usability problems.

The last point in particular has caused much debate in the scientific usability community because the effectiveness of a usability study is mainly determined by the number of uniquely discovered usability problems per (paid) test person. Over the last decades, the interest in conducting low-cost but effective usability studies has added authority to Nielsen's recommendation to use five users for a usability study.

Nielsen's recommendation is based on the following equation, which declares that the number of found usability problems by  $n$  users is:

$$N(1 - (1 - \lambda)^n), \quad (9.1)$$

where  $\lambda$  is the probability of finding a new usability problem with one person. On average,  $\lambda$  is approximately 31%, which is concluded from several usability studies [Nielsen & Landauer 1993].  $N$  is the total number of usability problems in the system that can be estimated by conducting evaluations of the system. To recapitulate the findings of Nielsen & Landauer [1993], one can assume that five users are likely to discover ca. 85% of the usability issues. This becomes clear if we examine the following example, with  $n = 5$  and  $N = 100$ :

$$100(1 - (1 - 0.31)^5) \approx 84.36 \quad (9.2)$$

Following Nielsen's observations and mathematical formalization, it can be shown that five subjects would discover most usability problems in a system. However, there is an ongoing controversy of conducting usability studies with such a small number of participants [Shneiderman & Plaisant 2005, cf. p. 148], which is also known as the "five users is (not) enough" debate [Schmettow 2012, p. 66]. For instance, Hwang & Salvendy [2010] conclude from a metastudy that  $10 \pm 2$  subjects are necessary to achieve a 80% overall discovery rate of usability problems.

Nielsen & Landauer [1993] and Hwang & Salvendy [2010] have in common that both rely on a "magic number" based on a heuristic to give advice on the design of usability tests. Schmettow [2012] criticizes this approach and doubts that 10 layperson or even expert users are sufficient to discover 80 % of the usability issues [Schmettow 2012, cf. p. 70]. He strongly argues against magic numbers because he considers them to be a poor model for the prediction of the effectiveness of usability studies as they do not take the usability problem visibility into account. Roughly speaking, the aforementioned approaches do not adequately address the re-observance of already discovered (obvious) usability problems. To give a very simple example, the swapping of the mouse buttons against the operating system's conventions would be a usability problem that is likely to be discovered by every test person. Such a repetitive finding does not contribute to a usability study's expressiveness.

## 9 Evaluation of the User Experience

On the whole, it is complicated to definitely predict the number of users required to find every usability problem, but there are heuristics that are used in practice.

Nevertheless, this does not mean that the presented usability studies in this dissertation are invalid. Instead, Schmettow [2012] cites Nørgaard & Hornbæk [2006] who report that (industrial) usability studies are often used to “confirm problems that are already known” [Schmettow 2012, p. 69]. From this point of view, the presented usability study does not strive for completeness regarding the number of discovered usability problems. Instead, as mentioned at the beginning of Chapter 9, it constitutes an initial user study during the user-centered development process. As such, it can only evaluate the usability of the prototype at a given point in time. A dissertation, by definition, cannot present an end-user ready software system, which is the typical usage domain of usability engineering. During a dissertation project one has neither access to the financial, personnel, nor the technical-methodical resources needed to carry out all usability tests that are state of the art in an industrial context. Hence, the presented studies have to take this conflict into account.

To discover a fair amount of usability problems of the prototypical Pythia MIR system, this dissertation presented a comprehensive, quantitative user study relying on 59 subjects at the beginning of this section that is extended with qualitative parts (see Section 9.4.1). In addition, five user observations are discussed in Section 9.5 in order to present a relatively complete picture of the usability problems of the Pythia MIR system. As mentioned in Section 6.1.3, the informal user studies and the expert workshops contribute their part by providing valuable information about the user needs and usability expectations.

In any case, one could object to the setup of this usability study: it has been conducted in a highly artificial environment, its work task is simulated, and the subjects sit in a computer lab during the study – although most are used to this kind of location. This setup contrasts to Cooper et al. [2007] who advise to conduct user studies in the form of *contextual inquiries*, i.e., users are interviewed and observed in their daily work environment. While this approach is certainly feasible for an industrial software development process, the needed resources (time, money, personnel) were not available to the author of this dissertation.

To come to an end, the presented results and conclusions have to be interpreted in the context of a controlled experiment in a laboratory environment. Moreover, the results are biased by the subjects’ backgrounds. Although some subjects from outside Academia could be recruited, over 91% are students (see Figure 9.7). Though, the quantitative results that have been presented before can serve as a baseline to measure the success of future changes to the Pythia MIR system [Preece et al. 2002, cf. p. 344]. The same holds true for the qualitative parts of the user study presented in the next section, which can be used to confirm the conclusions drawn on basis of the quantitative study.

## 9.4 Results of the Qualitative User Study

This section continues the discussion on the questionnaire presented in Section 9.2. That is, the results of the qualitative questions 10-17 which are available in Appendix C.3.1 and the open comment question.

When the test persons are asked directly after their preference between the three GUI variants (see Appendix C.3.1; 17), T4 is favored by most users. Figure 9.10 shows the histogram of the votes for a top-3 ordering of the GUI variants (the  $x$ -axis denotes the GUI variant, the  $y$ -axis shows the number of votes). Most subjects place T3 as second and like T2 as last.

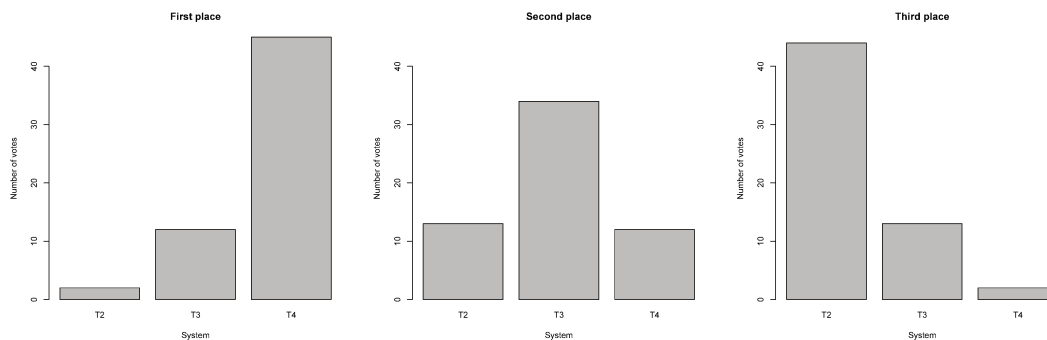


Figure 9.10: Ranking of the tested GUI variants; 1st to 3rd place from left to right

As described in Section 9.1.6, the GUI variants differ mostly by their GUI components. Figure 9.11 displays the general utility of different GUI components as judged by the subjects (see Appendix C.3.1; 10-16). The standard box plot<sup>156</sup> uses the German grading scale<sup>157</sup> to illustrate how the subjects judged the utility of the search input and result visualization widgets in isolation.

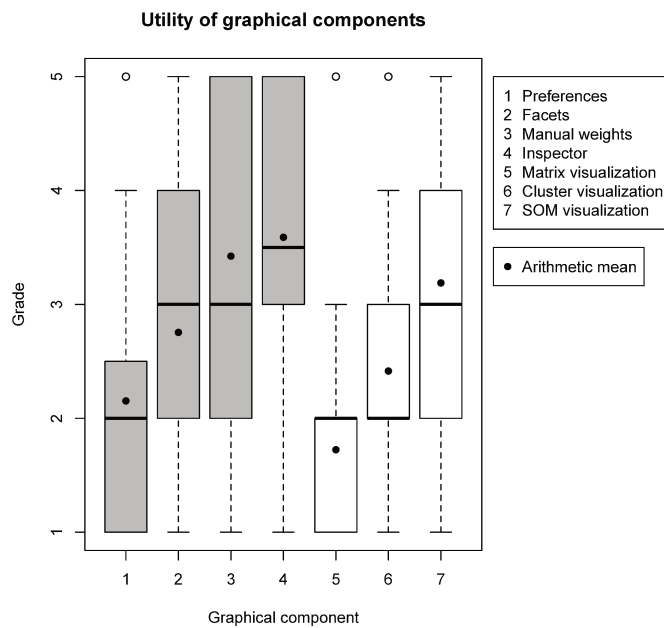
Regarding the input mechanisms, the preference rings widget (1) has received the best rating. On average, it is rated good. The faceted navigation (2) is perceived as mediocre by the test persons but clearly better than the manual weights widget. In any case, the judgements are widely spread. The manual weight setting dialog (3) widget is rated satisfactory with a trend towards a sufficient performance. Interestingly, the variance of the manual weights widget's grade is rather high with respect to the comparable variance of the grades for the other input widgets (see Figure 9.11, shaded boxes). The weight inspector (4) is not considered as very helpful by most test persons.

From the visualization methods, the subjects clearly prefer the matrix visualization. While the cluster visualization is still graded good, the utility of the SOM is mostly perceived satisfactory. Interestingly, the picture here is not as clear as with the other visualization variants. There are study participants that clearly like the SOM visualization, while other see it as nearly useless.

<sup>156</sup>For further information about the interpretation of box plots, see page 246.

<sup>157</sup>1= very good, 2= good, 3= satisfactory, 4= sufficient, 5= deficient

## 9 Evaluation of the User Experience



Gray boxes indicate input widgets.  
Horizontal bands indicate the median, solid circles the mean. Empty circles indicate outliers.

Figure 9.11: General utility of the different GUI components; 1 indicates best and 5 worst grade

### 9.4.1 Results of the Open Question

In addition to the closed questions asking for the subjective preferences of the subjects, the participants of the study had the chance to provide a textual comment on the systems in form of an open question. Approximately 62.71% of the test persons, i.e., 37 out of 59, used this opportunity to give additional feedback. Table 9.5 summarizes the findings of the evaluation of this open question grouped by their area of criticism. The full text is available as a supplement (see Appendix E). Most comments are neutral, although there are eight distinctly positive and three negative ones. Being highly subjective, these comments range from “[The system is] is rather needless. [...] Its use is not noticeable, working is not getting more effective.”<sup>158</sup> (subject 36) to “The software is a great tool for the targeted image search after motifs.”<sup>159</sup> (subject 49). The negative judgements are reflected in the outliers shown in Figure 9.11.

On a more objective level, approximately 70% of the subjects complain about the general controllability of the GUI variants. Most of them miss informative explanations,

<sup>158</sup>Translation from the original comment in German: “Eher überflüssig [...] Der Sinn ist nicht erkenntlich [sic!], die Arbeit wird nicht effektiver.”

<sup>159</sup>Translation from the original comment in German: “Die Software ist ein großartiges Werkzeug für die gezielte Bilder suche [sic!] nach Motiven.”

e.g., more descriptive labels for the GUI widgets or an online help. In particular, the naming of the representations in the manual weight setup dialog are incomprehensible to a large amount of the subjects. For example, one test person stresses that “the slider was completely incomprehensible, I could not understand which settings I would control with it”<sup>160</sup>. Furthermore, a “more specific description of the relation and impact of the various search parameters”<sup>161</sup> is requested.

To a lesser extent, the subjects are unable to cope with the complexity of the GUI variants T2 and T4. The following comments from subject 8 illustrate the most common points of criticism: “At all costs, T2 needs explanations about the control options.”<sup>162</sup> and “T4 is designed in a very bulky way. Sometimes, less is more.”<sup>163</sup>.

Only one subject had problems to revert carried out operations in case of an error.

Surprisingly, only six subjects complained specifically about the retrieval quality of the system. Two of these test persons delimited their criticism to the retrieval effectiveness of the faceted navigation whereas subject 44 emphasized a weakness of the person presence facet: “If I want to exclude images depicting persons [from the results], this should be simple.”<sup>164</sup>.

Although the subjects were explicitly asked not to judge the loading time of the GUI variants, three subjects name this as a problem.

Table 9.5: General user criticism

Category	Total	Specialization	Total
General Comments	37	Positive	8
		Neutral	26
		Negative	3
General Controllability	26	Missing explanations	19
		High complexity	3
		Non-self-explanatory	3
		Error correction	1
Retrieval quality	6	General	4
		Facets	2

## 9.5 Supplementary User Observations

To learn about specific usability problems [Shneiderman & Plaisant 2005, cf. pp. 147], four subjects<sup>165</sup> were chosen randomly, and their interactions with all three GUI vari-

<sup>160</sup>Translation from the original comment in German: “[...] Schieberegler war komplett unverständlich, mir war nicht klar, welche genauen Einstellungen ich dabei variere.”

<sup>161</sup>Translation from the original comment in German: “[...] genauere Beschreibung von Beziehung und Wirkung der einzelnen Suchparameter [...]”

<sup>162</sup>Translation from the original comment in German: “Bei T2 muss unbedingt eine Erklärung zu den Regelmöglichkeiten dazu.”

<sup>163</sup>Translation from the original comment in German: “T4 ist sehr umfangreich gestaltet. Weniger ist manchmal mehr.”

<sup>164</sup>Translation from the original comment in German: “Wenn ich Bilder mit Personen ausschließen will sollte das doch einfacher gehen.”

<sup>165</sup>That is, “j” of the pre-study and subjects 6, 30, and 58 of the main study.

## 9 Evaluation of the User Experience

ants were recorded using a screen grabber software. Additionally, the interaction with T4 of one randomly picked user<sup>166</sup> was recorded. None of the test persons were aware of the recording to avoid distortion effects due to the recording. The objective to avoid such effects made the creation of think-aloud protocols also impossible. The resulting soundless user observations are available as a supplement to this text (see Appendix E), the latter subject's video is available as the sample user interaction.

Following Nielsen's heuristic and the arguments presented in Section 9.3.1, the sample size of five is supposed to show and confirm the majority of the predominant usability problems of the Pythia MIR system under laboratory conditions. Further analyses were not possible due to the limited resources<sup>167</sup>.

In the following, the observations of usability problems made with the recorded subjects are summarized and presented in the order of their occurrence. The numbers given in parentheses next to the GUI elements refer to Figure 6.20 on page 193.

The videos show clearly that the test persons made frequent use of the *preference rings* (see Figure Figure 6.20, 11) – and hence *PrefCQQL* – to give relevance feedback. However, the subjects mostly gave positive feedback only by stating preferences. Only two out of five subjects used the trash bin (13) to provide negative feedback. A common erroneous user input consists of placing only two documents on the same ring, although the GUI provides various online help channels with labels or free floating tool tips (see Figure 9.12; 3) aiming at minimizing erroneous input. The resulting error message was usually closed before it was read. Only subject 30 seems to learn from the error message and adjusted his behavior, the remaining users learnt from trial and error. Furthermore, the accentuation of documents that already participate in a preference relation in the *result visualization panel* (7) was seldom noticed. As a consequence, users tried to drag documents again on the preference rings triggering a warning by the system.

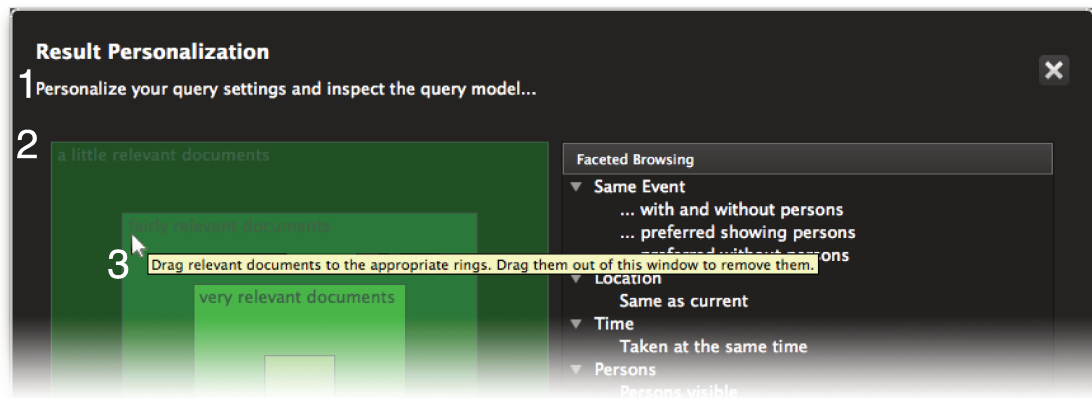


Figure 9.12: Online help information channels in preference rings widget

<sup>166</sup>That is, user 'm' of the pre-study.

<sup>167</sup>For instance, only the sighting of the screen recording of one subject takes approximately 30 minutes (see Table 9.3), fully neglecting the time consumption of the actual usability analysis.

The *manual weight setting dialog* (15) is hardly comprehensible for any test person. Subject 6 manipulated the weight value sliders for the high level representations, such as GPS or time, and relied on a trial and error approach for all other sliders. This interactive pattern is typical for all users. In particular, the function of the “all features” slider was commonly misunderstood as the subjects did not expect it to affect all weights at once.

The *weight inspector* (17) is also incomprehensible to all users. Although it was opened frequently, the users could not link its visualization to the manual weight setting dialog described before.

Users seldom extended or modified the initial query in the *multimedia query documents box* (3), only two subjects discovered this functionality.

Two subjects used the *faceted navigation panel* (12). Subject “j” navigated the facets in a trial and error manner, thus separating his interaction style from subject 58 who picked facets selectively.

Although the subjects tended to use trial and error tactics or browsing to solve the simulated work task, only one subject used the *breadcrumb search history navigation* (6) to undo and redo actions that did not yield the expected result.

Only subject 58 retrieved more than the initially defined 30 result documents with the help of the *search result limiter* (4). The same user opened the *polyrepresentation visualization* (16) but did not link the visualization with her actions.

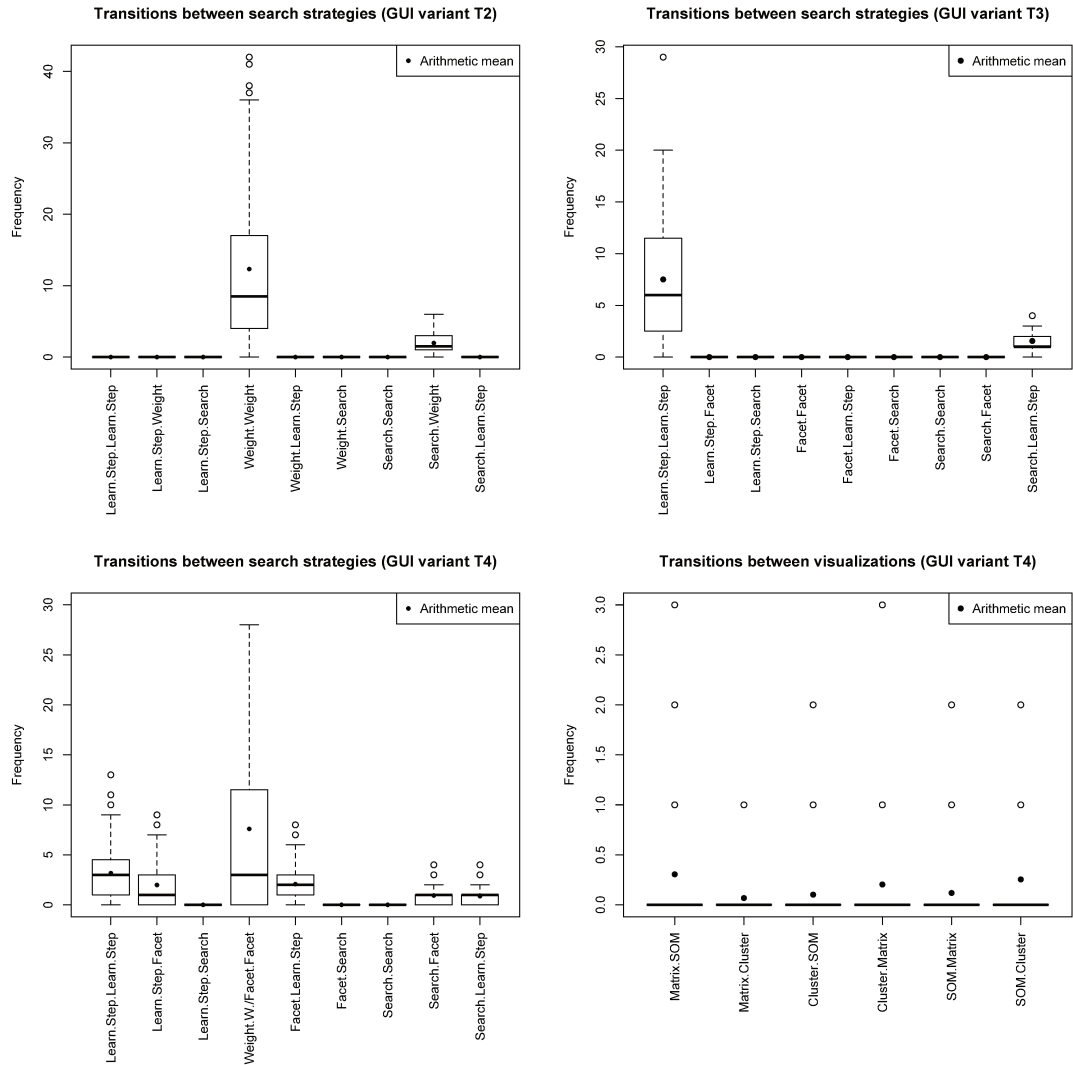
In general, the test persons made heavy use of the provided drag and drop functionality supported by most widgets in the prototype. Documents were dragged from the result visualization panel to the preference rings and trash bin or moved between the preference rings to alter stated preferences. Additionally, from time to time the subjects used the context menu providing the same functionality.

Techniques that support longer search sessions, such as the manual organization of documents in the result visualization panel or tagging were infrequently used. The same holds true for the multiple selection of documents in order to execute an operation on many documents simultaneously (e.g., to associate them with the same relevance level).

Regarding the result visualizations, the subjects clearly preferred the matrix and cluster visualization over the SOM. In fact, two subjects used the cluster visualization continuously during their interaction with the Pythia MIR system prototype. Only subject 58 used the SOM for a short while and dragged some documents from this result visualization in order to modify her preferences. Usually, the users retained one visualization after they had discovered their favorite.

This observation is supported by quantitative results from the user interactions of all 59 subjects of the main group. Figure 9.13 (bottom right) clearly shows that most test persons seldom changed the result visualization type. Table 9.6 serves as the legend for Figure 9.13 that shows the frequency of the search strategy and visualization transitions for each GUI variant. Furthermore, Table 9.6 lists the types of the involved search strategies following the terminology introduced in Section 3.3. The feedback mechanism of PrefCQQL is regarded as a hybrid between directed search and browsing because RF uses a directed search as its basis whose direction is then further optimized with the

## 9 Evaluation of the User Experience



Horizontal bands indicate the median, solid circles the mean. Empty circles indicate outliers. Transitions to a QBE query (.Search) are not possible due to the implementation of the Pythia MIR system.

Figure 9.13: Frequency of the transitions between search strategies and visualizations per GUI variant; see Table 9.6 for the legend



Table 9.6: Legend to Figure 9.13

Label	Search strategy transition from ... to ...
Learn.Step.Learn.Step	PrefCQQL learning → PrefCQQL learning (Browsing/directed search)
Learn.Step.Facet	PrefCQQL learning → facet or manual weight setting (Browsing/directed search → exploratory search)
Facet.Facet	Facet or manual weight setting → facet or manual weight setting (Faceted navigation)
Facet.Learn.Step	Facet or manual weight setting → PrefCQQL learning (Faceted navigation → browsing/directed search)
Search.Facet	QBE query → facet or manual weight setting (Directed search → faceted navigation)
Search.Learn.Step	QBE query → PrefCQQL learning (Directed search → browsing/directed search)
<b>Visualizations</b> (only for GUI variant T4)	
Matrix.SOM	Matrix → SOM
Matrix.Cluster	Matrix → cluster
Cluster.SOM	Cluster → SOM
Cluster.Matrix	Cluster → matrix
SOM.Matrix	SOM → matrix
SOM.Cluster	SOM → cluster

Due to technical reasons there is no discrimination between faceted navigation and manual weight setting. Transitions to a QBE query are not possible due to the implementation of the Pythia MIR system.

help of browsing the close neighborhood of the information space around the initial results (cf. Section 3.3.2).

If all search strategies of the Pythia MIR system are available<sup>168</sup> as in the T4 GUI variant (see Figure 9.13; bottom left), pure faceted navigation (Facet.Facet) was used most often in the examined simulated work task.

Second to the usage frequency of pure faceted navigation is the sole utilization of the PrefCQQL RF mechanism (Learn.Step.Learn.Step). The usage frequency of the transitions between PrefCQQL RF and faceted navigation (Learn.Step.Facet) and vice versa (Facet.Learn.Step) follow shortly after. In other words, if the subjects did not rely on faceted navigation or PrefCQQL RF alone, they alternated between these search strategies.

Generally speaking, new directed searches occurred relatively seldom (Search.Learn.Step and Search.Facet/Weight, respectively). This includes both the extension of the initial query with further QBE documents as well as the full replacement of the initial query – an effect that has also been observed in the screen recordings presented before. Transitions to a directed QBE search from another search strategy (\*.Search) are not possible due to the implementation of the Pythia MIR system. In all of the aforementioned cases, the retrieval engine is reset to the search state. Thus, only transitions from the directed search strategy (Search.\*) to other search strategies can be observed with the given raw data generated by the Pythia MIR system.

<sup>168</sup>This excludes the *k*-medoid clustering-supported exploratory search (see Section 6.3.2), which was disabled during the usability study.

## 9.6 Discussion of the General User Experience

The user studies presented in the last sections identified some usability issues of the Pythia MIR system prototype that become visible when the system is operated by un-experienced first-time users in a laboratory environment.

According to the results of the quantitative user study (see Section 9.3), of the three examined GUI variants T4, i.e., the Pythia MIR system, is regarded as the most suitable for use. It is voted first in four of the seven usability goals while T3 is voted first two times. T2 reaches the last position in all goals. Regarding their suitability for individualization, T3 and T4 hit a stalemate. This finding is reflected by the qualitative study (see Section 9.4) in which the subjects were directly asked for their preference between the three GUI variants.

However, as pointed out in Section 9.4.1, T4 constitutes a complex and feature-rich GUI, which makes it hard to control for some users. On the other hand, the higher complexity of T4 with respect to T2 and T3 is acknowledged by most users that at-test this GUI variant the best suitability for the given simulated work task. Despite its complexity, the subjects regard T4 as relatively easy to learn. This finding is somewhat surprising because this GUI variant integrates all features present in T2 and T3 which are considered as harder to learn. This contradiction becomes more obvious if one considers the self-descriptiveness of all GUI variants.

The study participants criticized all GUI variants for their lack of self-descriptiveness. For instance, many users asked for descriptive labels or an online help, which was unexpected because all GUI elements are labeled with inviting task-related labels (e.g., see Figure 9.12; 1). Furthermore, tool tips with similar pro-active formulations are available for all widgets, e.g., the preference rings (see Figure 9.12; 3). Concluding from the user feedback, the explanations are insufficient or misleading for first-time layperson users. Whether the tool tips were simply overlooked by the subjects cannot be ascertained on the basis of the available data. However, the screen recordings suggest that most users simply ignored tool tips and closed dialogs instantly. Although the provision of an online help fell out of the scope of the prototype's development, the study results show clearly that this factor should not be neglected – even at early development stages. In particular, the participants of the user study demand more specific descriptions of the various search parameters and understandable feedback. Moreover, the incomprehensibility of error messages becomes visible in the user observations, where some users needed several attempts to correct wrong input, e.g., if they place two documents only at one preference ring.

Regarding the controllability, the quantitative results are not very evident. As said in Section 9.3, the subjects consider T4 as easier to control than the other variants. The improved controllability is due to the more flexible and adaptable GUI that offers more functionality than T2 and T3 (see Appendices B.4.2 and B.4.3; 3a,b,d,e). This is no contradiction to the lack of self-descriptiveness of T4 attested at the same time. Instead, it seems that the users perceive the greater choice of tools as more important than their self-descriptiveness. In either case, the results of the open question emphasize that the controllability suffers from the lack of self-descriptiveness of the GUI variants.

The high complexity of T4 is also reflected in its conformity with user expectations. T3 is thought to conform more with the expectations of the subjects. This effect is most likely due to the lower number of functions provided by T3. Essentially, T3 only supports PrefCQQL input with the help of the preference rings. From that perspective, users can easily foresee the reactions of the system, while T4 offers many more functions with different outcomes that obviously cannot be predicted in every case.

Generally speaking, the subjects are very undecided about the error tolerance of all GUI variants. Although all variants offer a breadcrumb search history navigation that helps users to reverse their last actions, this functionality was infrequently used. This finding is also illustrated by the video-based user observations. In particular, this behavior is unexpected because a similar function via the browser's back button is also available in common commercial Web-based image search engines such as Google's image search, which is known to most subjects (see Figure 9.9).

Although all GUI variants share the same framework, T3 and T4 are regarded as more adjustable to the individual needs than T2. Interestingly, the individualization of the MIR system has not been addressed during development. The only means to individualize the GUI are basically the adjustment of the windows' sizes and the window layout – a functionality that is common to all GUI variants. If you will, the different result visualizations can be considered individualization tools, but they are only available in T4 and can therefore not explain the stalemate between T3 and T4 in terms of their suitability for individualization. Hence, it seems reasonable to assume that the negative perception of T2's suitability for individualization is actually due to its low usability in general. In a way, this usability goal of T2 might suffer from a masking effect due to other dominant usability problems.

Because the GUI variants differ mainly in the available widgets, the next part of this section focusses on the specific usability or visibility problems of the central GUI elements of the Pythia MIR system.

Although it is rated good by the test persons (see Figure 9.11), the preference rings widget is not free from usability problems. The widget is commonly used to give positive relevance feedback in form of inductive preferences. The opportunity to give negative RF with the help of the neighboring trash bin is mostly neglected. Whether this is due to the users' incapacibilities to provide negative RF or the unclear metaphor of the trash bin for irrelevant documents cannot be ascertained. Concluding from the erroneous input of placing only two documents on the same ring (see Section 9.5), which results in an uninterpretable preference (the documents are considered equal but no preference is given with respect to all other documents), it becomes obvious that the preference metaphor is ambiguous. This is most likely due to the current implementation of the preference rings that places the current QBE documents in the center (see Figure 9.14; 1) and the other relevant documents around it at different rings (2 and 3). Placing two documents at the same ring leads the subjects to believe that the system will interpret this as the preference "QBE document  $\succ$  all documents on ring X". Unfortunately, this is not the case and eventually causes user irritations. As said before, the corresponding error message is either misunderstood by most users or ignored,

## 9 Evaluation of the User Experience

and should therefore be revised. However, many users can cope with this problem and adjust their interactions appropriately after some failed attempts.

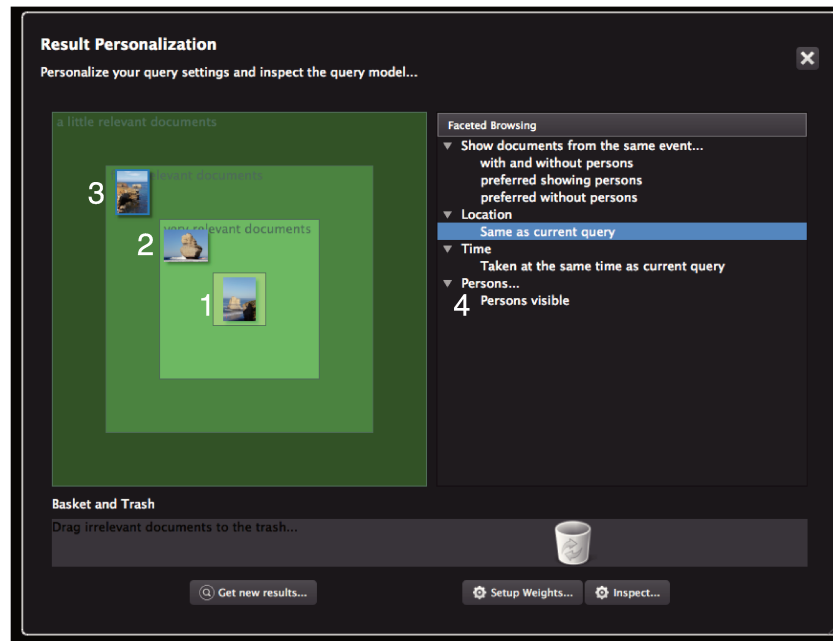


Figure 9.14: Result personalization window of the Pythia MIR system

Furthermore, the accentuation of relevant documents in the result visualization panel is often overlooked. Hence, the relation between documents that already participate in a preference relation has to be made visually much clearer. The currently used check sign icon at the border around the image document often stays unnoticed, which is reflected by superfluous drags from the result visualization panel; this can also be observed in the screen recordings. Although this usability issue does not interrupt the user interaction, it causes unnecessary mouse actions.

Unlike the widgets discussed before, the manual weight setting dialog is suffering from more serious usability issues that render it almost unusable for many subjects. Foremost, the widget suffers from its lack of self-descriptiveness. For instance, if the tool is opened in the screen recordings, it is never used in a manner other than trial and error – if used at all. Roughly speaking, only the weights for high level representations such as GPS are modified in a reasonable manner. Other weights are altered rather randomly by most subjects. This effect is also reflected in Figure 9.11 that shows a high variance in the grading of the manual weight setting dialog. This indicates that some subjects achieve good results when using the widget, while others fail completely. This phenomenon can also be observed in the supplementary screen recordings.

The related weight inspector that is supposed to visualize the impact of the representations on the used matching function is widely considered useless. The user observations show that this inspector is often opened, but it seems that the users cannot link its

visualization to the manual weight setting dialog described before. Furthermore, the users are irritated that the weight inspector visualizes the state of the weights at a given point in the search history and does not update automatically when they proceed with their search. Hence, the weight inspector should update itself according to the weight settings of the current matching function in order to emphasize the relation between the current weights and the user-adjustable weights in the manual weight setting dialog.

Similar problems can be observed with the polyrepresentation visualization. For instance, subject “j” opened the visualization but could not draw any conclusions from it. This is a prevalent problem with all tools and visualization that directly interact with the representations used internally by the system. Typically, the test persons had difficulties to associate the representations with properties of the seen images.

The manual weight setting dialog, the weight inspector, and the polyrepresentation visualization were designed as expert tools. From this perspective, it is not surprising that the layperson users in the user study avoided these tools because of their complexity or failed to operate them correctly. However, it is astonishing that the subjects also seldom modified the set of QBE documents, although they had learned about this functionality. One possible reason might be that the drop region of the multimedia query documents box is too small to be hit by the user. This problem is illustrated by subject 30 (see screen recording from minute 9:45 onward), who repeatedly tried to drag a new QBE document to the query box.

Roughly speaking, the faceted navigation panel, which is placed second to the preference rings in terms of utility, does not suffer from direct usability problems. The subjects could choose the facets they were interested in and retrieved the corresponding result documents. Nevertheless, the usability of the faceted navigation is negatively affected by its varying retrieval effectiveness. In particular, this phenomenon becomes obvious whenever facets that require a perfectly working face detection (e.g., see Figure 9.14; 4) are chosen. In such a case, users obviously expected perfect results motivated by the clear naming of the facets. As the retrieval engine cannot guarantee the expected perfect results in every case, the subjects tended to use the faceted navigation in a trial and error manner. This becomes evident if one reconsiders the high frequency of faceted navigation to faceted navigation transitions (see Figure 9.13; bottom left), or if one examines the screen recording of subject “j”. This problem of faceted navigation is known, and commonly attributed to a gap between the user’s mental model and the conceptual model of the IR system [Hearst 2009, cf. p. 195]. That is, the user’s expectations (a sharply bound set of results) conflict with the “unsharp” results of the IR system, which works with probabilities to assess document relevances. This is also why facets should normally be chosen manually (as it is the case with the Pythia MIR system) and not automatically because the latter cannot guarantee comprehensible facets that support the user.

Because of the nature of the simulated work task user study with a time limit of 15 minutes, functionalities that are meant to support users during longer search sessions, such as the manual organization of documents in the result visualization panel or tagging, were infrequently used. The same holds true for the multiple selection of documents in order to execute an operation on many documents simultaneously (e.g.,

## 9 Evaluation of the User Experience

to associate them with the same relevance level).

Regarding the result visualizations, the matrix visualization, known to most subjects from other CBIR engines and also present in all GUI variants, is preferred by most users (see Figure 9.11). The cluster visualization is also frequently used, whereas the SOM visualization is avoided by most subjects. As mentioned before, most subjects used only one result visualization to solve the work tasks which is also illustrated by the video-based user observations. Most likely, the relatively short duration of the simulated work tasks causes the infrequent transitions between the provided result visualizations. Interestingly, the different result visualizations were criticized in the open question only by a few test persons, which suggests that many users might not have noticed these functionalities.

To recapitulate, the central functionalities of the Pythia MIR system could be used by layperson users without considerable usability problems when confronted with the GUI variants for the first time. However, a future version of the MIR system should emphasize the search history and the query modification facilities in order to make the users aware of the functions that provide support for correcting errors and finding more relevant documents.

The general problem of the current prototype is the missing self-descriptiveness for layperson users. If this issue is addressed, one can expect also an increased level of controllability and an establishment of a link between the various control mechanisms and visualizations. This step is crucial to establish a link between the user's cognitive space and the information space as postulated in Section 5.4.2 (see Figure 5.4). At the current stage of development, the more advanced features of the Pythia MIR system such as the weight modification are basically incomprehensible and unusable for layperson users. Whether this applies to expert users could not be examined within the scope of this dissertation because no significant amount of experts could be recruited for the user study.

Concluding from the generally good evaluation results of the quantitative usability study of GUI variant T4 (the Pythia MIR system) presented in Section 9.3, it can be said that most usability problems can be addressed with a combination of appropriate user trainings, an online help, and more descriptive labels. From an end-user's perspective, the system's error tolerance and individualization options in particular should be improved. However, these usability goals were not in the primary focus of the prototype's development, which is meant to investigate the utility of the PrefCQQL approach in MIR.

## **Part V**

# **Conclusions and Outlook**





## 10 Discussion and Conclusions

This chapter juxtaposes the results from the evaluations described in Chapters 8 and 9 in order to place them in the context of the research results presented throughout this dissertation and to discuss them briefly in a joint manner. To come full circle, Section 10.2 gives answers to the research questions posed in Section 1.2.

### 10.1 Discussion

This dissertation relies on *CQQL* and *PrefCQQL* to implement the principle of *polyrepresentation in MIR*. Consequently, it examines whether the hypotheses of the PoP hold in different MIR scenarios. A scenario, as presented before, is defined by the retrieval process in different document collections with various thematic backgrounds (see Section 8.4.1). Moreover, the examined scenarios can be based on PrefCQQL-based relevance feedback or do not require interactions in form of RF.

The PoP follows the cognitive viewpoint on IR, i.e., the acknowledgment that information processing during the IR process takes place in all participants of this process, e.g., users or systems, although at different levels (see Section 3.1.2). In other words, the participants of the IR process interactively affect each other. A similar perspective is taken by PrefCQQL's user-centered preference model for MIR (see Section 5.4.2) which recognizes the differences in information processing and reasoning about preferences at different levels, i.e., the cognitive and the information space in particular. Furthermore, the PrefCQQL-based RF process is seen as a continuous dynamic information flow (see Figure 5.4) between the user and the system. These assumptions align PrefCQQL with the core ideas of the cognitive viewpoint as outlined by Ingwersen [1992, cf. p. 17].

The PoP of the information space, i.e., the system's perspective on IR, relying on the formation of cognitive overlaps to improve retrieval effectiveness is strongly related to the field of feature fusion approaches which typically assume that an appropriate combination of multiple low-level features will model higher-level semantical concepts [Del Bimbo 1999; Feng et al. 2003; Lew et al. 2006]. This hypothesis is supported by the relevant research results in the MIR community. However, typical fusion approaches do not make generalizable statements on how an effective feature fusion can be achieved. This discriminates the PoP from such approaches as it suggest explicitly the formation of cognitive overlaps.

In addition to the system's perspective on IR, the global model of polyrepresentation (see Section 6.2.1) also addresses the cognitive states a user experiences during the retrieval process. According to Ingwersen [1996], these cognitive states manifest in four IN extreme cases (see Figure 6.4). These IN cases motivate the user need for different

information seeking strategies, namely *directed and exploratory search*. To support these ISS and to bridge the gap between the the user's mental and the system's conceptual model, the *Pythia MIR system*, which has been developed as a part of this thesis, offers different widgets and communication paths that will be described below.

In order to investigate the validity of the hypotheses of the PoP in combination with CQQL/PrefCQQL, different experiments have been conducted. According to Fuhr [2012], the presented experiments have to be regarded as *why-experimentation* because they aim at validating the PoP's assumptions about the development of retrieval effectiveness. From a methodical point of view, this discriminates the experiments from typical IR experimentation mostly aiming at an optimization of retrieval parameters (how-experimentation). The usability-related experiments are based on an initial but comprehensive user study with 59 subjects, which compares the usability of three GUI variants (see Section 9.2). The study has been carried out in a laboratory environment and must be interpreted respectively (see Section 9.3.1).

An additional methodical contribution of this dissertation is its focus on *reproducibility* which is ensured by providing the full source code of the Pythia MIR system as well as the retrieval effectiveness examination toolchain as a supplement to this thesis (see Appendix E).

To begin with the discussion on the results of the experiments in more detail, we approach the Pythia MIR system and the underlying CQQL/PrefCQQL-based retrieval model's effectiveness from a user's point of view.

### 10.1.1 User Experience

As mentioned before, CQQL and PrefCQQL form the core of the polyrepresentative user interaction model of the Pythia MIR system prototype (see Section 6.2.1). PrefCQQL constitutes a *hybrid preference approach* that both features qualitative inductive preferences and quantitative preferences in form of the weighting variables supported by CQQL (see Section 5.6). In the scope of (M)IR, PrefCQQL can be interpreted as a relevance feedback approach – or more precise – a learning to rank technique. Unique features of the PrefCQQL approach range from the support of negative QBE documents at query formulation time as well as during the interactive retrieval process. Furthermore, PrefCQQL allows the formulation of weak preferences between result documents to express gradual levels of relevance. In addition, inductive preferences can be used from query formulation time onward to learn new CQQL queries (see Section 5.5). As a consequence, the general user interaction can be based, in principle, on inductive preferences alone.

To cope with the different IN cases assumed by the PoP and to facilitate a transition between them, the Pythia MIR systems supports other ISS than the aforementioned PrefCQQL-based RF. These ISS range from directed search to exploratory search techniques such as faceted navigation or browsing. To assist users during information seeking, further supportive mechanisms, e.g., a search history, are available. It is generally believed that the support of multiple ISS increases a MIR system's utility.

However, critics of RF approaches often deny them user acceptance. Nevertheless,

the workshops held as part of the user-centered design process (see Section 6.1.1) of the Pythia MIR system indicate that users are willing to provide RF if it guarantees a fast adaptivity while remaining controllable (see User Stories 1 and 2). This demand is satisfied as the good usability attested to the PrefCQQL-based RF approach illustrates (see Section 9.4). For instance, the examined GUI variants featuring PrefCQQL-based RF are perceived as the most controllable variants by the subjects (see Section 9.3). Concerning the fast adaptivity of the RF approach, the retrieval effectiveness evaluation clearly shows that the PrefCQQL-based retrieval engine reacts fast at the first RF iteration (see Figure 8.21 and Appendix B.3.2). Furthermore, the PrefCQQL approach does not rely on a high number of input preferences to cause an impact on the retrieved rank (see Section 8.5.5 and Figure 8.29). Unfortunately, the change in the retrieved ranks does not necessarily result in an increase of the retrieval effectiveness (see below).

Another important aspect of the user experience that is directly related to the fast adaptivity of the MIR system is its responsiveness. As Section 5.2 shows, the PrefCQQL approach reaches roughly 90% of the retrieval performance of the first placed machine-based learning approach in the ImageCLEF 2013 Personal Photo Retrieval subtask, while only taking a tenth of the first placed approach's processing time and requiring significantly less powerful hardware. Although further investigations in this area are necessary, a definite trend is evident indicating the feasibility of PrefCQQL for interactive MIR.

From a more abstract point of view, the users involved in the development of the Pythia MIR system demanded that the control and the training of the provided algorithms must function intuitively (see Section 6.1.1). If one interprets intuitive handling as a combination of self-descriptiveness, controllability, user expectation conformity, and suitability for learning, the GUI variants featuring PrefCQQL-based RF as well as the complete Pythia MIR system obtain mostly good ratings for these usability criteria (see Section 9.3). However, regarding self-descriptiveness, the trend is only slightly positive. In general, all examined GUI variants are criticized for their lack of self-descriptiveness. This criticism becomes particularly obvious in the open question of the usability study (see Section 9.4.1).

In any case, the Pythia MIR system is considered the best of all examined GUI variants (see Section 9.4). Regarding the ISS user interface mechanisms, the PrefCQQL preference rings widget receives the best rating. On average, it is rated good. The faceted navigation is rated mediocre (which is mainly due to the issues with the retrieval accuracy of the face recognition) by the subjects, but clearly better than the manual weights widget.

The relatively poor evaluation results of the manual weight setting dialog and the result explanation mechanisms are not surprising considering the fact that they are placed on interface layer 3, which primarily addresses the needs of the professional and the research persona. However, corresponding user groups could not be recruited as subjects in the user study. In other words, even the more complex UI elements have only been tested by layperson users. Thus, it is not surprising that layperson users avoided these tools because of their complexity or failed to operate them correctly. Unfortunately, these UI elements are meant to establish the link between the information and

the cognitive space which is obviously not possible – at least for layperson users. As a consequence, the gap between the user’s mental and the system’s conceptual model of the current IN cannot be bridged reliably. This might result in unpredictable results and user dissatisfaction that poses a potential usability risk for the examined version of the Pythia MIR system prototype.

Concerning the usage of different visualization methods, matrix visualization is preferred by most test persons. While the cluster visualization’s utility is still graded good, the utility of the SOM is rated satisfactory by the majority of the subjects. Although a SOM-based visualization is demanded by User Story 11, its poor grade is not surprising given the fact that its main functionality, the visualization of similarities and dissimilarities by placing documents in direct spatial neighborhood, was demanded mostly by professional users that participated in the user workshops during development (see Section 6.1.1). As said before, this user group is not represented by the subjects in the usability study. Moreover, the data indicates that users seldom change the type of visualization during the search (see Figure 9.13). Whether this is due to the differences in the perceived utility of the various visualizations or the relatively short interaction time with the MIR system cannot be answered on the basis of the obtained data and therefore needs further research.

To conclude, the conducted user study reveals a preference for exploratory search by most subjects. This finding is consistent with other research results presented in Section 6.1.4. Generally speaking, new directed searches could be observed relatively seldom (see Section 9.5). However, the PrefCQQL RF mechanism is used often – its usage frequency is only surpassed by faceted navigation. This can be interpreted as an additional indicator of its good usability. Moreover, the data shows that if the subjects did not rely on faceted navigation or PrefCQQL RF alone, they frequently alternated between these two search strategies. Interestingly, although the faceted navigation is only rated mediocre regarding its utility, its straightforward usability seems to compensate its deficits in retrieval accuracy (see above).

### 10.1.2 Retrieval Effectiveness and Further Properties of PrefCQQL

The retrieval effectiveness of various feature fusion approaches, including such that follow the recommendations of the PoP, is compared against different baselines in Section 8.4.3. The results clearly show that poorly performing representations affect the retrieval effectiveness of a matching function negatively – even if the matching function is consonant with the PoP. This effect is consistent with the unstable retrieval performance of the PoP observed in some cases by Larsen [2004, cf. pp. 36f.] and Larsen et al. [2009]. In contrast, the positive results of the PoP reported by Skov et al. [2004], who use exact matching, and Larsen et al. [2009], who rely on best matching systems, could not be verified in non-interactive MIR as the arithmetic mean (best features variant) typically outperforms the conjunction (best features variant), which is the logical equivalent to the cognitive overlap in CQQL (see Table 8.16). Moreover, main experiment I (see Section 8.4.3) shows that the fusion of multiple representations is no guarantee to improve the retrieval effectiveness in MIR. In fact, there are matching functions relying

on various representations that perform worse than the color structure baseline or the collection-dependent best performing single representation. This effect becomes particularly visible with the minimum and maximum matching functions. Although the minimum and maximum are in principle fusing matching functions, they suffer from the *dominance problem* (see Section 4.5.3) that makes them behave more like a single representation. To recapitulate, the arithmetic mean and the conjunction are not necessarily outperforming the best performing single representation in every collection. However, these two matching functions are effectiveness-stable (see Definition 8.9), which makes their performance predictable – a factor that is desirable from a user’s perspective.

During interactive MIR based on PrefCQQL, the situation changes (see Section 8.5.2). Although the experiments reveal a gradual increase of the weighted arithmetic mean’s effectiveness during the RF iterations, its effectiveness is surpassed by the effectiveness development of the conjunction from RF iteration 1 onward. In other words, in the interactive MIR scenario, the predictions of the PoP can be verified. Moreover, the results provide sufficient evidence that the PrefCQQL approach is effective when used with the described user simulation in MIR. To further support this claim, the conjunction has been examined at an extreme case with 15 RF iterations which also does not provide evidence for overfitting problems with this matching function. However, the PrefCQQL approach tends to stabilize after five RF iterations. That is, the retrieval effectiveness stagnates (see Section 8.5.3). This effect is most likely due to the fact that, after a number of RF iterations, no more informative preferences are input by the utilized user simulation. Consequently, we recommend to suggest a new query after subsequent RF iterations that do not result in significantly different ranks. For instance, such effects can be measured with the help of rank correlation coefficients or statistics based on the weighting variables of CQQL (see below).

Another interesting observation is the *disjunctive effect*, i.e., a decrease of the retrieval effectiveness after the first RF iteration of matching functions with pre-dominant disjunctive characteristics. This observation further supports the PoP as disjunctions form the logical counterpart to conjunctions and thus cognitive overlaps.

Nevertheless, the dominance of conjunctive characteristics in a matching function alone does not guarantee a gradual effectiveness increase during RF. The development of the retrieval effectiveness is also affected by the number of representations present in the matching function – at least if PrefCQQL is involved. That is, if only a small number of representations and thus weighting variables is available in a matching function, the effectiveness is likely to decrease during RF. This phenomenon is an indication of an underfitting of the IN in its PrefCQQL-based model. Roughly speaking, the effectiveness of a matching function during RF depends on the logical structure of the matching function *and* the number of representations used to model the user’s IN. Concerning the logical structure and the number of addressed representations, a definite trend is evident: *matching functions consistent with the PoP and a sufficient number of weighting variables will outperform alternatives and will act predictably*. Alas, when the “magical” point of sufficient weighting variables to support PrefCQQL is reached could not be revealed in this dissertation.

Because of the novelty of the PrefCQQL approach, additional properties of PrefCQQL and CQQL need further attention, e.g., how the elicitation of inductive preferences and the resulting machine-based learning affects both weighting variables and retrieval effectiveness. Additional properties include preference conflicts or statistics that might indicate the need for the suggestion of a new query that better reflects the user's dynamic information need.

Besides increasing the validity of statements about a matching function's retrieval effectiveness as recommend by Voorhees [2008], the examination of a matching function with different test collections also allows insights into the development of the underlying CQQL weighting variables. For instance, Figures B.12 to B.16 illustrate that the relation between the weighting variables' values is not constant over the examined collections, i.e., the learnt weight values are characteristic for each collection. Interestingly, when the collections are examined separately, a typical weight value pattern for each collection can be observed (see Appendix B.3.2). In most cases, the weight values gradually *and* directly converge towards this characteristic pattern during PrefCQQL-based RF. Hence, there is evidence that PrefCQQL adapts well to different retrieval scenarios.

In order to reveal whether statistics of the weight values can be exploited to assist users during information seeking, Section 8.5.4 examines various statistical figures. To start with, the intuitive understanding that *only* a big change in the weighting variables' values between two RF iterations will have a big impact on the retrieval effectiveness cannot be justified. Figure 8.26 shows that many RF iterations result into a small change of the weight values learnt by PrefCQQL's weight learning algorithm on the basis of the preferences input by the simulated user. That is, even small changes of the weight values can have a tremendous effect on the change of the retrieval effectiveness – whether it will be positive or negative. This finding has to be reflected in the UI of a PrefCQQL-based system; for instance via an appropriate mapping of the manual weight values slider onto the corresponding weighting variables' values. This is particularly important because the manual weight value setting dialog already receives poor usability ratings, which might be affected negatively even more if this phenomenon is not addressed appropriately.

The dynamics of the user's IN has been discussed extensively in this thesis. Hence, it is reasonable to search for statistics or effects that can serve as indicators of a recent IN change that may be better addressed by a newly formulated query. If a changed IN can be detected reliably by the MIR system, the system can learn new queries on the basis of the input preferences (see Section 5.5) or suggest to manually reformulate the current query in order to assist the user during information seeking. As argued in Section 5.4.7, query incompatibilities and preference conflicts might serve as valuable indicators of a tremendous or spontaneous IN change.

Although a correlation between the weighting scheme distance between two RF iterations and the retrieval effectiveness (see Section 8.5.4) cannot be observed, there is evidence that the weighting scheme distance might be utilized as an indicator for the stagnation of the current RF process (see Section 8.5.3). In such a case, the system can suggest a reformulation of the initial query or the modification of the currently used

preferences. However, to make resilient statements about this statement's general validity, further experiments with user simulations that use different preference elicitation strategies are needed.

Additional conclusions cannot be drawn from the statistics of the examined data.

To conclude, the least surprising effect is presented. The experimental data clearly shows that the chance to observe a query incompatibility rises with the number of stated preferences (see Section 8.5.5). Interestingly, the risk of observing query incompatibilities rises significantly between the fifth and sixth RF iteration corresponding almost perfectly to the moment from which on the retrieval effectiveness during RF stagnates. Most likely, this is due to the fact that the Pythia MIR system continues to use the last valid weights from the point on a query incompatibility is detected.

Moreover, there is evidence that matching functions with few weighting variables are likely to become subject to frequent query incompatibility issues as the Eidenberger variants or Q10 clearly illustrate. In conjunction with the findings presented above, it is reasonable to conclude that matching functions used within the PrefCQQL approach must feature a number of weighting variables, which yet has to be discovered.

## 10.2 Conclusions

To conclude the thesis, we will answer its core research questions posed in Section 1.2. As said before, the research questions aim at investigating whether the hypotheses of the PoP can be verified in the context of CQQL/PrefCQQL.

### 10.2.1 Can the Hypotheses of the Principle of Polyrepresentation Be Verified in Multimedia Information Retrieval?

Evidence of the PoP's utility in textual IR has been presented [Skov et al. 2004; Larsen et al. 2006, 2009]. However, an investigation in MIR was missing. Roughly speaking, one of the core predictions of the PoP is that cognitive overlap and hence conjunctive matching functions should outperform other kinds of matching functions. To answer this question, Section 1.2 suggested to break it into three main parts.

#### Can CQQL Be Used to Implement a Formal IR Model on the Basis of the Principle of Polyrepresentation?

As elaborated in Section 4.6, CQQL can be used to implement the PoP in MIR. As a consequence, all subsequently made statements apply to the PoP's implementation using the quantum-logic based and probabilistic IR model provided by CQQL (see Section 4.5.5).

### **Do the Hypotheses of the Principle of Polyrepresentation Apply in Multimedia Information Retrieval?**

In non-interactive MIR, i.e., when no RF is given, the experimental data does not provide sufficient justification for the statement that PoP-based matching functions will always surpass single features or other matching functions. For instance, the arithmetic mean surpasses the conjunction/geometric mean and hence the cognitive overlap of multiple representations in terms of retrieval effectiveness (see Figure 8.11). Nevertheless, the matching function following the PoP is effectiveness stabler than the best performing single representations per collection. Hence, the cognitive overlap's retrieval performance is more reliable than the usage of single representations.

In any case, the conjunction still performs better than any other examined matching function, besides the arithmetic mean, as Table 8.16 illustrates. Moreover, there is evidence that the conjunctive variant of a matching function always performs better than its disjunctive counterpart (see Section 8.4.3). This finding complies with the predictions of the PoP.

To recapitulate, if RF does not have to be supported by the retrieval engine, the usage of the arithmetic mean as the primary matching function has to be recommended. If only logic-based matching functions are available, the conjunction clearly surpasses all other matching functions.

### **Can Polyrepresentation Compensate the Weak Retrieval Effectiveness of Some Low-Level Features in Multimedia Information Retrieval?**

Unfortunately, there is no evidence that the PoP can compensate the weak retrieval effectiveness of some representations. That is, weakly performing representations, e.g., those that rely on some sort of automatic image segmentation, have a negative effect on the retrieval effectiveness, even if PoP-motivated matching functions are used. In fact, this phenomenon is present in all examined matching functions as the comparison between their normal and best features variants shows (see Figures 8.14 and 8.15). Hence, the choice of representations to be used during retrieval has to be examined carefully. However, the experiments indicate that the results of the single representation effectiveness study can be directly transferred to the fusion-based experiments. In other words, representations that perform already weak in isolation, should be excluded from the formation of cognitive overlaps or other feature fusions.

#### **10.2.2 Do the Hypotheses of the Principle of Polyrepresentation Hold in a Preference-based Interactive Multimedia Search Process?**

Although the usage of single representations as a matching function surpasses the retrieval effectiveness of fusion approaches such as the PoP in some cases, single representations cannot be used during RF. In such an interactive MIR scenario, which is realized with PrefCQQL in the context of this dissertation, the data shows that the conjunction outperforms all other fusion-based matching functions from RF iteration 1 on-



ward. This includes the arithmetic mean. As a consequence, the predictions of the PoP can be verified in the examined PrefCQQL-based interactive MIR scenario. However, it is important to note that also the number of available representations has an impact on the retrieval outcome. That is, if too few representations are present in a matching function, the corresponding IN model in PrefCQQL obviously becomes subject to underfitting eventually lowering its retrieval effectiveness (see Section 8.5.2).

### **10.2.3 Can a Usable Multimedia Information Retrieval System Be Built on the Basis of the Principle of Polyrepresentation and PrefCQQL?**

This question can be answered to a large extent by the proof of concept, the Pythia MIR system prototype, described in Chapter 6. The combination of CQQL and PrefCQQL supports different ISS and a seamless transition between them (see Section 6.2.1). The support of various ISS is important to address the different user needs during the information seeking process and is commonly believed to increase a (M)IR system's utility.

The motivation to develop a prototypical interactive MIR system is the objective to test its usability in a real-world scenario. Besides its lack of self-descriptiveness, which has not been prioritized during the prototype's development, the 59 subjects of the usability study (see Chapter 9) attest the Pythia MIR system a high level of usability.

To come to an end, the objective of why-experimentation is to justify a theoretic model's validity, i.e., the validity of the predictions made by the principle of polyrepresentation. This verification process is required to assess a model's validity and to investigate whether the PoP forms a theory. The answer to this question is particularly interesting in conjunction to the experiments conducted in textual IR [Skov et al. 2004; Larsen et al. 2006, 2009]. However, although the experiments presented in this thesis were conducted in a way that allows a generalization of the resulting conclusions, we believe that this dissertation cannot answer this question exhaustively. This is mainly due to the different outcomes of the interactive and non-interactive experiments. While the former presents evidence for the validity of the PoP's assumptions, the latter only shows a good performance of the PoP in MIR which is surpassed by the utilization of the arithmetic mean as a matching function. To ultimately answer the question whether the PoP constitutes a theory for IR and MIR, further research is needed. Nevertheless, the presented research results provide sufficient evidence of the PoP's general utility in MIR as it definitely allows predictions of the retrieval effectiveness development of a matching function.

### 10.3 Future Work

We can only see a short distance ahead, but we can see plenty there that needs to be done.

Alan Turing; 1950

As frequently mentioned throughout this dissertation, there are many areas that need further research. Unfortunately, this section can only present a small selection of ideas for future work. The area of future work can be subdivided into four parts: further evaluation opportunities (see Section 10.3.1), improvements of the Pythia MIR system's usability (see Section 10.3.2), functionality (see Section 10.3.3), and technical improvements addressing the performance of the system (see Section 10.3.4). By no means the discussed areas for future work are meant to be exhaustive. Instead, the following sections shall give a rough orientation of fields that can profit from further research.

To conclude the thesis, Section 10.3.5 sketches additional application areas for the presented polyrepresentative CQQL/PrefCQQL approach.

#### 10.3.1 Evaluation Opportunities

The user study presented in Chapter 9 is positioned as an initial study. That is, to obtain resilient results, it has to be extended to allow insights into all aspects of the Pythia MIR system. For instance, to assess the usability of tagging or manual resorting functionalities, additional long enduring simulated work tasks are needed. Moreover, to obtain usability results for all addressed personas, test persons with a domain expert or professional background should be recruited and the simulated work task(s) should be repeated with this group of users.

Long enduring simulated work tasks are also required to examine whether the alternation frequency between different ISS changes over time or if different visualizations are more likely to be used during longer information seeking sessions.

Another perspective on the retrieval effectiveness of the PrefCQQL approach can be taken by repeating the experiments described in Section 8.5 with different user simulations. As described above, with the examined user simulation, the PrefCQQL RF process tends to stagnate after approximately five RF iterations. To ultimately find out whether this effect is due to PrefCQQL or, as assumed, the utilized user simulation, further user simulations are needed.

As this dissertation only examines PrefCQQL's effectiveness during explicit RF, an additional investigation of PrefCQQL's feasibility for pseudo RF (see Section 5.1.2) is reasonable. First experiments show a trend that PrefCQQL is not subject to query drifting during pseudo RF when the conjunction is used (see Figure 10.1). However, the figure also illustrates that pseudo RF limits the retrieval effectiveness development. Most likely this is due to the similarity between the user simulation, which is used from RF iteration 2 onward, and the pseudo RF strategy which also only switches preferences. This approach resembles the strategy used in the second step of the user simulation with the difference that the ground truth is not available to the pseudo RF algorithm (see Section 8.5.1).

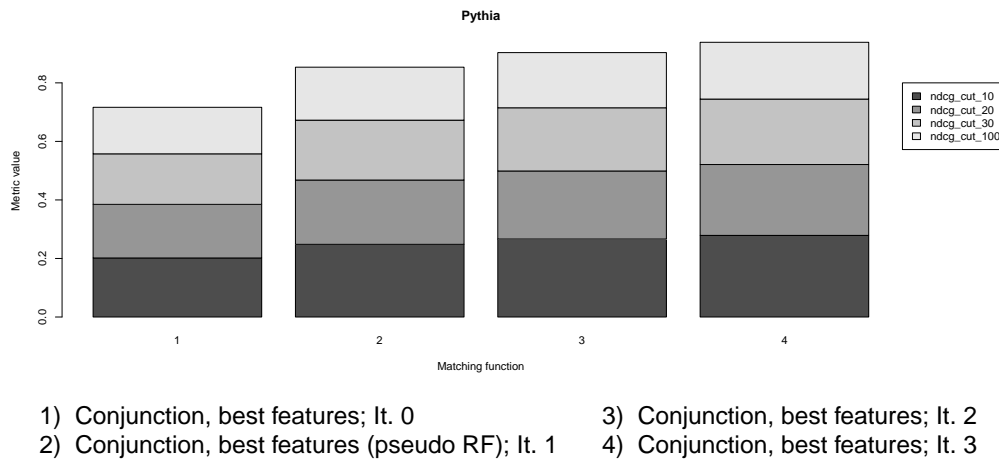


Figure 10.1: Performance comparison of RF-enabled representation combinations and standard aggregations

From a strict statistical point of view, all experiments should be repeated with a higher number of documents, collections, matching functions, and subjects to increase their level of expressiveness. In particular, this is true for the choice of matching functions because the presented data cannot give a definite answer to the question of how many representations are needed to guarantee an increase of retrieval effectiveness for a conjunctive matching function during RF. Moreover, a different choice of matching functions might give answers to the questions posed in Section 8.5.2 regarding the impact of a matching function's dominant logical characteristic onto its behavior during relevance feedback.

To further investigate PrefCQQL's weight learning algorithm, a Monte Carlo experiment could be conducted. The utilized weight learning algorithm (see Section 5.4.4) is based on pseudo-random numbers using a fixed random seed to ensure the reproducibility of the experiments. However, this also means that the random numbers used to start the simplex algorithm only cover a small region of the parameter space. In order to obtain statistical resilient results about the learning algorithm's behavior, a Monte Carlo experiment can be used to investigate how the simplex algorithm behaves if the parameter space is sampled evenly. Roughly speaking, the core idea of the suggested Monte Carlo experiment is to repeat the user simulation as long as the learning algorithm's simplex has been defined by a number of vertices that are distributed evenly over the whole parameter space. As long as this form of experiment has not been conducted, it cannot be inferred safely whether the weighting variable values used in this thesis are optimal. As a result, this thesis cannot give a definite answer to the question of how the algorithm will perform on average in every scenario. Instead, it has to be stated that the measured performance might increase or decrease if other pseudo-random number sequences are used.

Alternatively, all experiments could be repeated with different mapping strategies for

## 10 Discussion and Conclusions

unavailable representations (see Section 8.4.3) or using statistically standardized similarity scores for each representation. At the current stage, a similarity score regarding a representation has a distinct statistical distribution. From a statistical point of view, this complicates the comparison of such variables because of their different distribution. To address this issue, a z-transformation could be used [Falk et al. 2002].

As these examples illustrate, the opportunities for additional evaluations and experimental designs are almost unlimited. Although this list is far from being complete, we will end its presentation here and leave room for the reader's imagination.

### 10.3.2 Usability Improvements

As mentioned before, the Pythia MIR system prototype has been criticized for its lack of self-descriptiveness by a high number of study participants. As a result, this usability factor has to be addressed in future releases along with the provision of an online help.

Furthermore, a future version of the system should emphasize the search history and the query modification facilities in order to make the users aware of these functions as they provide support for correcting errors and finding more relevant documents.

To establish a link between the user's cognitive space and the information space as postulated in Section 5.4.2, the usability of the corresponding widgets such as the poly-representation visualization (see Section 6.2.9) has to be improved. Concluding from the study results, the more advanced functionalities suffer particularly from their lack of self-descriptiveness and remain invisible or uncontrollable for most users.

To conclude, the faceted navigation's retrieval effectiveness regarding persons has to be improved to enhance the user experience of the prototype.

### 10.3.3 Functional Extensions

Although preference aging is supported by the PrefCQQL approach in principle (see Section 5.4.7), the inclusion of this functionality is desirable because of its convincing utility reported by Campbell [2000] and Liu et al. [2009].

Furthermore, a functionality aiming primarily at advanced users is a direct modification of the PrefCQQL preference graph. For instance, the preference graph could be displayed in an additional dialog enabling users to state more complex preferences than the ones possible with the concentric circles metaphor (see Section 6.2.5).

Another functional extension that might assist users during information seeking would be the suggestion of preferences that result in meaningful restriction of the information space, e.g., by bisecting the result list.

As outlined in Section 6.3.2, the handling of PDF documents in the current prototype is only available in a rudimentary manner. At the current stage of development, it only constitutes a proof of concept of the extraction of different media from composite PDF documents. It remains open how the extracted media objects have to be combined in a meaningful way to be of use during the information seeking process.

To come to an end, another functional extension, whose implementation is particularly compelling, would be the realization of CQQL-based polythetic clustering as

sketched in Section 6.2.4.

### 10.3.4 Technical Improvements

Although the Pythia MIR system prototype already uses various optimizations (see Section 8.4.1), the notion of a prototype implies a variety of possible technical improvements. The most obvious optimization area is the continuous integration of parallelization into the application in order to fully exploit the parallel processing power of modern CPUs and GPUs.

For instance, parallelization can be used to search for useful preferences (see above) in the background without disturbing the user interaction. Moreover, the most probable next queries can be submitted before the user manually triggers them in order to increase the system's responsiveness or to suggest new queries.

In a similar manner, parallelization can be used to render the various result visualizations in the background to allow an instant change between the different views on the results. At the current stage of development, each switch between the visualizations causes a new rendering of the corresponding widget.

On an even lower level, the distance calculations between the representations can be optimized by avoiding superfluous calculations or by caching results whenever it is appropriate. In addition, further improvements of the system's performance are expected from changing the representations' internal storage format from file-based XML to a less complex storage structure or a feasible database system.

Although all these exemplary improvements will have an impact on the system's responsiveness and thus on its user experience, they have been neglected because of this dissertation's focus on the principal design and evaluation of an interactive polyrepresentative MIR system. As a result, there are many opportunities to fine-tune the source code as well by revising the internal storage and data access functionalities.

### 10.3.5 Future Directions and Application Areas

Besides this dissertation's focus on visual MIR, the presented Pythia MIR system and its underlying retrieval model in particular are not limited to this domain.

For instance, CQQL has also been used successfully in music retrieval to model musical genres within the "GlobalMusic2one" research project<sup>169</sup>. In this research project, CQQL is used to formulate matching functions on the basis of descriptions provided by musicologists characterizing musical genres as logical sentences of propositions about automatically extracted representations such as a song's instrumentation, rhythm, approximate length, or the geographic origin of a song [Schiela 2010; Zellhöfer & Schmitt 2011b].

Alternatively, PrefCQQL can be used in traditional IR or relational databases as mentioned in Section 5.4. To give an example, in a supervised bachelor's thesis, Buckow [2009] uses CQQL with manually pre-determined weighting variable settings to retrieve auction lots from a fine-arts database. In this scenario, the weighting variables

<sup>169</sup>[http://www.globalmusic2one.net/en\\_summary.html](http://www.globalmusic2one.net/en_summary.html)

are set by domain experts and PrefCQQL is not used. However, PrefCQQL support can be easily added in order to personalize the search results.

A similar application area is the search in library catalogs, e.g., during retroconversion (the mapping of an outdated catalog syntax to a modern one) or revision (the correction or addition of catalog information). In this use case, an incomplete catalog entry, which happens to be interpretable as a logical CQQL query, could be used to retrieve similar catalog entries to find missing information. Using PrefCQQL, this process can be repeated until all missing information is collected from similar catalog entries. Moreover, the sketched query learning capabilities of PrefCQQL (see Section 5.5) can be used to learn queries if a manual statement is too complex.

As mentioned before, CQQL has also been used successfully to process combined textual and CBIR queries in order to retrieve data from the Wikipedia at ImageCLEF 2011 [Zellhöfer & Böttcher 2011].

To come to an end, the most interesting area for future research on a formal level is the inclusion of quantum entanglement into CQQL. In quantum mechanics, quantum entanglement describes the phenomenon that two particles cannot be described independently. That is, these quantum particles constantly interact with each other. Similar phenomena are imaginable in (M)IR. For instance, certain representations interact with each other such as the author and the literary genre or the color layout and the contours present in an image. However, at the time of writing this thesis, none of the quantum mechanics-inspired IR models discussed in Section 4.5.4 address this open issue.

**Part VI**  
**Appendices**





# A Mathematical Foundations

The main objective of this appendix is to sketch the mathematical foundations needed for the understanding of this thesis. As such, it cannot replace a full mathematical introduction on the topics becoming relevant. This is in particular true for the field of statistics and probability theory which is more thoroughly covered by Mendenhall et al. [1999]; Miller & Miller [1999], or Ash [2008].

Basic overviews of the mathematical concepts being used in IR are available by a number of authors, e.g., by Dominich [2008] or Croft et al. [2009].

## A.1 Basics

### A.1.1 Partially Ordered Sets (Posets)

**Definition A.1 Partially ordered set (Poset):** A partially ordered set (*poset*) is a set with a partial order. A (weak or non-strict) partial order is a binary relation  $\mathcal{R}$ , e.g., denoted by  $\leq$ , defined over a set  $\mathbf{S}$  ( $\mathcal{R} \subseteq \mathbf{S} \times \mathbf{S}$ ) which is anti-symmetric, reflexive and transitive for all  $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbf{S}$ . ◇

**Definition A.2 Anti-symmetry:** If  $\mathbf{a}\mathcal{R}\mathbf{b}$  and  $\mathbf{b}\mathcal{R}\mathbf{a}$ , then  $\mathbf{a} = \mathbf{b}$  holds. ◇

**Definition A.3 Reflexivity:** The following holds:  $\mathbf{a}\mathcal{R}\mathbf{a}$ . ◇

**Definition A.4 Transitivity:** If  $\mathbf{a}\mathcal{R}\mathbf{b}$  and  $\mathbf{b}\mathcal{R}\mathbf{c}$ , then  $\mathbf{a}\mathcal{R}\mathbf{c}$ . ◇

**Definition A.5 Strict partially ordered set:** A partially ordered set which is irreflexive, i.e.,  $\neg(\mathbf{a}\mathcal{R}\mathbf{a})$ , is called a strict poset. A typical binary order relation  $\mathcal{R}$  for a strict poset is  $<$ . ◇

### A.1.2 Total Order

**Definition A.6 Total order:** A total order is a partial order  $\mathcal{R}$  over a set  $\mathbf{S}$  that is also total, i.e.:  $\forall \mathbf{a}, \mathbf{b} \in \mathbf{S} \mid \mathbf{a}\mathcal{R}\mathbf{b} \vee \mathbf{b}\mathcal{R}\mathbf{a}$  holds. ◇

### A.1.3 Lattices

A lattice is a set of elements over which a poset is defined. Every two elements of the lattice have one supremum (*join*) and one infimum (*meet*). In addition, the laws of absorption, associativity, and commutativity hold for all elements of the lattice.

Figure A.1 depicts a lattice with the partial order  $\subset$  ("subset of") using a Hasse diagram.

**Definition A.7 Hasse diagram:** A Hasse diagram is an ordered, directed and acyclic graph. Following the direction of the arrows illustrates the join operation of two sets. Moving against this direction illustrates the meet operation of two sets. Each node of the graph depicts a set of letters while  $\emptyset$  denotes the empty set.

Generally speaking, a Hasse diagram is constructed by drawing a line pointing from an element  $\mathbf{A}$  to  $\mathbf{B}$  if  $\mathbf{A} < \mathbf{B}$  while  $\mathbf{A}$  appears at a lower position than  $\mathbf{B}$  in the graph. Additionally, there must be no element  $\mathbf{C}$  with  $\mathbf{A} < \mathbf{C} < \mathbf{B}$ . That is, a Hasse diagram visualizes the transitive reflexive reduction of a poset.  $\diamond$

A complete lattice is a lattice whose subsets have all a supremum and an infimum.

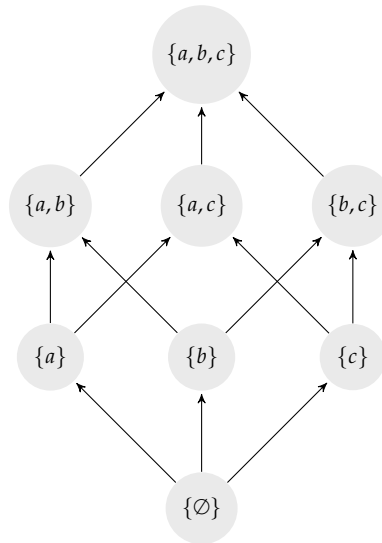


Figure A.1: Hasse diagram of a lattice with the ordering criterion “subset of”; arrows point to the greater element

### A.1.4 What is Logic?

According to the Oxford English dictionary *logic* can be described as

“the systematic use of symbolic and mathematical techniques to determine the forms of valid deductive argument.”

In this thesis, we only discuss formal logical systems that are used for reasoning like the classical Boolean logic. Comparable definitions are available in Schönig [1989]; Siefkes [1990], or Dominich [2008].

Roughly speaking, a formal logic consists of a language that operates on symbols in order to assess the truth value of a *sentence*. A sentence consists of *propositions* (which are denoted by using a capitalized letter, e.g.,  $A$ ) and *logical connectors* (e.g. conjunction ( $\wedge$ ), disjunction ( $\vee$ ), and negation ( $\neg$ ) in Boolean algebra (see below)), whereas a

proposition can also be a sentence. To infer the truth of a sentence, one replaces the propositions (or elementary propositions [Wittgenstein 2004, cf. Sec. 5 and 5.01]) with their distinct truth value, i.e., *true* or *false* in the case of Boolean logic. We denote *true* with 1 and *false* with 0. Then, one applies the rules defined by the logical connectors as listed in Table A.1. These rules are defined by axioms or can be deduced from them. Two example axioms are the law of the excluded middle, i.e.,  $A \vee \neg A$ , or the law of identity, i.e.,  $A = A$ . Both aforementioned sentences are true.

Table A.1: Semantics of Boolean logical connectors and propositions **A** and **B**

<b>A</b>	<b>B</b>	<b>A ∧ B</b>	<b>A ∨ B</b>	<b>¬A</b>
0	0	0	0	1
0	1	0	1	1
1	0	0	1	0
1	1	1	1	0

### A.1.5 Boolean Algebra

A *Boolean algebra* is a formal logical systems with the following properties. For the propositions  $A, B, C \in \{0, 1\}$  holds. To form sentences, two binary operators (*conjunction/and/meet*  $\wedge$  and *disjunction/or/join*  $\vee$ ) and one unary operator (*negation*  $\neg$ ) are defined.

When considered as a poset with the ordering relation  $\leq$ , 1 is the supremum and 0 the infimum of this structure.

The following axioms hold in a Boolean algebra:

**Definition A.8 Axiom of Associativity:**

$$(A \wedge B) \wedge C = A \wedge (B \wedge C) \quad \text{and} \quad (A \vee B) \vee C = A \vee (B \vee C) \quad (\text{A.1})$$

◇

**Definition A.9 Axiom of Commutativity:**

$$A \wedge B = B \wedge A \quad \text{and} \quad A \vee B = B \vee A \quad (\text{A.2})$$

◇

**Definition A.10 Axiom of Distributivity:**

$$A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C) \quad \text{and} \quad A \vee (B \wedge C) = (A \vee B) \wedge (A \vee C) \quad (\text{A.3})$$

◇

**Definition A.11 Axiom of Identity:**

$$A \wedge 1 = A \quad \text{and} \quad A \vee 0 = A \quad (\text{A.4})$$

◇

Definition A.12 **Axiom of Complements:**

$$A \wedge \neg A = 0 \quad \text{and} \quad A \vee \neg A = 1 \quad (\text{A.5})$$

◇

Definition A.13 **Axiom of Idempotence:**

$$A \wedge A = A \quad \text{and} \quad A \vee A = A \quad (\text{A.6})$$

◇

### A.1.6 Dirac notation

The Dirac notation can be regarded as the *lingua franca* in quantum mechanics and has been originally introduced by Dirac [1958]. The following definitions resemble the overview of the notation given in Schmitt [2008, cf. Sec. 2] and van Rijsbergen [2004, cf. App. I]. Please refer to these contributions for more examples of the notation.

Definition A.14 **Ket vector:** A column vector  $\mathbf{x}$  in a Hilbert space  $\mathbf{H}$  (see Section 4.1) is represented by a ket:  $|x\rangle$ . ◇

Definition A.15 **Bra vector:** The transpose of  $|x\rangle$  yields a row vector bra:  $\langle x|$ . ◇

Definition A.16 **Inner product in bra-ket form:** The inner (or scalar) product of two ket vectors  $|x\rangle$  and  $|y\rangle$  is stated by a bra(c)ket:  $\langle x|y\rangle$ . ◇

Definition A.17 **Norm of a ket vector:** The norm  $\| |x\rangle \|$  of a ket  $|x\rangle$  is defined as:  $\sqrt{\langle x|x\rangle}$ . ◇

Definition A.18 **Outer product in bra-ket form:** The outer product of two kets  $|x\rangle$  and  $|y\rangle$ , which generates a linear operator in form of a matrix, is denoted by:  $|x\rangle\langle y|$ . ◇

## A.2 Probability Theory

This informal introduction of probability theory is oriented on Ash [2008]. It aims at summarizing classical probability theory being based on Kolmogorov's axioms of probability theory. Appendix A.2.4 extends this section by discussing Bayes' theorem that deals with conditional probabilities.

### A.2.1 Basic Terminology

Classical probability theory can be best described with the help of an observation of a random experiment, e.g., the rolling of a die.

The *sample space*  $\Omega$  represents all possible outcomes of a random experiment, i.e.,  $\Omega = \{1, 2, \dots, 6\}$  for the rolling of a die.

*Events* are sentences about the experiment involving a number of outcomes that can be assigned with a truth value, e.g., that an even number has been rolled,  $A = \{2, 4, 6\}$ . Hence,  $A \subseteq \Omega$  holds.

The *probability* of an event  $P(A)$  is a *probability measure*  $P$  that assigns an event  $A$  a value out of  $[0,1]$ , i.e.,  $P : A \subseteq \Omega \rightarrow [0,1]$ , while  $P(\Omega) = 1$  and  $P(\emptyset) = 0$ , i.e., an event that is not in  $\Omega$  cannot occur during the random experiment. Furthermore, for a probability measure countable additivity must hold. That is, for all countable sequences of disjoint events  $\{A_i\}$  (see below) the following holds:

$$P\left(\bigcup_{i \geq 1} A_i\right) = \sum_{i \geq 1} P(A_i). \quad (\text{A.7})$$

To simplify the handling of  $P(A)$ , we commonly assign it a probability based on the frequency or chance of the event, e.g.,  $P(\text{"1 is rolled"}) = \frac{1}{6}$ .

### A.2.2 Event Algebra

Being a Boolean algebra, the algebra of events and probabilities resembles the algebra of real numbers, i.e., a union ( $\cup$ ) corresponds to an addition while an intersection ( $\cap$ ) corresponds to a multiplication, if the events are independent.

Two events  $A$  and  $B$  are *independent* iff

$$P(A \cap B) = P(A) \cdot P(B). \quad (\text{A.8})$$

That is, the occurrence (or non-occurrence) of  $A$  has no effect on the occurrence of  $B$ . Otherwise, Equation (A.12) applies.

The probability of mutually exclusive/disjoint events, i.e.,  $E_i \cap E_j = \emptyset$ ; for  $i \neq j$ , is defined as follows:

$$P(E_1 \cup E_2 \cup \dots \cup E_n) = \sum_{i=1}^n P(E_i) \quad (\text{A.9})$$

If two events are *not* mutually exclusive the following applies:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (\text{A.10})$$

To conclude, let the complement be:

$$P(\Omega \setminus A) = P(\neg A) = 1 - P(A) \quad (\text{A.11})$$

**Conditional Probabilities** If two probabilities are affected by each other, they are called *conditional*. That is, the probability of  $A$  on the condition that  $B$  has occurred is defined as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{if } P(B) \neq 0. \quad (\text{A.12})$$

### A.2.3 Material Implication versus Implication as Conditional Probability

For the material implication, we make use of the notation used by van Rijsbergen [1986b].

$$P(A \supset B) = P(\neg A \vee B) \tag{A.13}$$

In contrast, the implication in which the soundness criterion holds is defined as follows:

$$P(A \rightarrow B) = P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{A.14}$$

That is, it is impossible for the premise of the inference to be true while its conclusion becomes improbable. In other words, only true statements can be proven. This contrasts to the material implication as Table A.2 clearly shows.

Table A.2: Truth table for the material implication and propositions **A** and **B**

A	B	A $\supset$ B
0	0	1
0	1	1
1	0	0
1	1	1

Another difference between both implications becomes clear if we consider the following example extending the illustration by van Rijsbergen [1986b]. Let us consider the implication “if *A* is true, then *B*” as a conditional probability in a classical die rolling experiment. Let *A* be the event that a number less than 3 is rolled, while *B* represents the event that an even number is rolled, i.e.:

$$P(A) = \frac{2}{6} \quad \text{and} \quad P(B) = \frac{3}{6} \tag{A.15}$$

If we substitute the variables in both implications, we obtain different results.

$$P(A \supset B) = P(\neg A \vee B) = \neg \frac{2}{6} + \frac{3}{6} = \neg \frac{2}{6} \cdot \frac{3}{6} = \frac{5}{6} \tag{A.16a}$$

$$P(A \rightarrow B) = P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{2}{6} \cdot \frac{3}{6}}{\frac{2}{6}} = \frac{\frac{1}{6}}{\frac{2}{6}} = \frac{1}{2} \tag{A.16b}$$

This observation, in addition to the soundness criterion that is not violated by Equation (A.14), leads to van Rijsbergen’s conclusion that only a conditional implication is appropriate in the field of IR [van Rijsbergen 1986b].

### A.2.4 Bayes’ Theorem

The Bayes’ theorem can be interpreted as a means to calculate the conditional probabilities of two events *A* and *B* by “inverting” them. The following transformation

can become useful in the domain of IR (as well as in other fields) if certain conditional probabilities are easier to calculate than others.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad (\text{A.17})$$





## B Evaluation Appendix

### B.1 Formulae of the Examined Matching Functions

The following equations use an abbreviated syntactical form to denote a (weighted) CQQL query which can be derived from relational tuple calculus. Let *collection* be a relation containing all document representations,  $d'$  a tuple variable, and *query* a relation containing the query representation  $q'$ . Then, the following query in extended<sup>170</sup> tuple relational calculus [Codd 1970]

$$\{d' \mid collection(d') \wedge (\exists q')(query(q') \wedge d'.cedd \approx q'.cedd \wedge d'.fcth \approx q'.fcth)\}$$

can be abbreviated as

$$q = (CEDD_{q'} \approx CEDD_{d'}) \wedge (FCTH_{q'} \approx FCTH_{d'}),$$

whereas  $CEDD_{q'} \approx CEDD_{d'}$  denotes the value of the CQQL proposition (or condition) that is obtained by calculating the similarity of a query and a document from the collection with respect to its representation using the CEDD descriptor (respectively FCTH) will be abbreviated with

$$q = R_3 \wedge R_9.$$

The index of  $R$  refers to the indices listed in Table B.1. Let  $R$  be the set of available representations and  $|R|$  be the total number of representations, i.e., 19 in the case of this dissertation (see Table B.1). In other words,  $R_i$  denotes the probability of relevance of a document in reference to a representation.

A weighted CQQL query  $q_\theta$  will be abbreviated accordingly with its weighting variables denoted as  $\theta_i$ , i.e., the weighted counterpart of the query above is:

$$q_\theta = R_3 \wedge_{\theta_1, \theta_2} R_9.$$

This query can also be stated as in n-ary form:

$$\bigwedge_{\substack{\theta_i \\ i \in S}} R_i \quad ; \quad S = \{3, 9\}.$$

The same convention applies to disjunction and negation.

<sup>170</sup>Note that we extend the tuple relational calculus by a binary similarity operator  $\approx$  on two attributes.

## B Evaluation Appendix

Table B.1: Available Features and Origin; high-level features are shaded gray

R#	Name	Type	Origin
1	Auto Color Correlogram	color-related, global	Huang et al. [1997]
2	BIC	color-related, global	Stehling et al. [2002]
3	CEDD	texture/color-related, global	Chatzichristofis & Boutalis [2008a]
4	Color Histogram	global	512 bin RGB histogram (own implementation)
5	Color Layout*	color-related, global	Cieplinski et al. [2001]
6	Color Structure*	color-related, global	Cieplinski et al. [2001]
7	Dominant Color*	color-related, global	Cieplinski et al. [2001]
8	Edge Histogram*	edge-related, global	Cieplinski et al. [2001]
9	FCTH	texture/color-related, global	Chatzichristofis & Boutalis [2008b]
10	Scalable Color*	color-related, global	Cieplinski et al. [2001]
11	Tamura	texture-related, global	Tamura et al. [1978]
12	Color Histogram (region based)	color-related, pseudo-local	Balko & Schmitt [2012]
13	Contour-based Shape*	global	Cieplinski et al. [2001]
14	Region-based Shape*	global	Cieplinski et al. [2001]
15	Gabor	texture-related, global	Zhang et al. [2000]
20	Date of creation	temporal	Exif
21	Time of creation	temporal	Exif
22	GPS coordinate	spatial	Exif
23	Camera model	metadata	Exif

\* denotes features in the scope of MPEG-7 [Manjunath et al. 2002]

### Best Features Variants

Many matching functions are available as a normal and a *best features* variant. The best features variant uses all available representations but the three worst performing ones (see Table 8.10), i.e., Gabor ( $R_{15}$ ), contour-based ( $R_{13}$ ), and region-based shape ( $R_{14}$ ). This naming convention applies to all matching functions.

#### B.1.1 Matching Functions Based on CQQL

This section lists logical CQQL queries that are already normalized. All matching functions based on CQQL can be transformed into an arithmetic expression using the rules presented in Chapter 4 (see Figure B.1).

In accordance with Section 8.4.2, the representations are limited to low-level features only because Exif or IPTC data is not available in all examined collections. Thus,  $n = 15$ .

## B.1 Formulae of the Examined Matching Functions

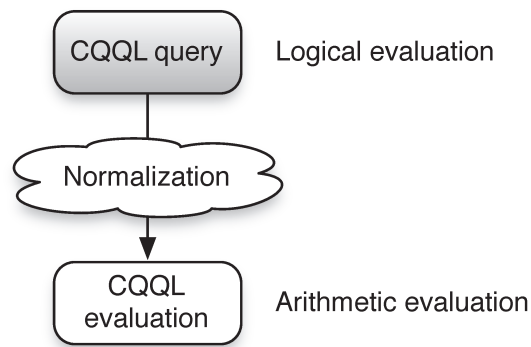


Figure B.1: Transformation of CQQL queries

Matching Function 1 **Conjunction**:

$$\bigwedge_{i=1, \theta_i}^n R_i \quad ; n = 15 \quad (\text{B.1})$$

*Produces the same total order as Matching Function B.16.*

*Source code:* dbis/weightlearning/evalfunction/PythiaGlobalConjunction.cpp

Matching Function 2 **Conjunction, best features**:

$$\bigwedge_{i=1, \theta_i}^n R_i \quad ; n = 12 \quad (\text{B.2})$$

*Produces the same total order as Matching Function B.17.*

*Source code:* dbis/weightlearning/evalfunction/BestAND.cpp

Matching Function 3 **Disjunction**:

$$\bigvee_{i=1, \theta_i}^n R_i \quad ; n = 15 \quad (\text{B.3})$$

*Source code:* dbis/weightlearning/evalfunction/PythiaGlobalDisjunction.cpp

Matching Function 4 **Disjunction, best features**:

$$\bigvee_{i=1, \theta_i}^n R_i \quad ; n = 12 \quad (\text{B.4})$$

*Source code:* dbis/weightlearning/evalfunction/BestOR.cpp

## B Evaluation Appendix

### Matching Function 5 Eidenberger conjunction, variant 1:

$$\bigwedge_{\theta_i} (R_5, R_7, R_8, R_{11}) \quad (\text{B.5})$$

Eidenberger [2003] suggests to combine only color layout, dominant color, and some texture representation(s). This choice is based on results of a retrieval effectiveness study relying only on MPEG-7 features that revealed that the other available representations only contribute redundant information and can therefore be ignored.

Source code: dbis/weightlearning/evalfunction/Eidenberger1AND.cpp

### Matching Function 6 Eidenberger conjunction, variant 2:

$$\bigwedge_{\theta_i} (R_5, R_7, R_8) \quad (\text{B.6})$$

This matching function resembles Matching Function 5 but the Tamura representation  $R_{11}$  has been removed because of its relatively weak performance with the examined collections (see Table 8.10).

Source code: dbis/weightlearning/evalfunction/Eidenberger2AND.cpp

### Matching Function 7 Eidenberger disjunction, variant 1:

$$\bigvee_{\theta_i} (R_5, R_7, R_8, R_{11}) \quad (\text{B.7})$$

This is the disjunctive counter-part of Matching Function 5.

Source code: dbis/weightlearning/evalfunction/Eidenberger1OR.cpp

### Matching Function 8 Eidenberger disjunction, variant 2:

$$\bigvee_{\theta_i} (R_5, R_7, R_8) \quad (\text{B.8})$$

This is the disjunctive counter-part of Matching Function 6.

Source code: dbis/weightlearning/evalfunction/Eidenberger2OR.cpp

### Matching Function 9 Q10:

$$(R_3 \vee_{\theta_1, \theta_2} R_9) \wedge (R_5 \vee_{\theta_3, \theta_k} (R_8 \wedge_{\theta_4, \theta_5} R_{11})) \quad ; \theta_k = 1 \text{ (const.)} \quad (\text{B.9})$$

Q10 is suggested by Zellhöfer & Schmitt [2011a] who report of a pre-study that revealed the generally good retrieval effectiveness of the CEDD ( $R_3$ ), FCTH ( $R_9$ ), color layout ( $R_5$ ), and Tamura ( $R_{11}$ ). Moreover, Q10 reflects the finding from Deselaers et al. [2008] that a combination of texture and edge detectors ( $R_8$ ) can improve the retrieval quality (see Section 8.4.3).

Source code: dbis/weightlearning/evalfunction/Q10.cpp

### Matching Function 10 Semantic group conjunction:

$$\bigwedge_{\theta_i} \left( \left( \bigvee_{\theta_j} (R_3, R_9) \right), \left( \bigvee_{\theta_i} (R_2, R_3, R_{10}, R_{12}, R_6, R_1, R_5, R_4, R_7) \right), \left( \bigvee_{\theta_k} (R_8, R_{11}, R_{15}) \right), \bigvee_{\theta_i} (R_{13}, R_{14}) \right) \quad (\text{B.10})$$

## B.1 Formulae of the Examined Matching Functions

This matching function groups representations into so-called “semantic groups” that fuse texture and color properties, that are color-related, or that model edge and texture properties. In other words, functionally similar representations are grouped. The representations in each group are connected with a disjunction modeling the assumption that they correlate as stated in other studies [Eidenberger 2003; Deselaers et al. 2008] and thus model the same aspects of an image. Hence, it would be enough, if one of the examined representations shares a high degree of similarity with the query in order to assess a document as highly relevant.

Source code: `dbis/weightlearning/evalfunction/SemanticGroupAND.cpp`

**Matching Function 11 Semantic group conjunction, best features:**

$$\bigwedge_{\theta_i} \left( \left( \bigvee_{\theta_j} (R_3, R_9) \right), \left( \bigvee_{\theta_k} (R_2, R_3, R_{10}, R_{12}, R_6, R_1, R_5, R_4, R_7) \right), \left( \bigvee_{\theta_l} (R_8, R_{11}, R_{15}) \right) \right) \quad (\text{B.11})$$

Source code: `dbis/weightlearning/evalfunction/SemBestAND.cpp`

**Matching Function 12 Bielefeld conjunction:**

$$\bigwedge_{\theta_i} (R_1, R_2, R_3, R_4, R_5, R_6, R_7, R_8, R_9, R_{10}, R_{11}, R_{12}, R_{14}, R_{20}, R_{21}, R_{22}, R_{23}, \text{Person}) \quad (\text{B.12})$$

This conjunction is named after the city the domain expert workshops took place in (see Section 6.1.1) and has been used in the user studies. In contrast to all other matching functions, it has not been used in the experiments presented in Chapter 8.

The matching function is multimodal and includes a Boolean predicate for person detection (Person), a spatial proximity predicate for the GPS coordinate ( $R_{22}$ ), and temporal difference predicates ( $R_{20+21}$ ).

The person detection is based on OpenCV using a Haar wavelet-based face detection algorithm<sup>171</sup>. This representation indicates whether one or more persons are present on an image document or not. The similarity between the Camera models ( $R_{23}$ ) is determined with the help of the Levenshtein distance<sup>172</sup>.

Source code: `dbis/weightlearning/evalfunction/BielefeldFeatureConjunction.cpp`

### B.1.2 Matching Functions Based Standard Aggregations

Unlike the matching functions presented before, the following functions constitute arithmetic expressions.

**Matching Function 13 Arithmetic mean:**

$$\frac{1}{n} \sum_{i=1}^n R_i \quad ; \quad n = 15 \quad (\text{B.13})$$

<sup>171</sup>See <http://opencv.willowgarage.com/wiki/FaceDetection> as accessed on August 5th, 2013. For the actual implementation in C++, see class `dbis::extraction::extractors::FaceDetectionExtractor`.

<sup>172</sup>See <http://xlinux.nist.gov/dads//HTML/Levenshtein.html> as accessed on November 18th 2013.

## B Evaluation Appendix

The sum is omitted because it will generate the same total order as the arithmetic mean because the of the multiplication with a constant factor  $\frac{1}{n}$ .

Source code: dbis/weightlearning/evalfunction/ArithmeticMean.cpp

### Matching Function 14 Arithmetic mean, best features:

$$\frac{1}{n} \sum_{i=1}^n R_i \quad ; \quad n = 12 \quad (\text{B.14})$$

The same arguments as stated before (see Matching Function 13) apply.

Source code: dbis/weightlearning/evalfunction/ArithmeticMeanBest.cpp

### Matching Function 15 Weighted arithmetic mean, best features:

$$\frac{1}{n} \sum_{i=1}^n \theta_i R_i \quad ; \quad n = 12, \quad \sum_{i=1}^n \theta_i = 1 \quad (\text{B.15})$$

This function extends Matching Function 14 with representation-specific weighting variables  $\theta_i$ .

Source code: dbis/weightlearning/evalfunction/WeightedArithmeticMeanBest.cpp

### Matching Function 16 Geometric mean:

$$\sqrt[n]{\prod_{i=1}^n R_i} \quad ; \quad n = 15 \quad (\text{B.16})$$

A CQQL conjunction (see Matching Function 1) produces the same order as the geometric mean because of its arithmetic evaluation that differs only from the geometric mean because of the missing root function. The root function is strictly monotonic and has therefore no effect on the produced total order.

Source code: dbis/weightlearning/evalfunction/GeometricMean.cpp

### Matching Function 17 Geometric mean, best features:

$$\sqrt[n]{\prod_{i=1}^n R_i} \quad ; \quad n = 12 \quad (\text{B.17})$$

The same arguments as stated before (see Matching Function 16) apply.

Source code: dbis/weightlearning/evalfunction/GeometricMeanBest.cpp

### Matching Function 18 Max:

$$\max(R_1, \dots, R_n) \quad ; \quad n = 15 \quad (\text{B.18})$$

The maximum represents the disjunction in the original publication of fuzzy logic [Zadeh 1988].

Source code: dbis/weightlearning/evalfunction/MaxCondition.cpp

## B.1 Formulae of the Examined Matching Functions

Matching Function 19 **Max, best features:**

$$\max(R_1, \dots, R_n) \quad ; n = 12 \quad (\text{B.19})$$

Source code: dbis/weightlearning/evalfunction/MaxConditionBest.cpp

Matching Function 20 **Min:**

$$\min(R_1, \dots, R_n) \quad ; n = 15 \quad (\text{B.20})$$

*The minimum represents the conjunction in the original publication of fuzzy logic [Zadeh 1988].*

Source code: dbis/weightlearning/evalfunction/MinCondition.cpp

Matching Function 21 **Min, best features:**

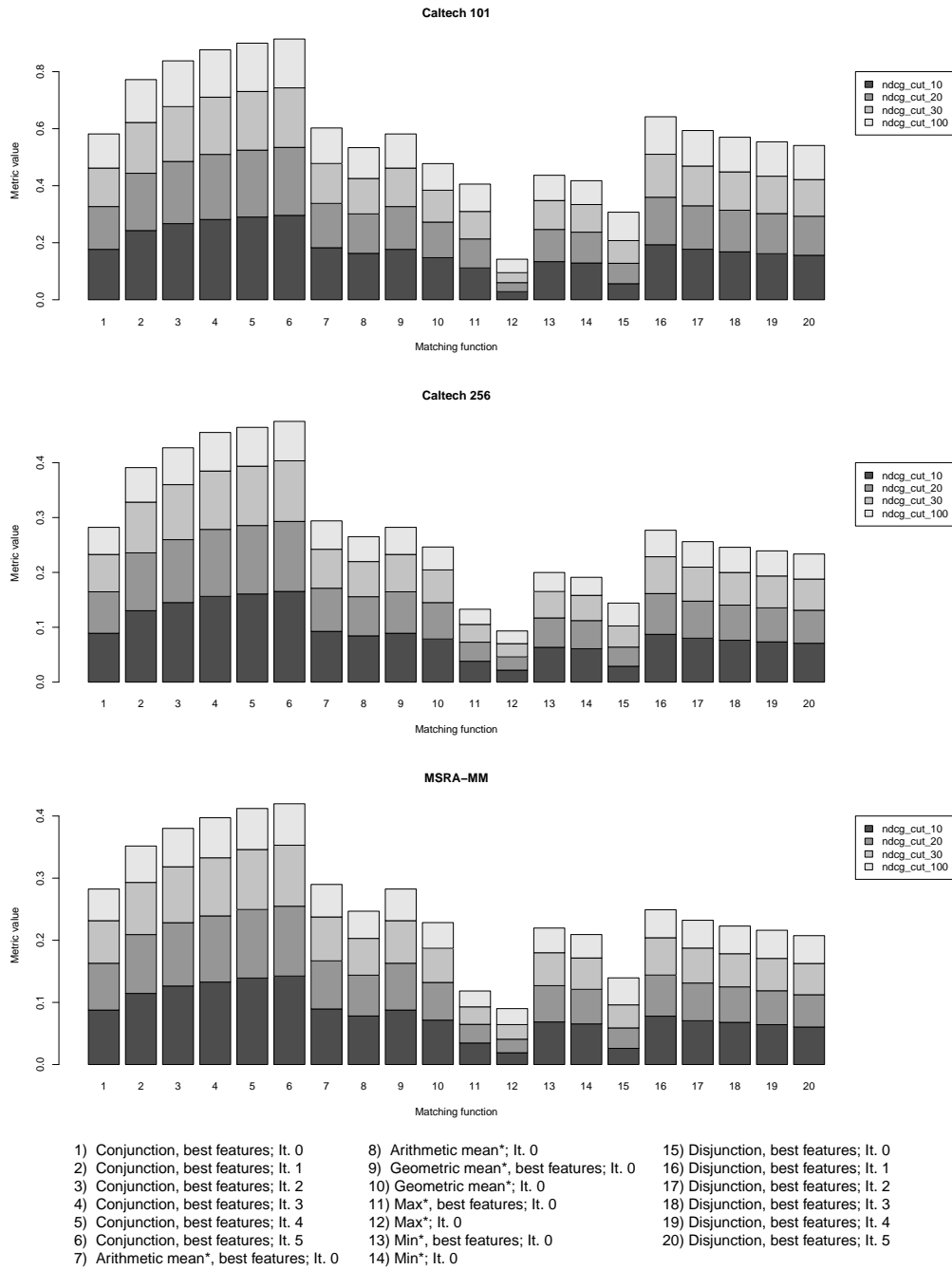
$$\min(R_1, \dots, R_n) \quad ; n = 12 \quad (\text{B.21})$$

Source code: dbis/weightlearning/evalfunction/MinConditionBest.cpp

## **B.2 Retrieval Effectiveness Comparisons of Different Matching Functions**



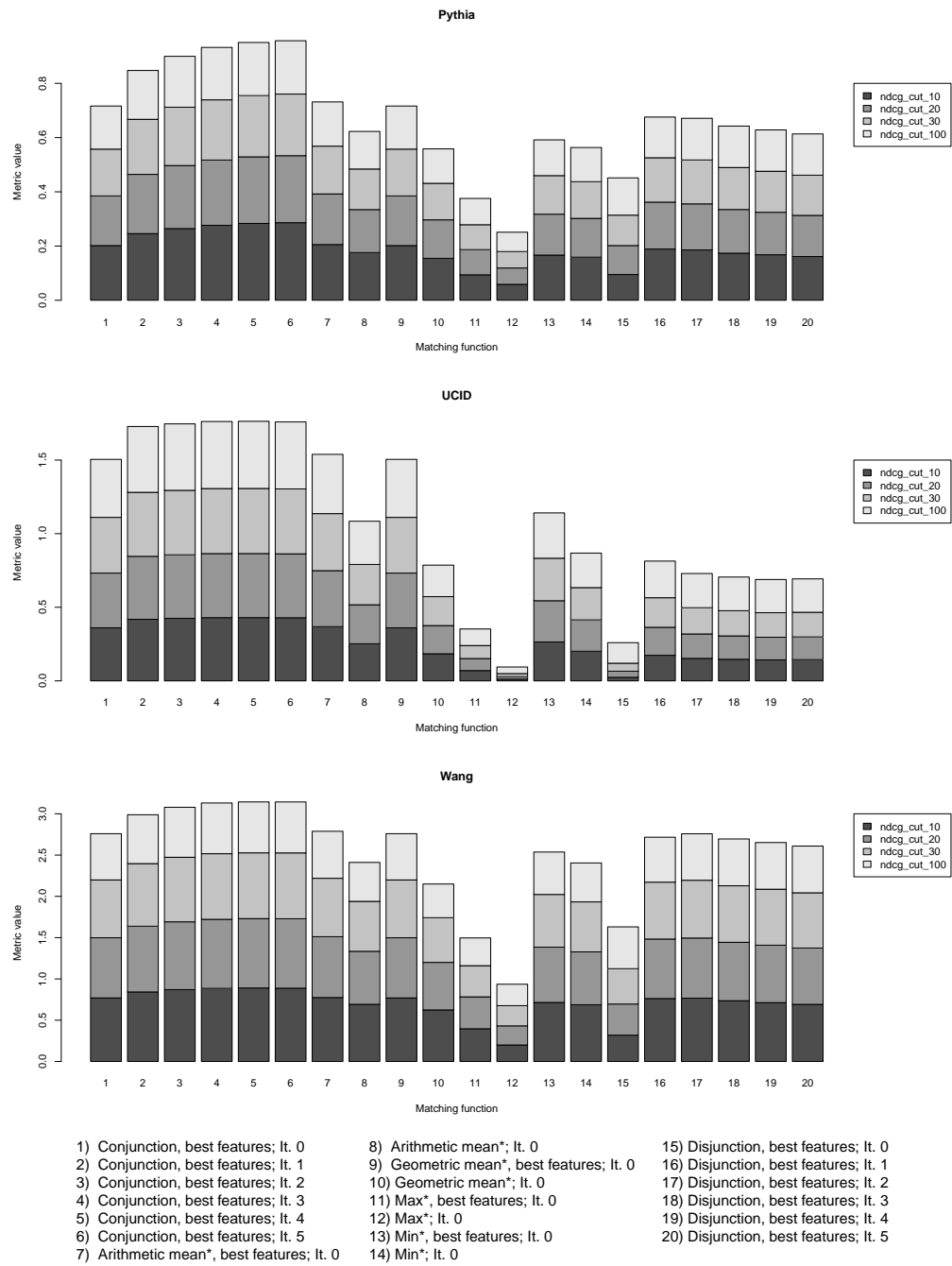
## B.2 Retrieval Effectiveness Comparisons of Different Matching Functions



\* indicates standard aggregations.

Figure B.2: RF performance comparison of characteristic matching functions and standard aggregations, part 1

## B Evaluation Appendix



\* indicates standard aggregations.

Figure B.3: RF performance comparison of characteristic matching functions and standard aggregations, part 2

## B.2 Retrieval Effectiveness Comparisons of Different Matching Functions

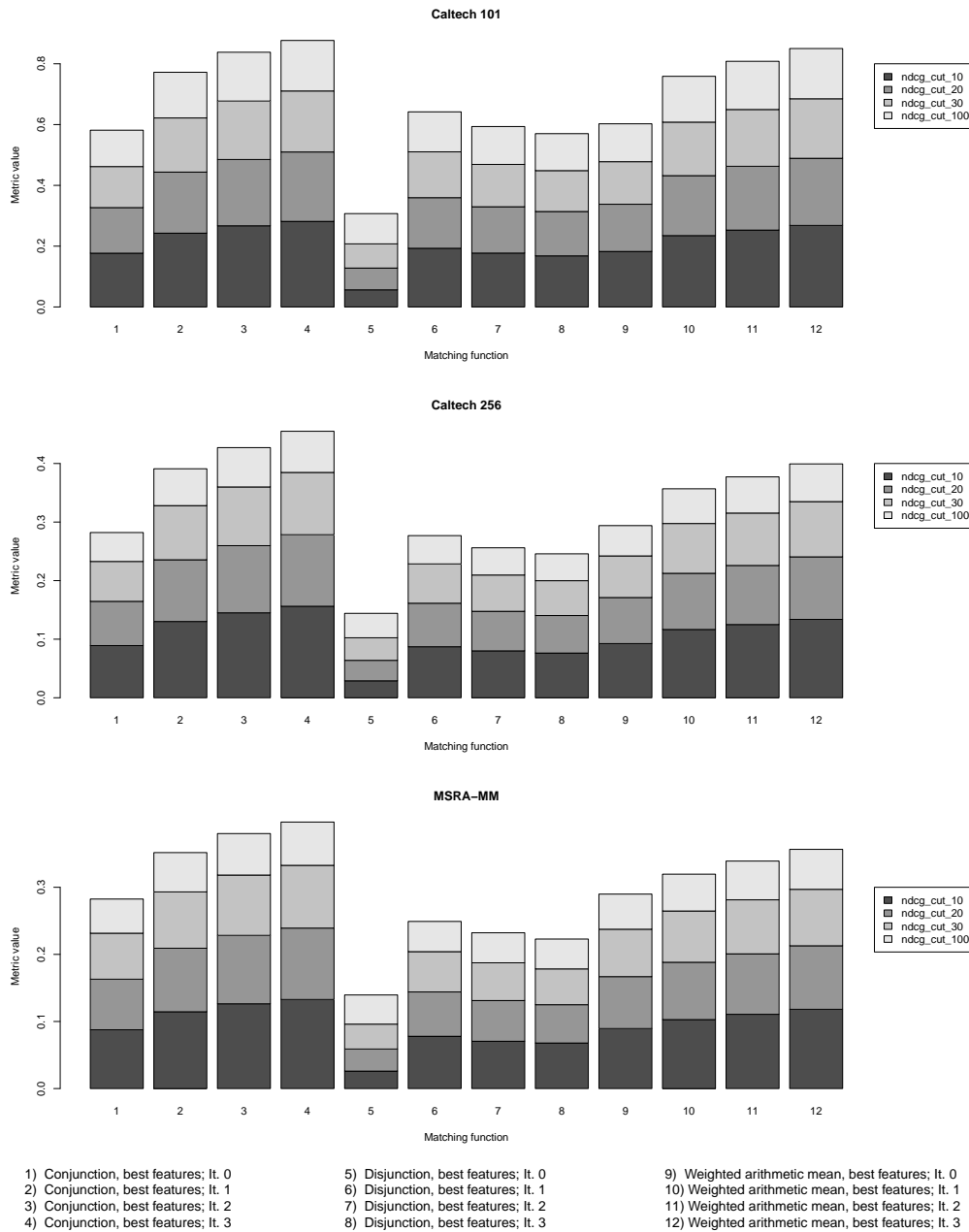


Figure B.4: RF performance comparison of conjunction, disjunction and weighted arithmetic mean, part 1

## B Evaluation Appendix

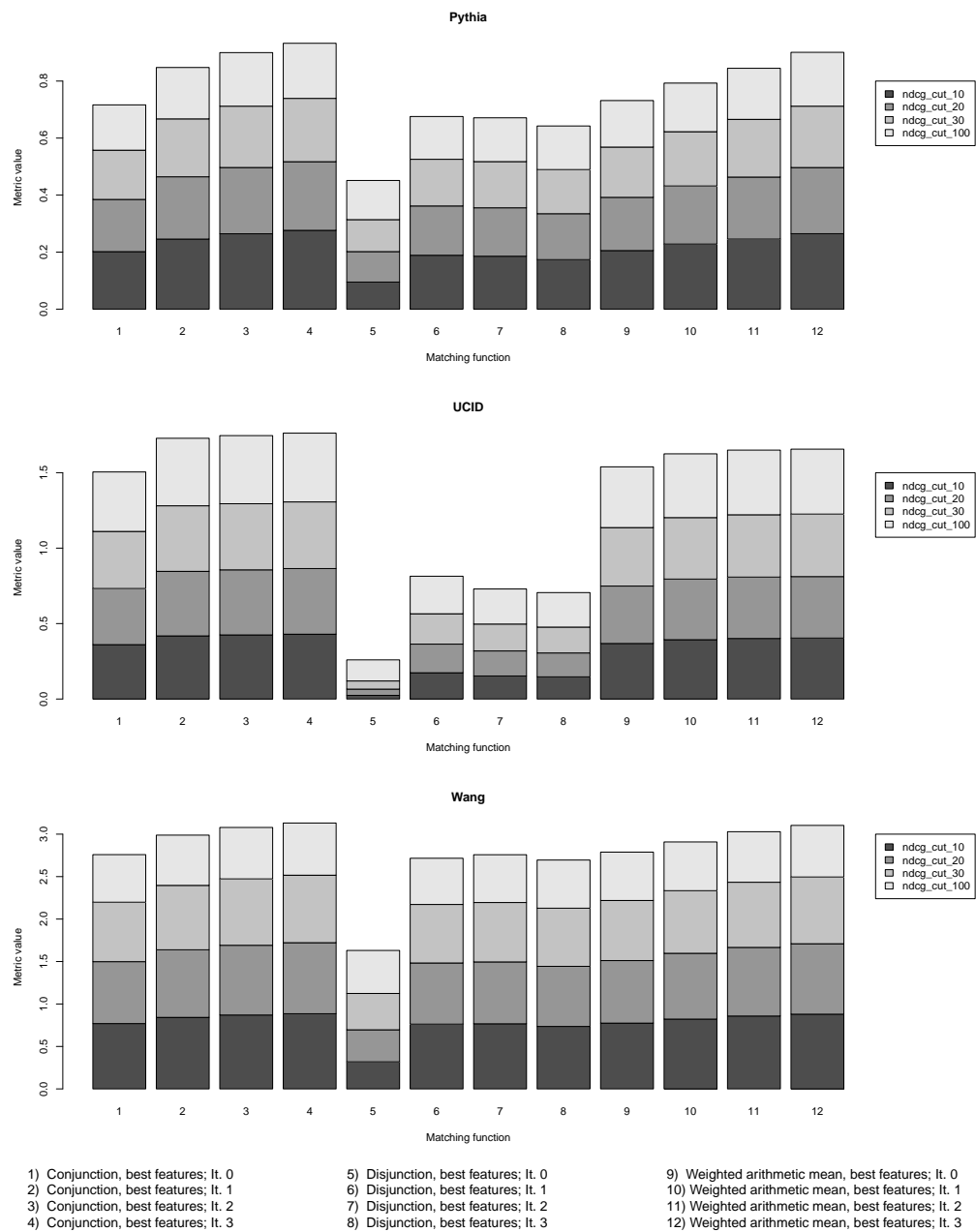


Figure B.5: RF performance comparison of conjunction, disjunction and weighted arithmetic mean, part 2

## B.2 Retrieval Effectiveness Comparisons of Different Matching Functions

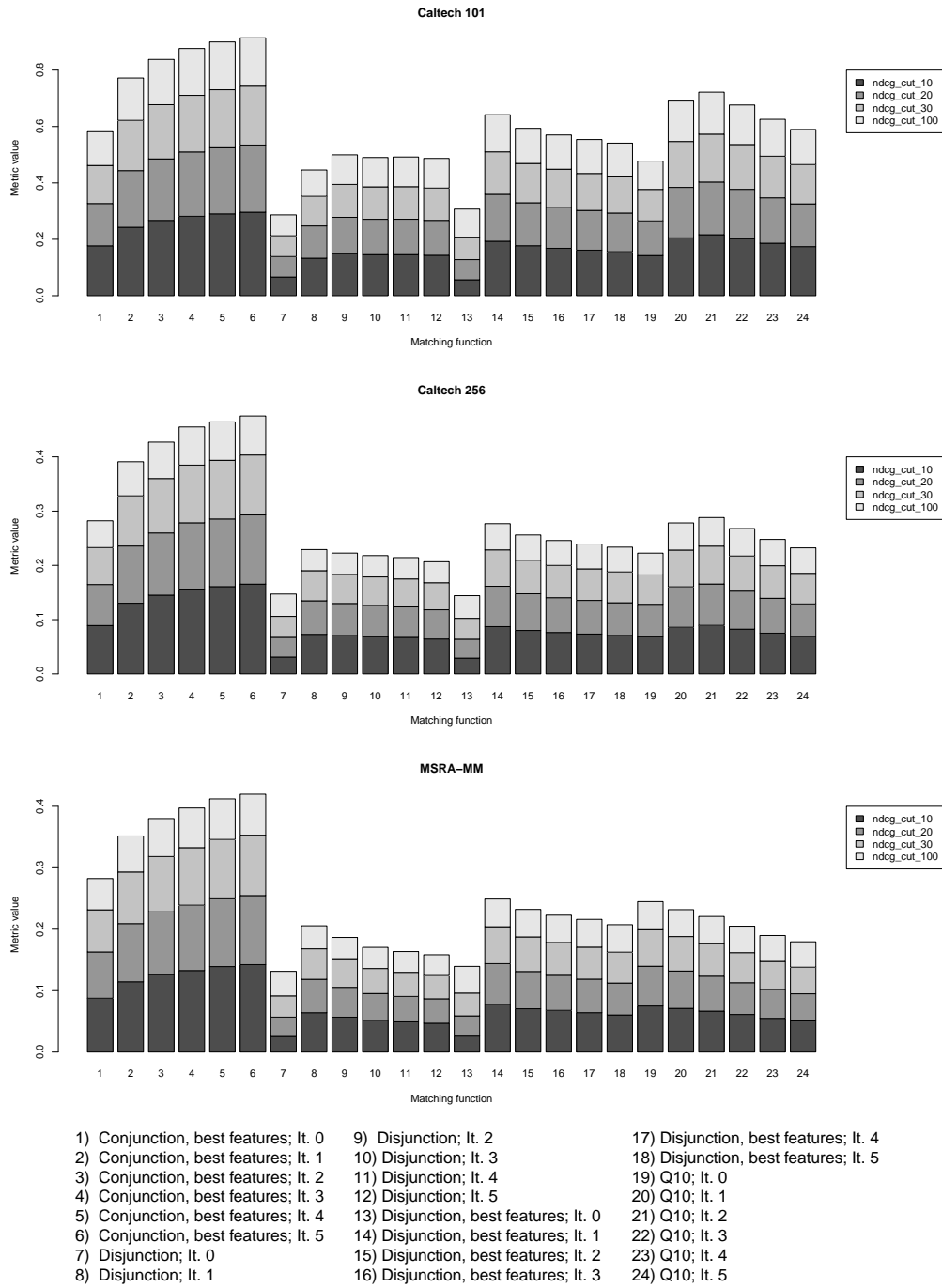


Figure B.6: RF performance comparison of characteristic matching functions (conjunctions, disjunctions, and Q10), part 1

## B Evaluation Appendix

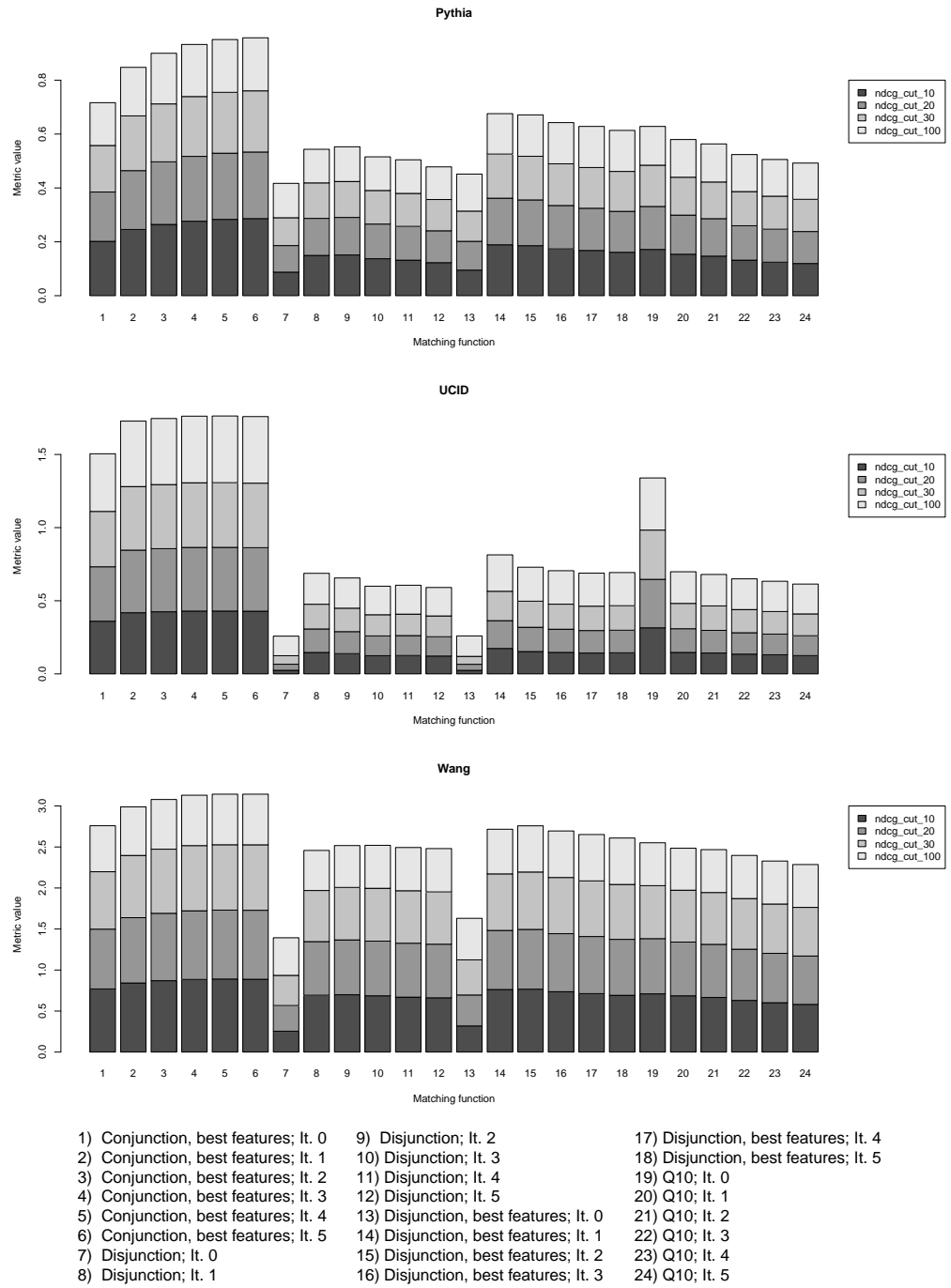


Figure B.7: RF performance comparison of characteristic matching functions (conjunctions, disjunctions, and Q10), part 2

## B.2 Retrieval Effectiveness Comparisons of Different Matching Functions

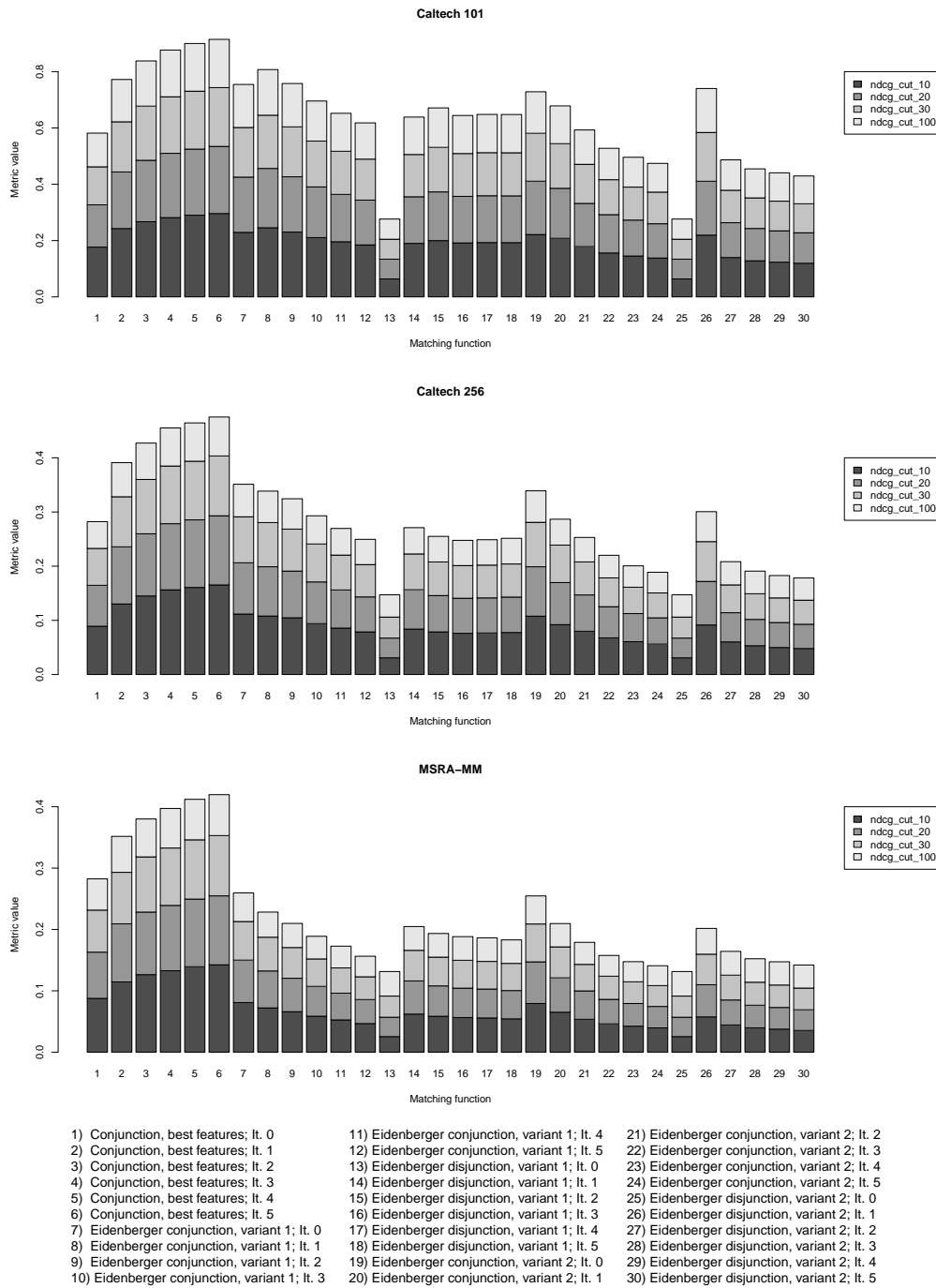


Figure B.8: RF performance comparison of conjunction and Eidenberger variants, part

1

## B Evaluation Appendix

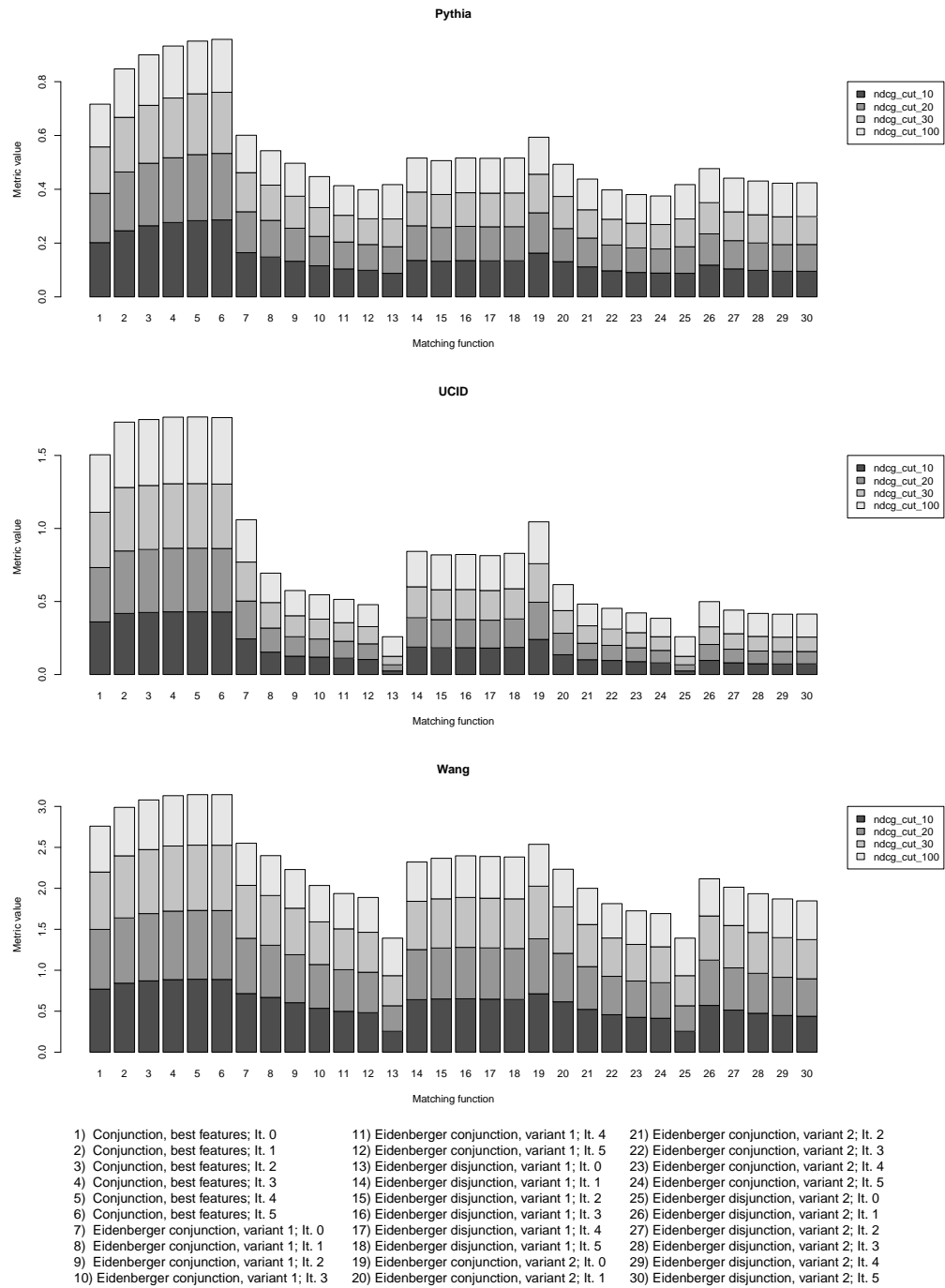


Figure B.9: RF performance comparison of conjunction and Eidenberger variants, part 2



## B.2 Retrieval Effectiveness Comparisons of Different Matching Functions

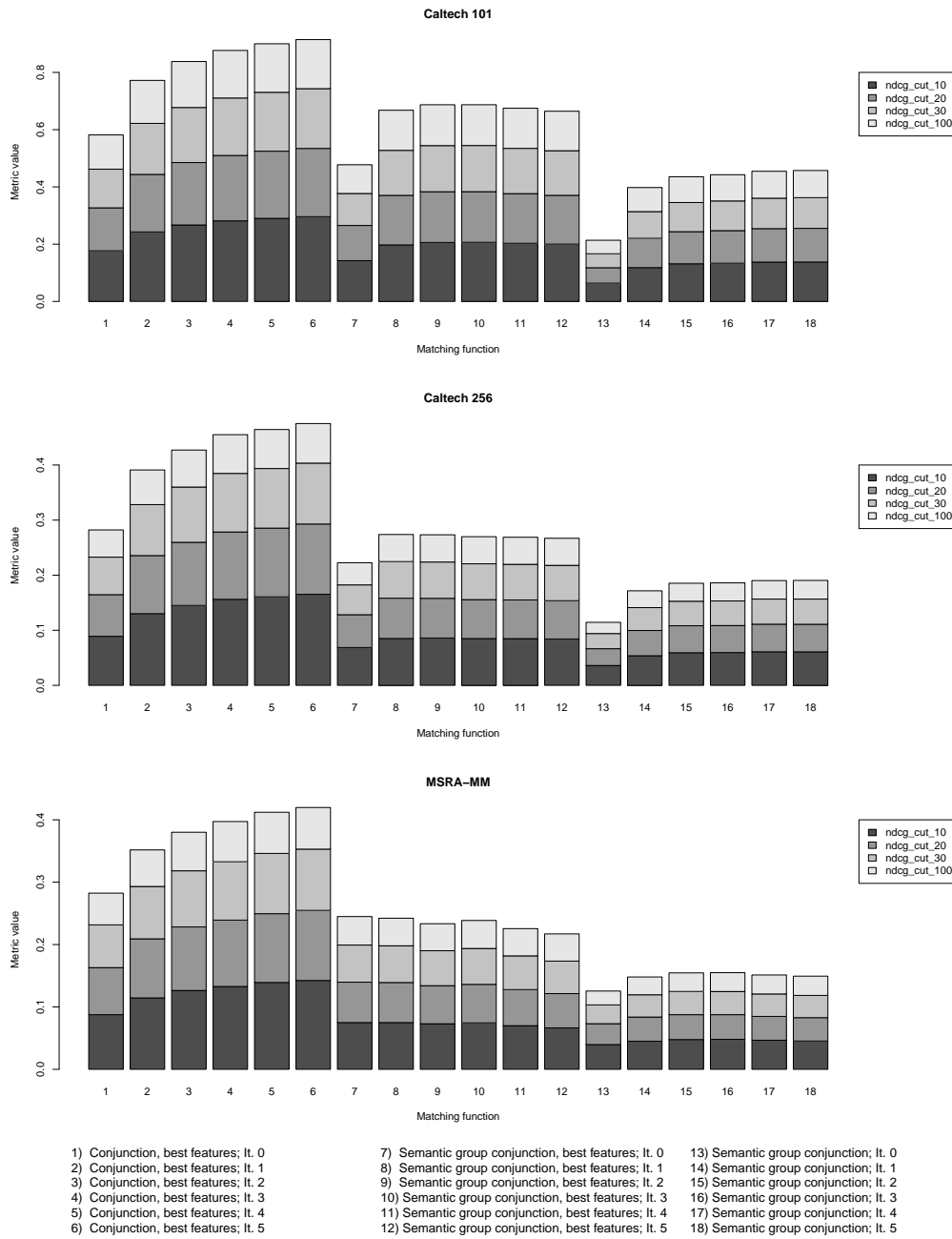


Figure B.10: RF performance comparison of conjunction and semantic group conjunctions, part 1

## B Evaluation Appendix

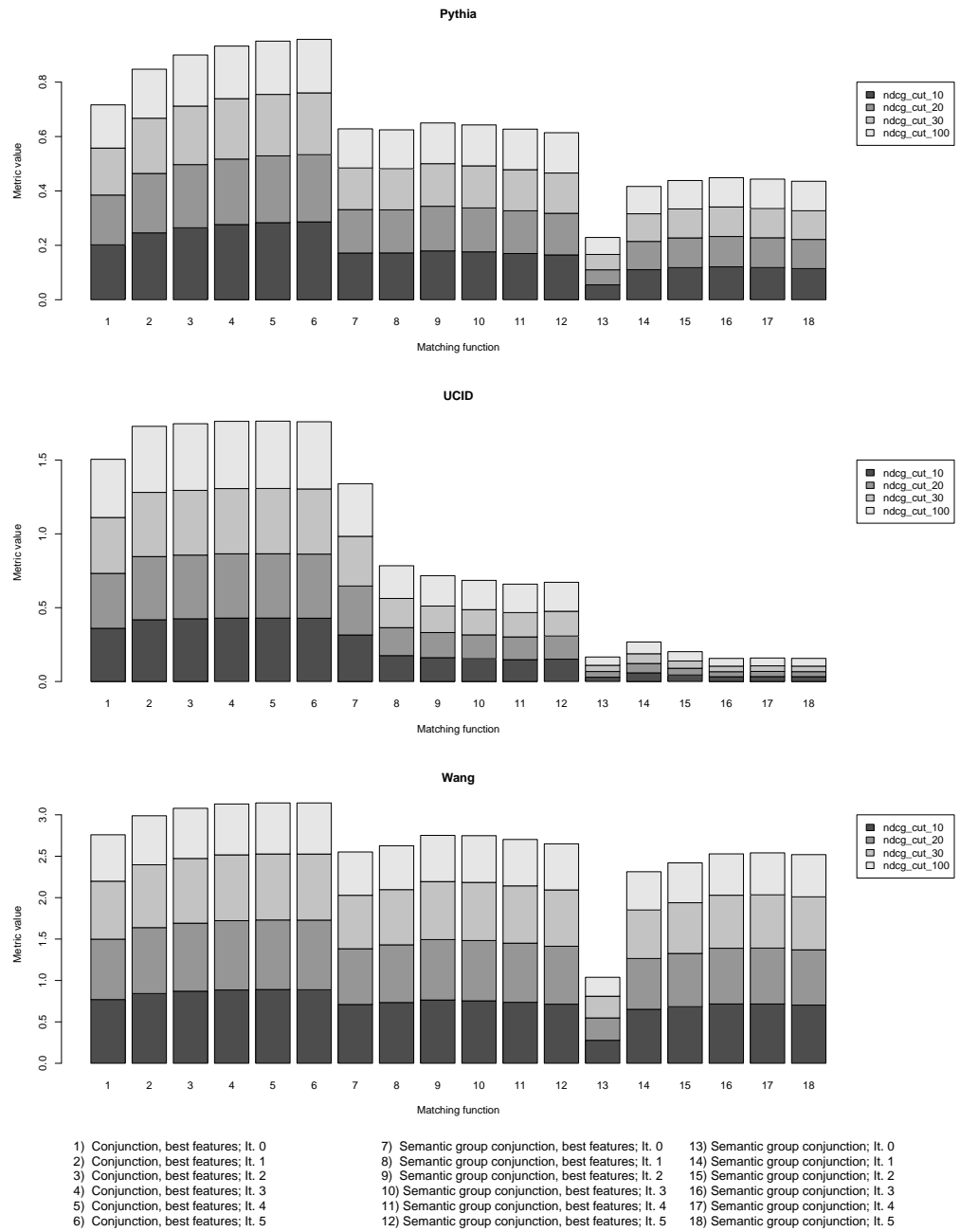


Figure B.11: RF performance comparison of conjunction and semantic group conjunctions, part 2

## B.3 Weight Development of Different Matching Functions

### B.3.1 Averaged Weighting Variable Development per Collection

For the weight variable indices, see Appendix B.1.

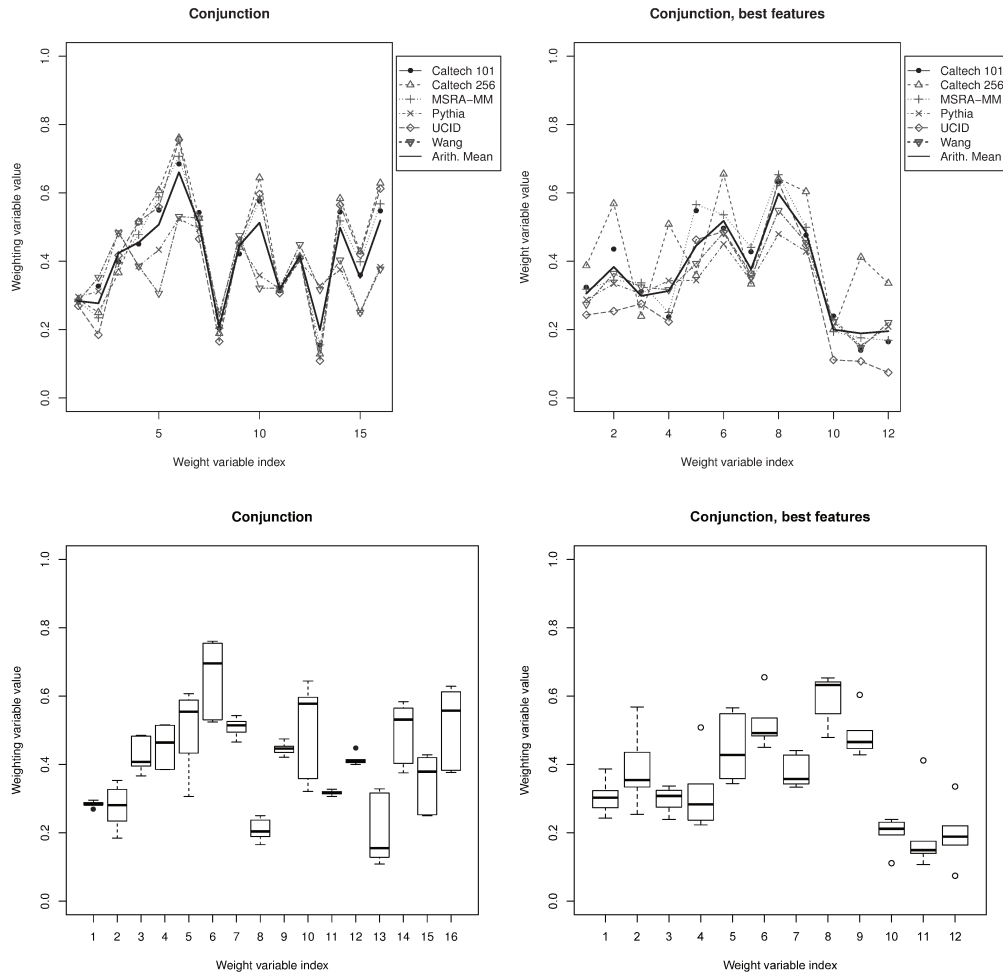


Figure B.12: Averaged weight variable development per collection and averaged over all collections for conjunction and best feature conjunction

## B Evaluation Appendix

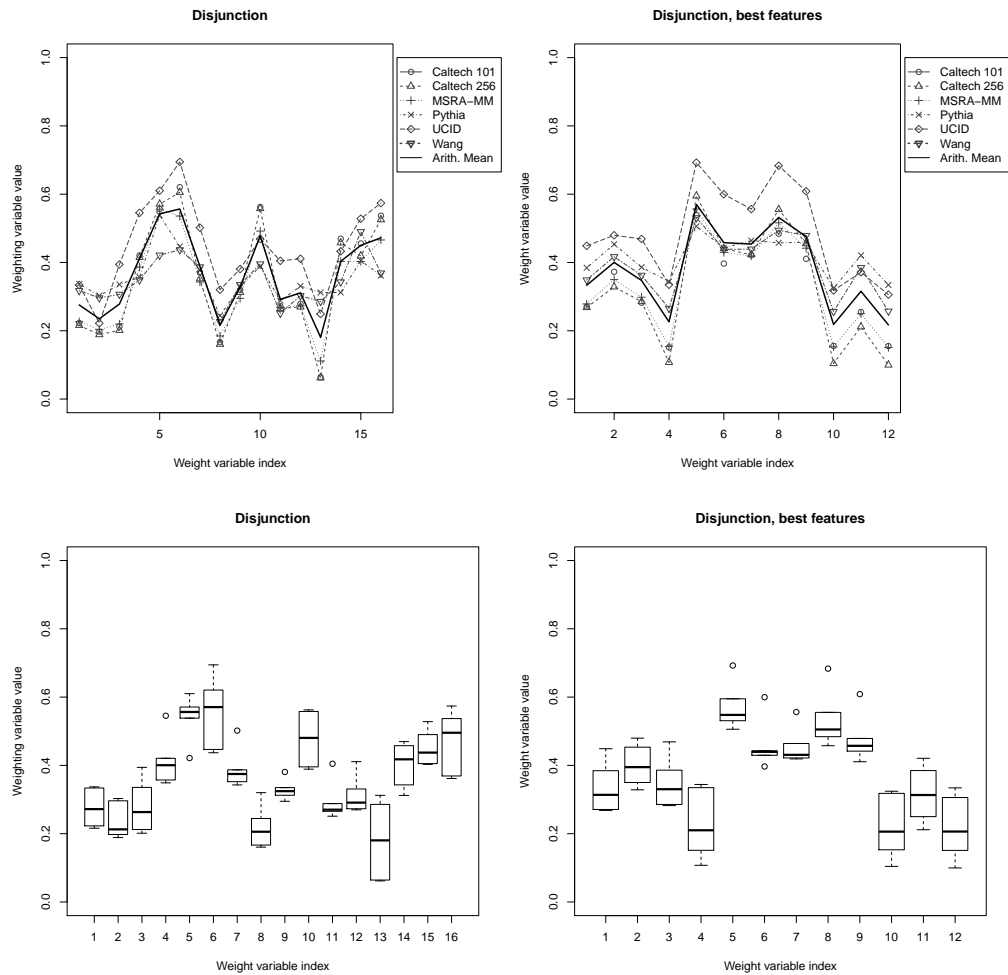


Figure B.13: Averaged weight variable development per collection and averaged over all collections for disjunction and best feature disjunction

### B.3 Weight Development of Different Matching Functions

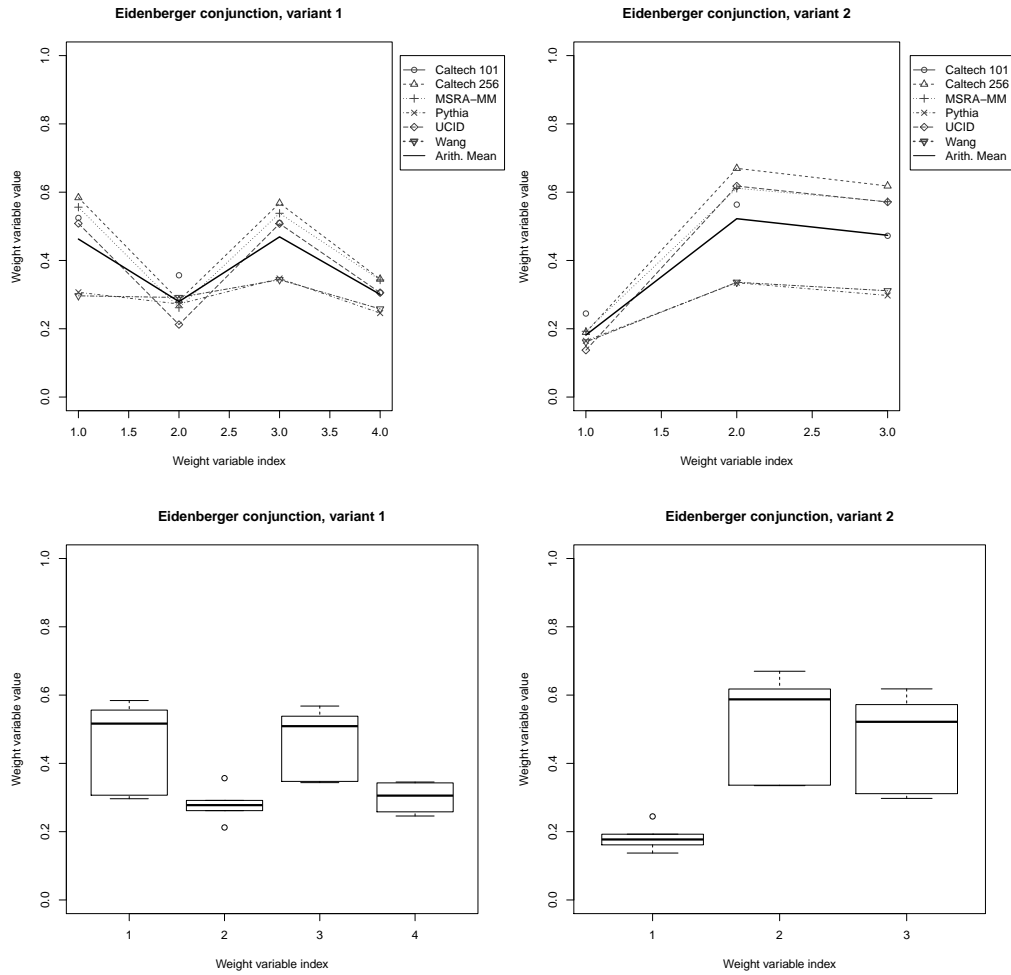


Figure B.14: Averaged weight variable development per collection and averaged over all collections for Eidenberger conjunction 1 and conjunction 2

## B Evaluation Appendix

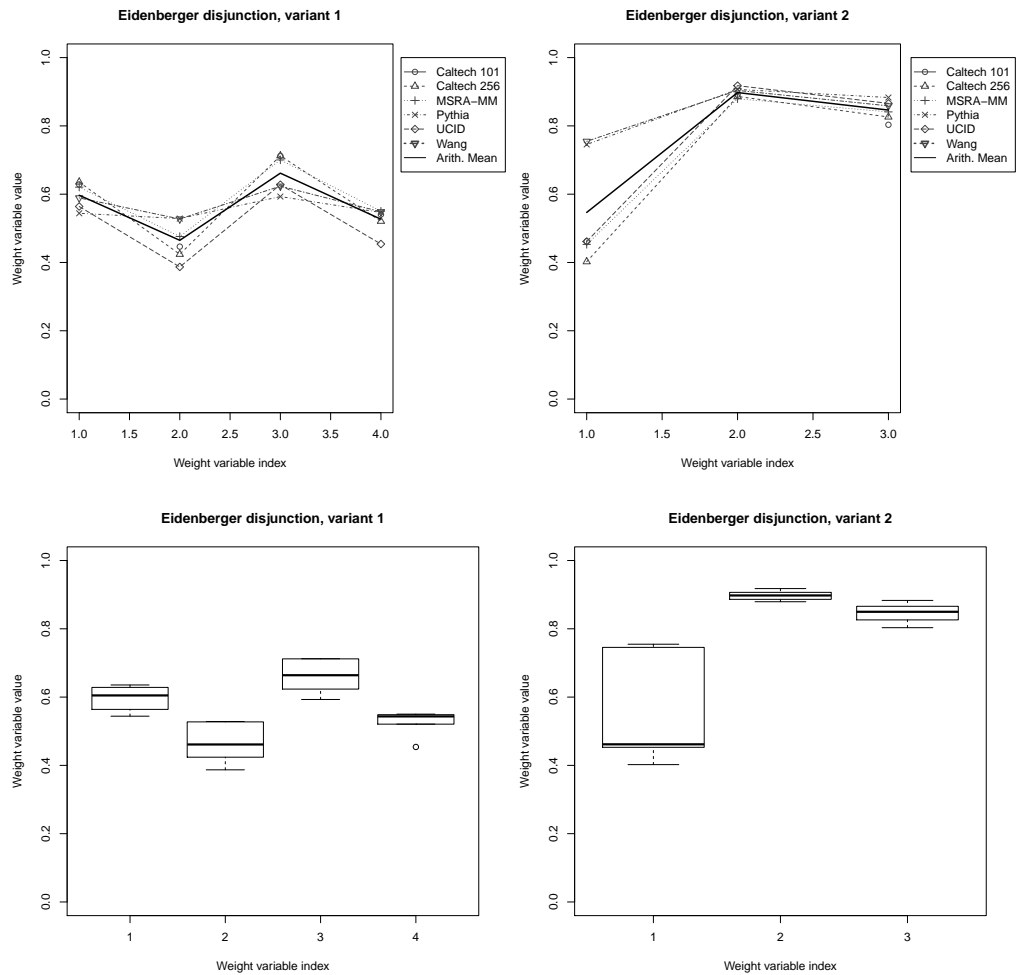


Figure B.15: Averaged weight variable development per collection and averaged over all collections for Eidenberger disjunction 1 and disjunction 2

### B.3 Weight Development of Different Matching Functions

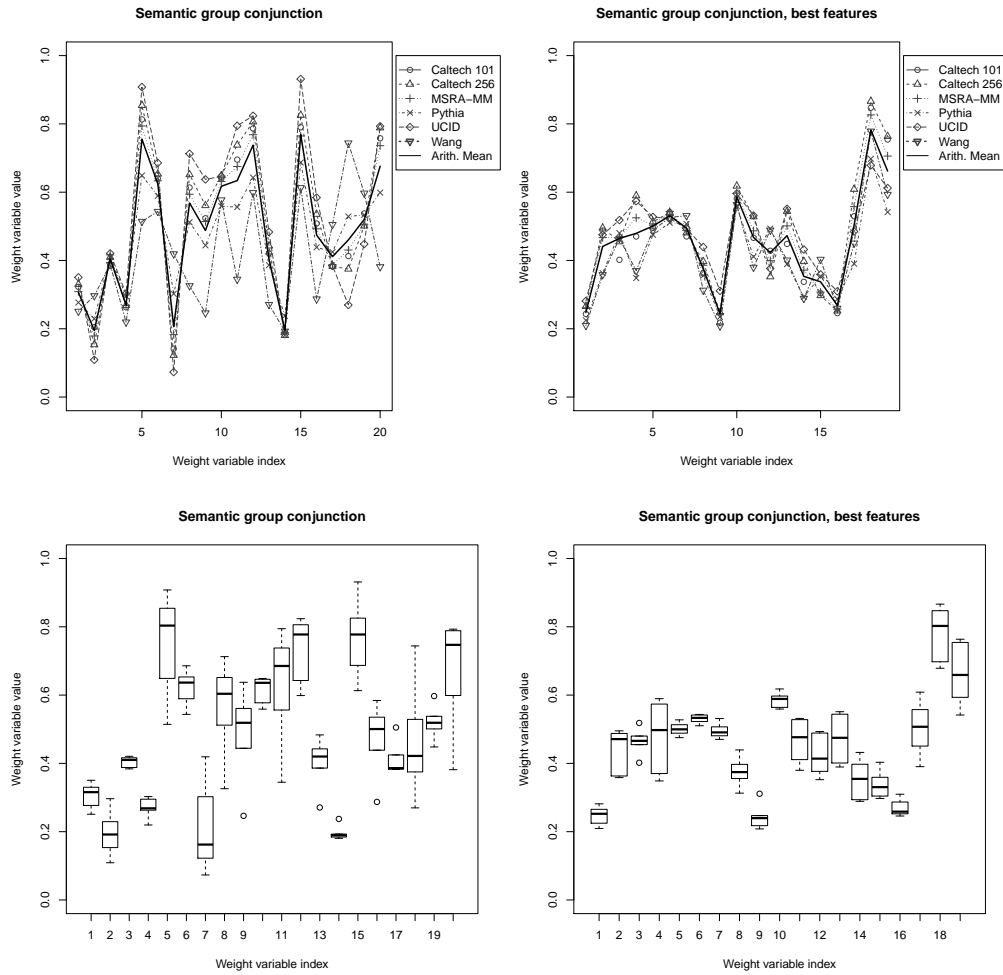


Figure B.16: Averaged weight variable development per collection and averaged over all collections for semantic group conjunction and best feature variant

## B Evaluation Appendix

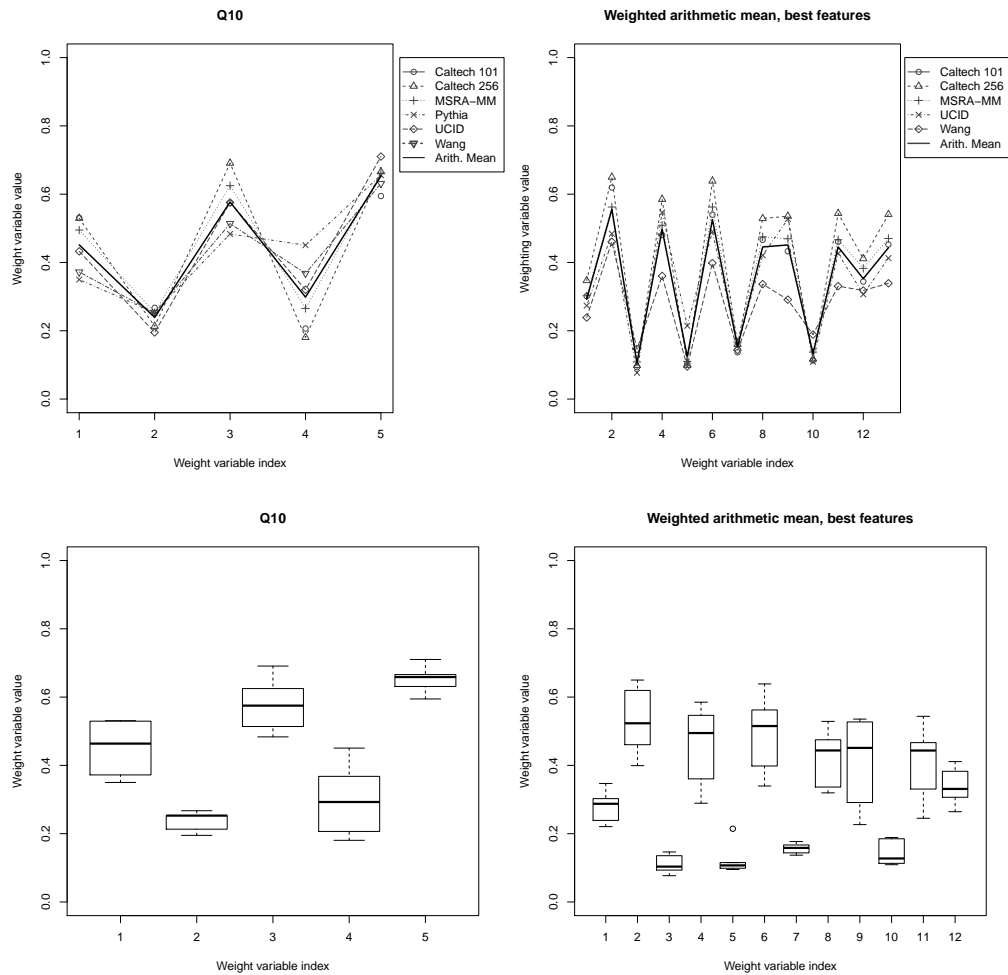


Figure B.17: Averaged weight variable development per collection and averaged over all collections for Q10 and weighted average (best features variant)



### B.3.2 Weighting Variable Development and Distribution during Relevance Feedback

For the weight variable indices, see Appendix B.1.

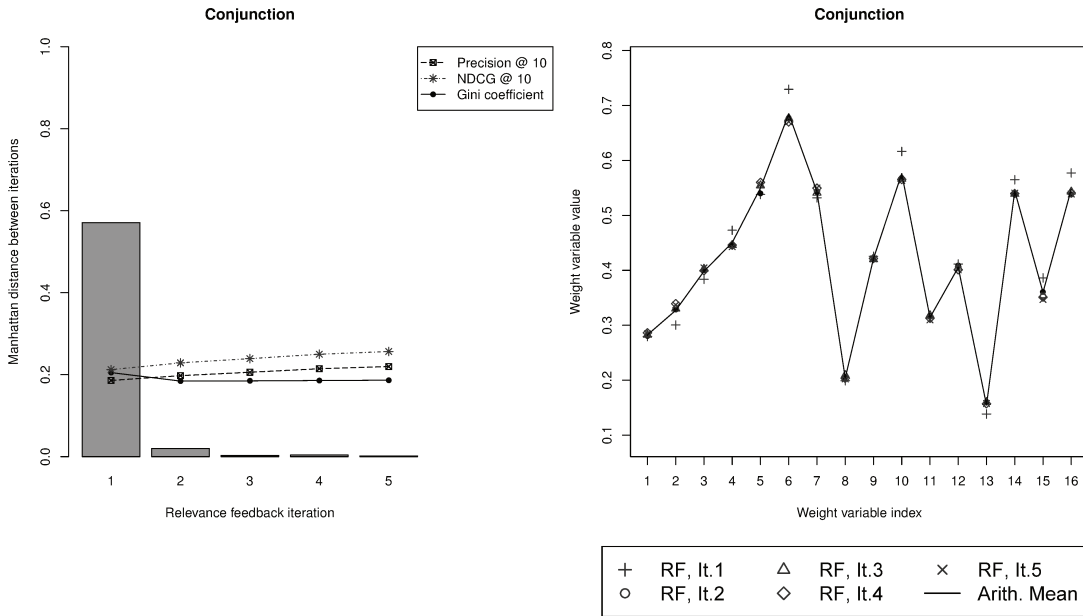


Figure B.18: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 101*

## B Evaluation Appendix

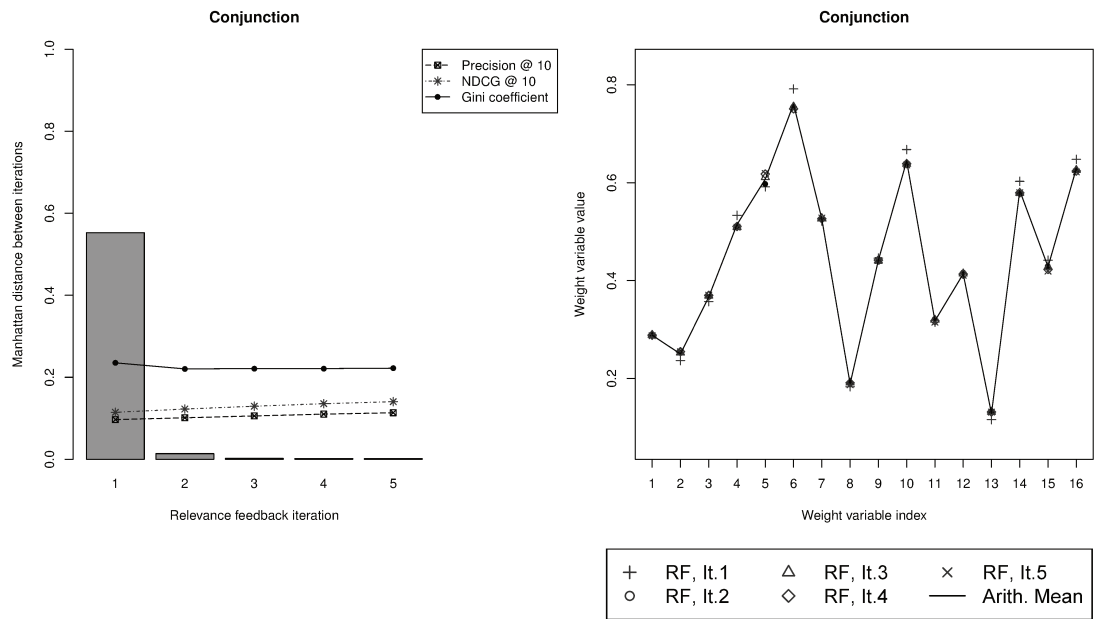


Figure B.19: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 256*

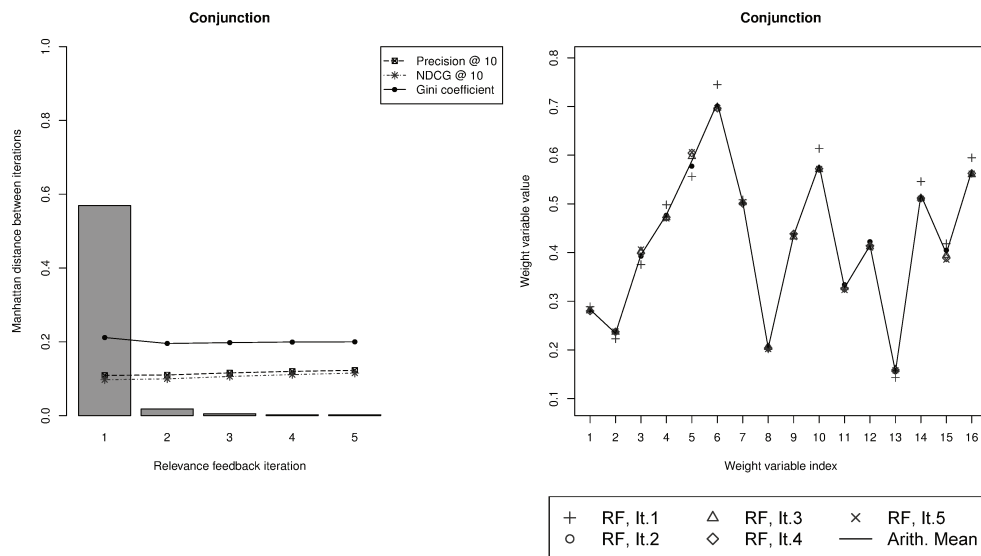


Figure B.20: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *MSRA-MM*

### B.3 Weight Development of Different Matching Functions

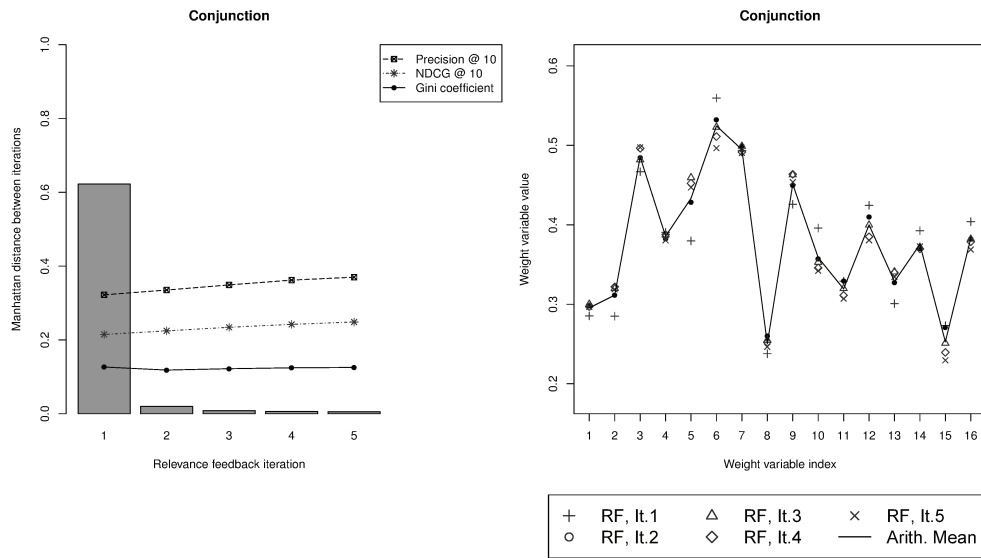


Figure B.21: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Pythia*

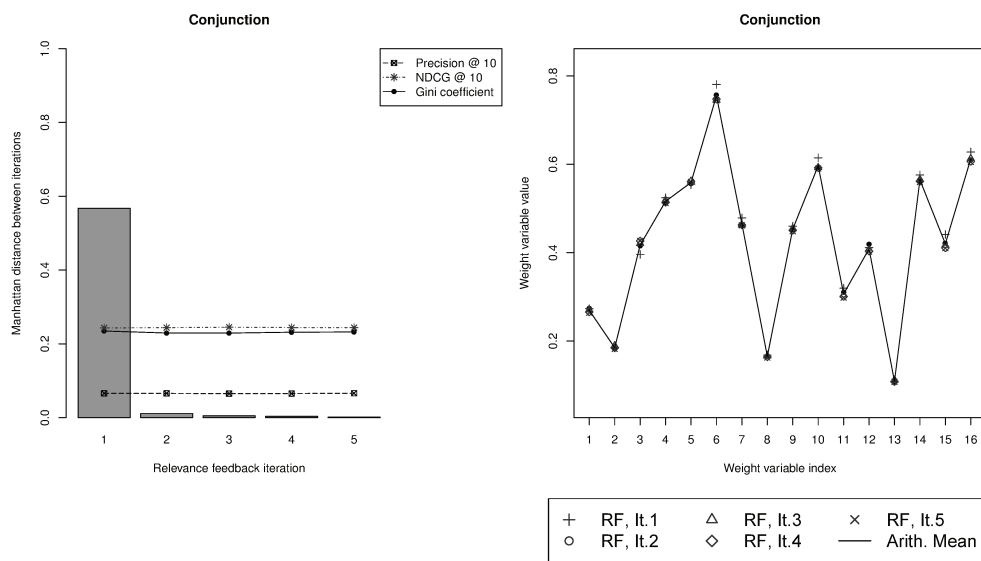


Figure B.22: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *UCID*

## B Evaluation Appendix

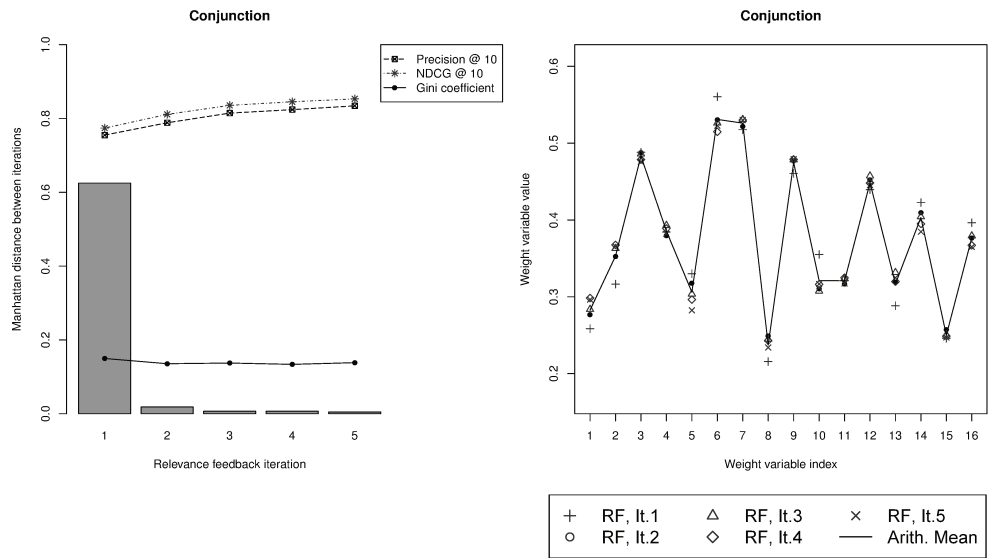


Figure B.23: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Wang*

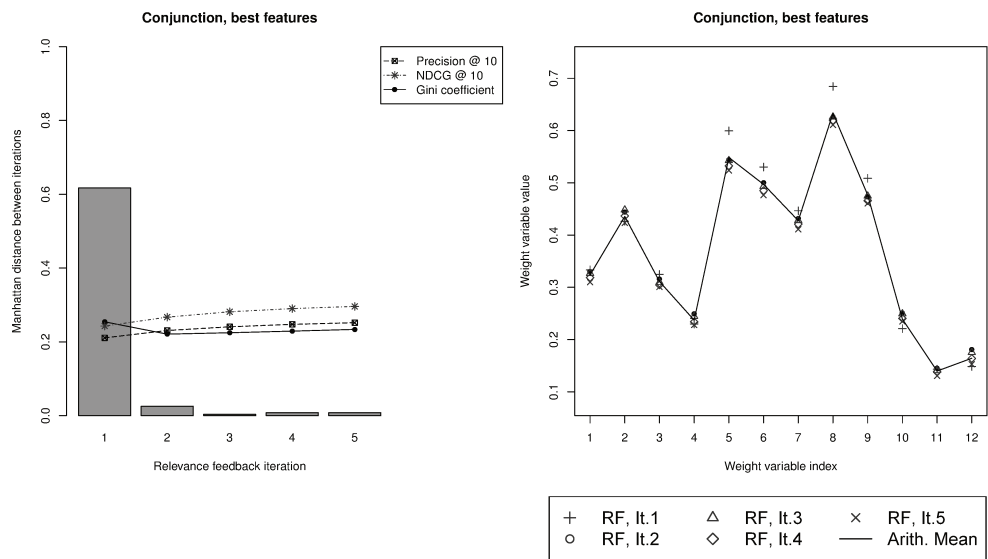


Figure B.24: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 101*

### B.3 Weight Development of Different Matching Functions

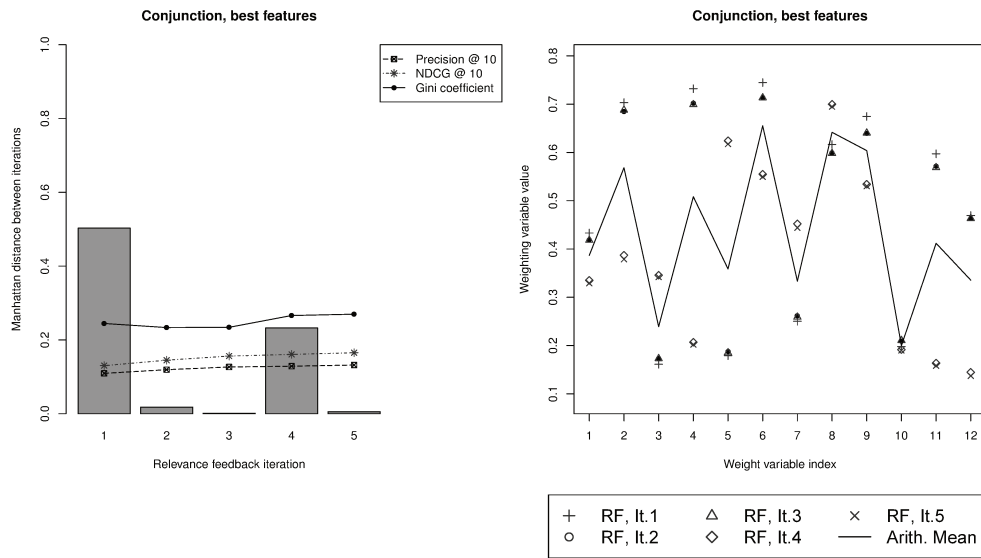


Figure B.25: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 256*

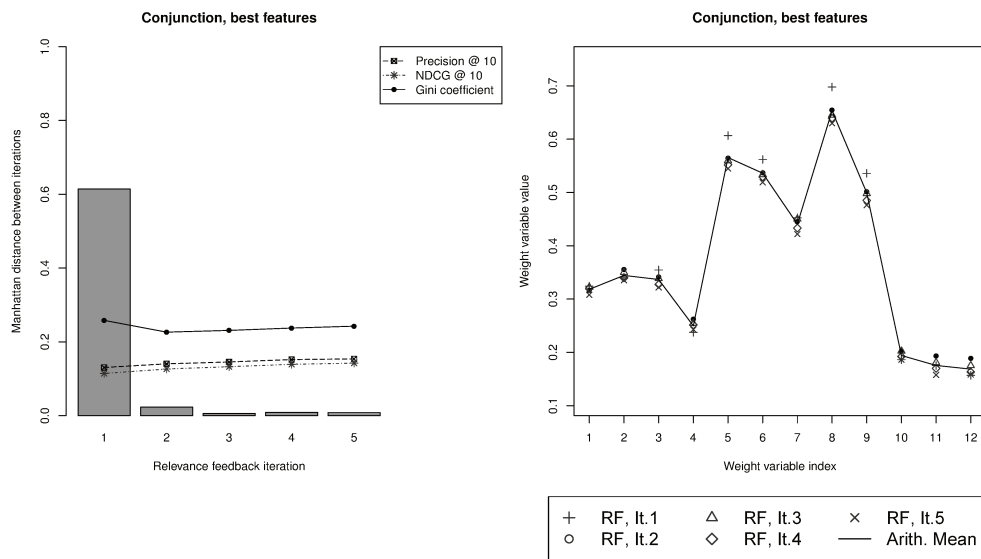


Figure B.26: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *MSRA-MM*

## B Evaluation Appendix

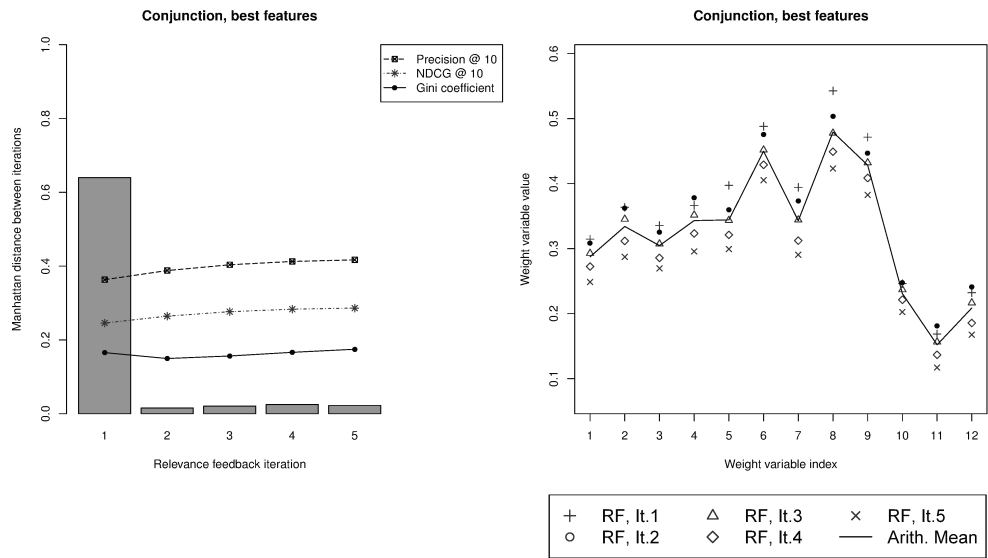


Figure B.27: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Pythia*

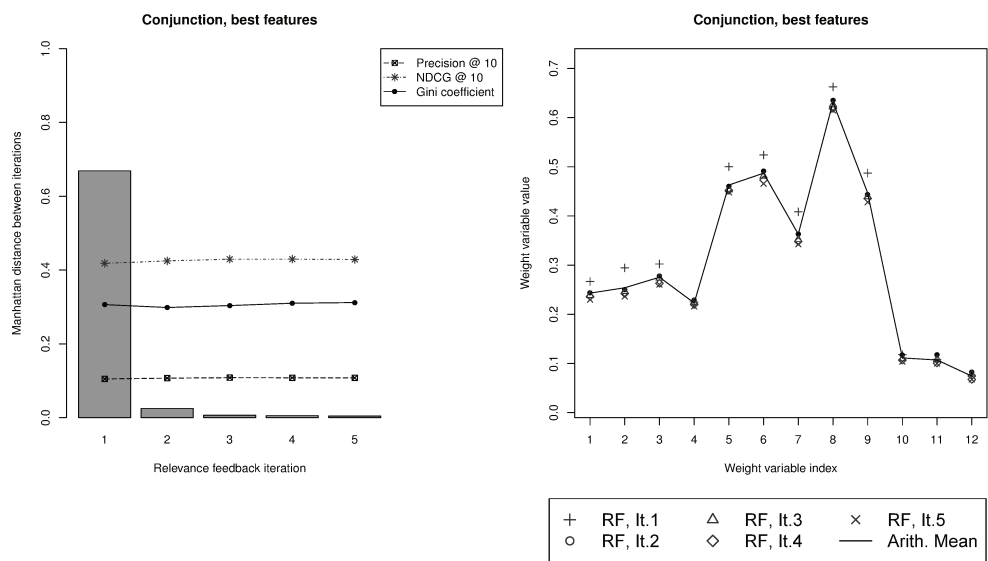


Figure B.28: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *UCID*

### B.3 Weight Development of Different Matching Functions

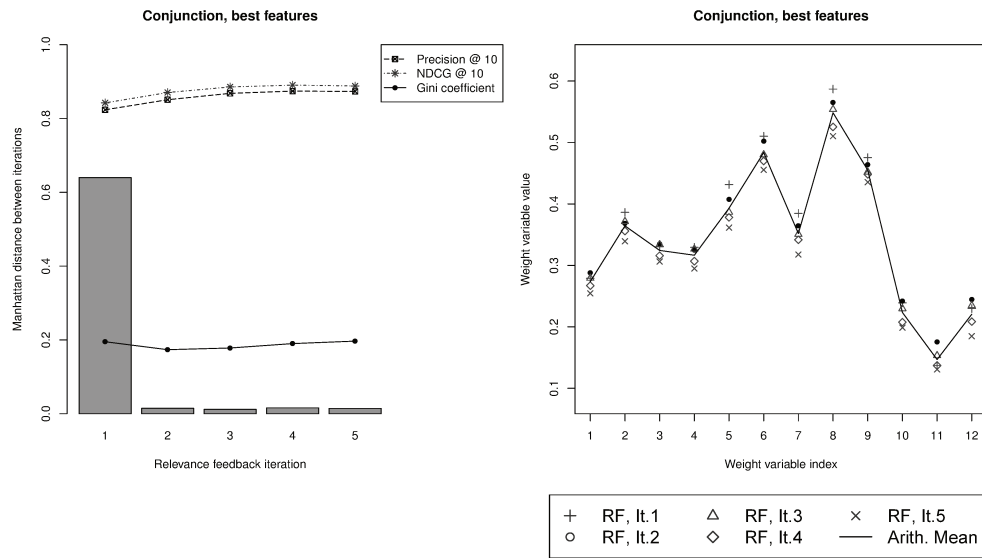


Figure B.29: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Wang*

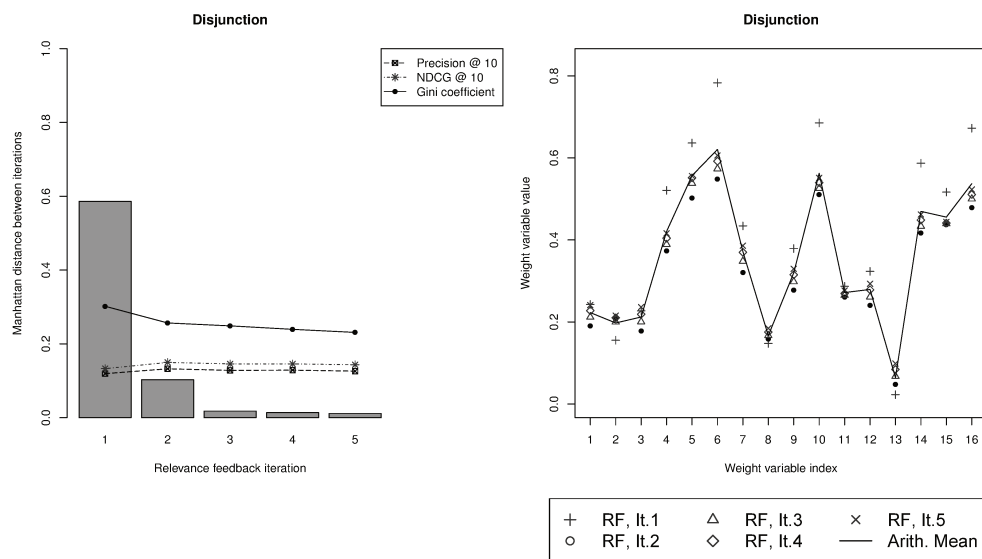


Figure B.30: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 101*

## B Evaluation Appendix

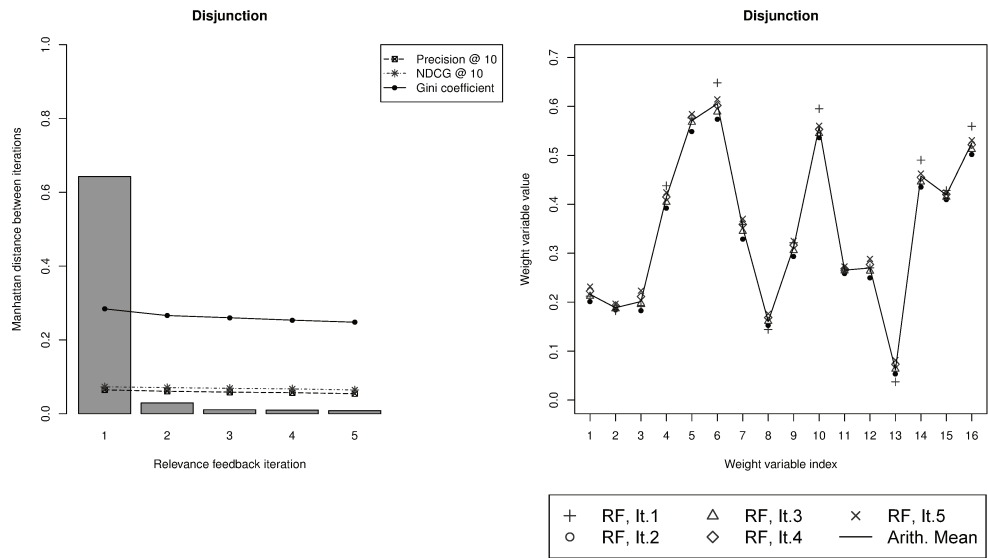


Figure B.31: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 256*

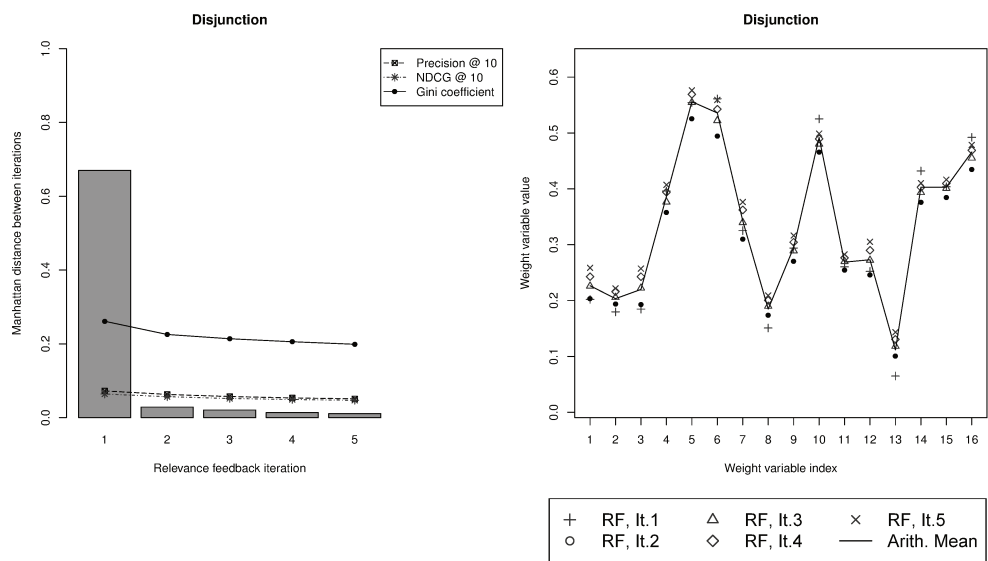


Figure B.32: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *MSRA-MM*



### B.3 Weight Development of Different Matching Functions

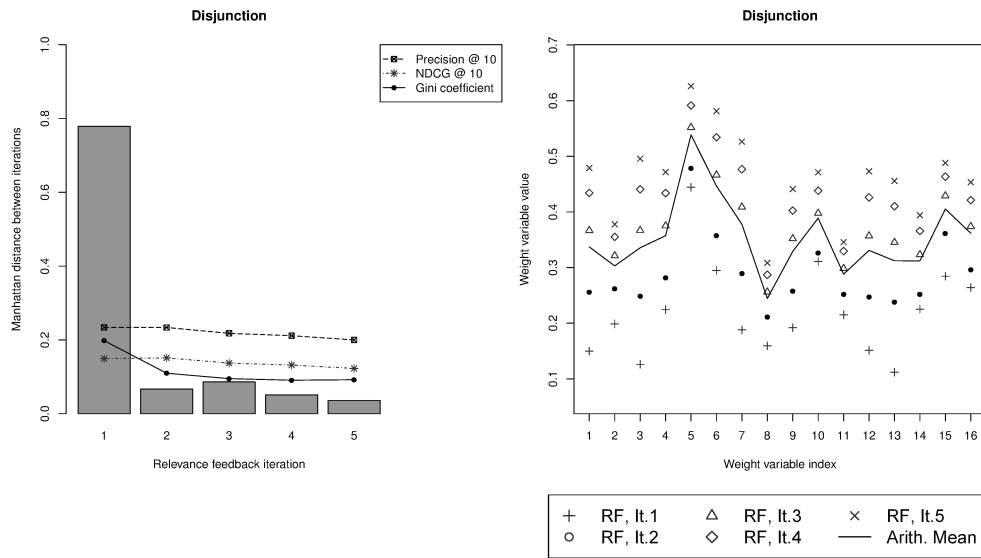


Figure B.33: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Pythia*

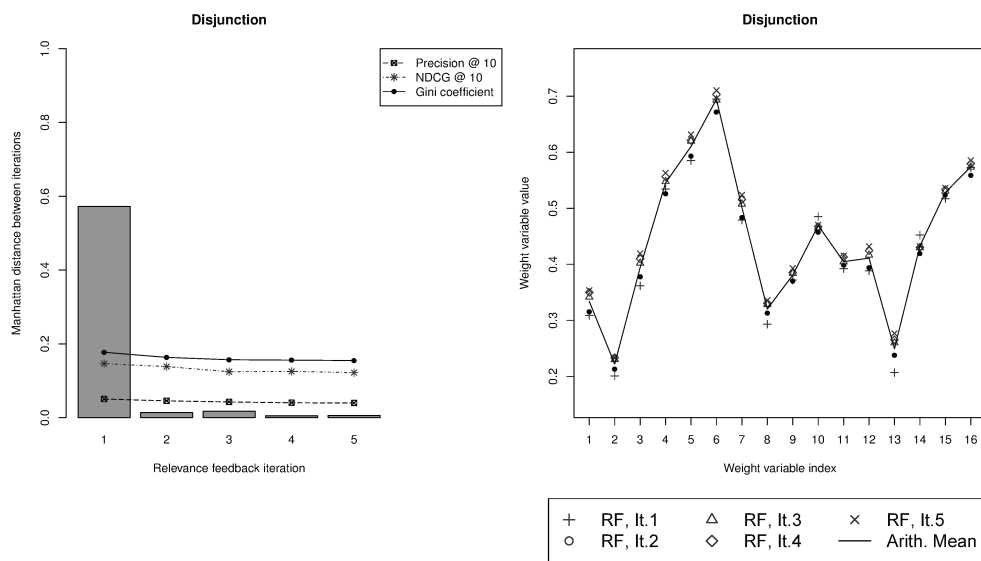


Figure B.34: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *UCID*

## B Evaluation Appendix

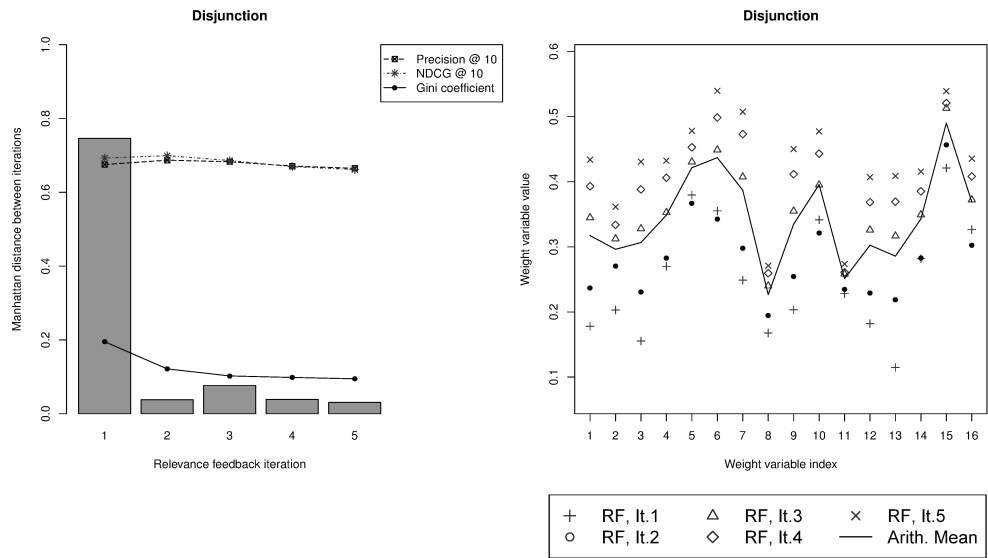


Figure B.35: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Wang*

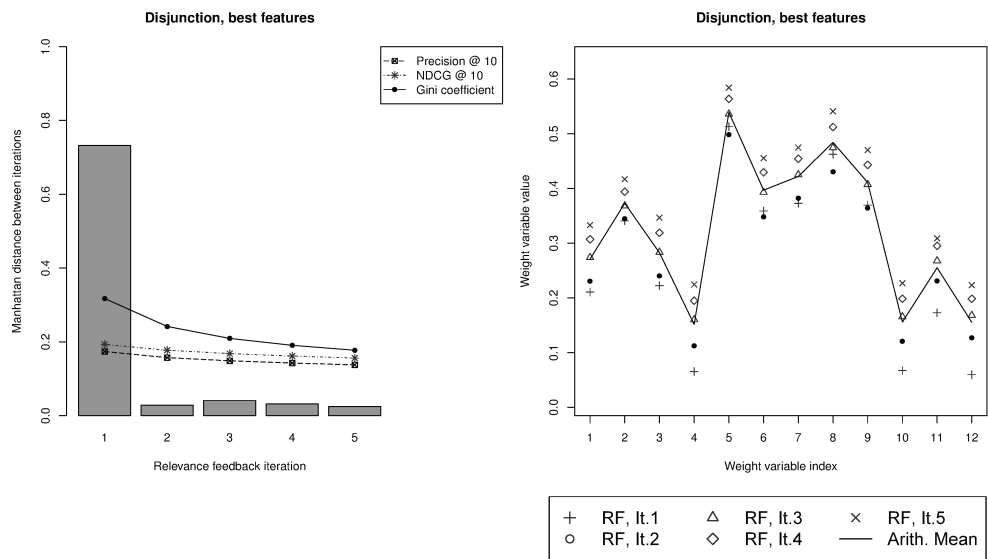


Figure B.36: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 101*

### B.3 Weight Development of Different Matching Functions

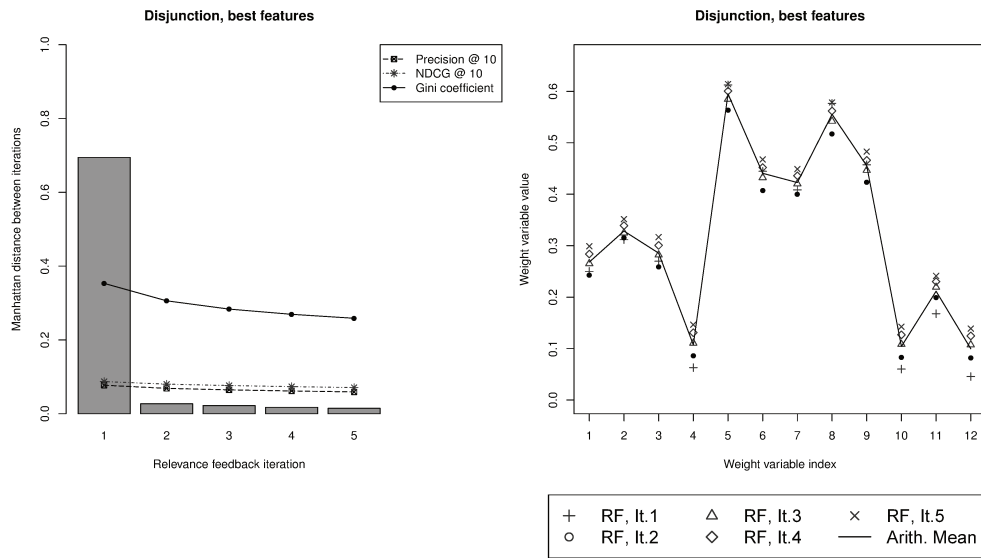


Figure B.37: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 256*

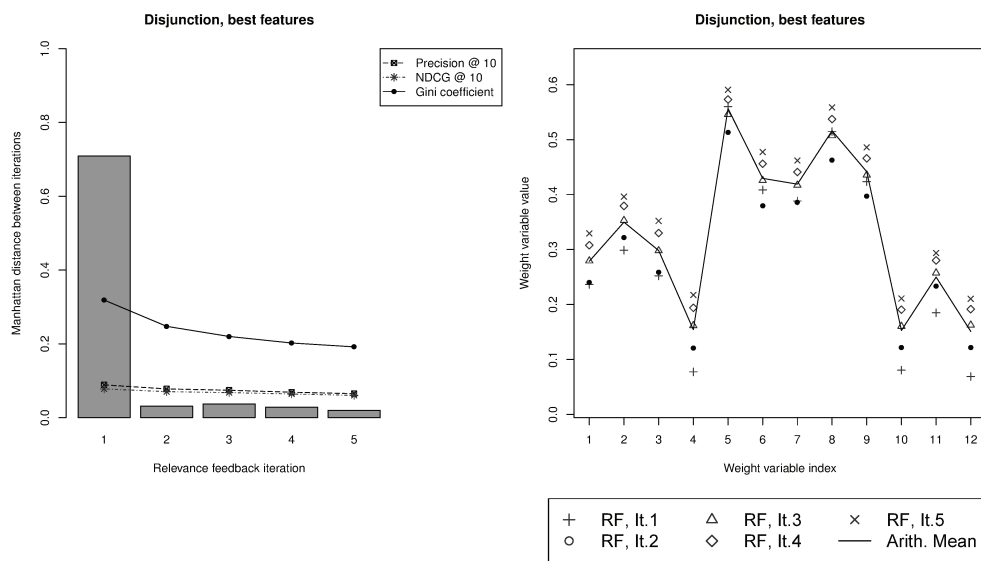


Figure B.38: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *MSRA-MM*

## B Evaluation Appendix

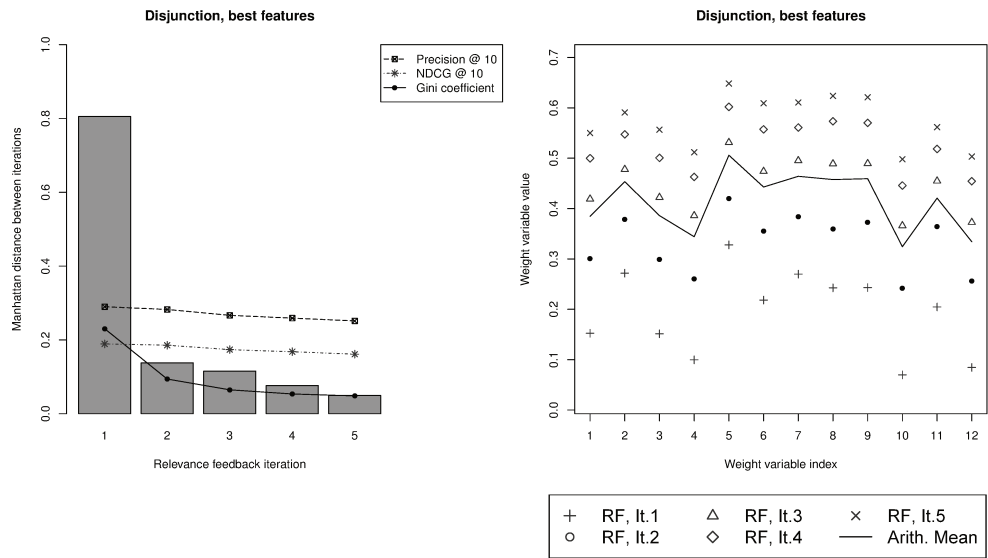


Figure B.39: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Pythia*

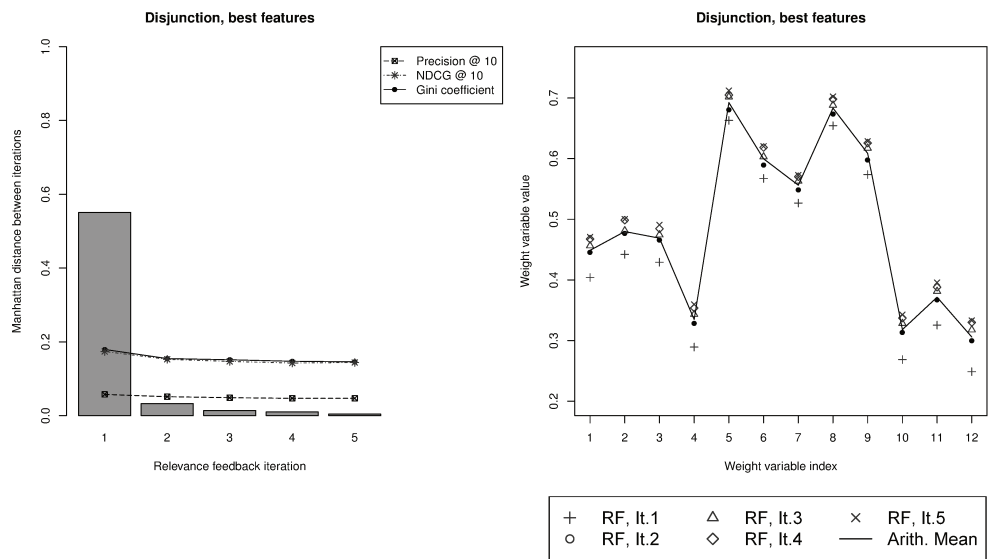


Figure B.40: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *UCID*

### B.3 Weight Development of Different Matching Functions

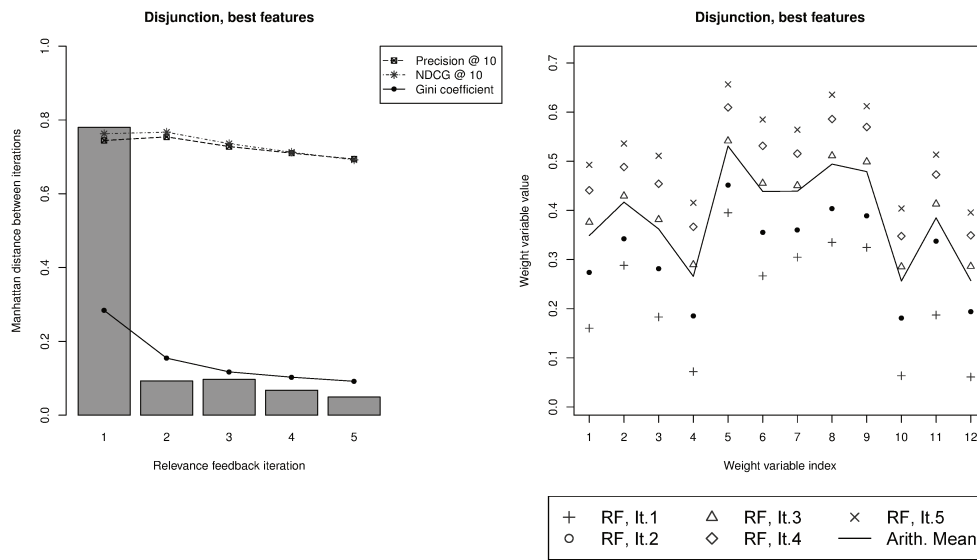


Figure B.41: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Wang*

## B Evaluation Appendix

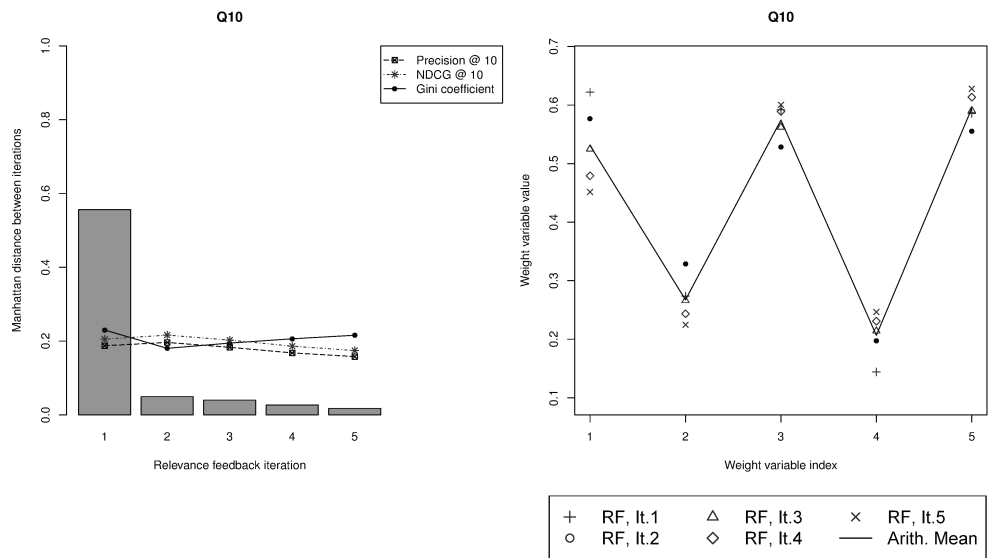


Figure B.42: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 101*

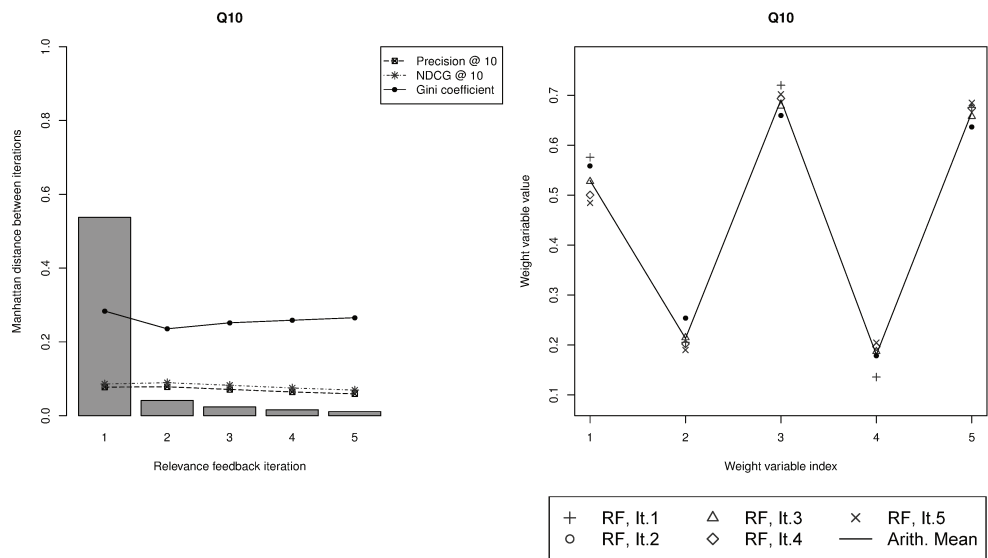


Figure B.43: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 256*

### B.3 Weight Development of Different Matching Functions

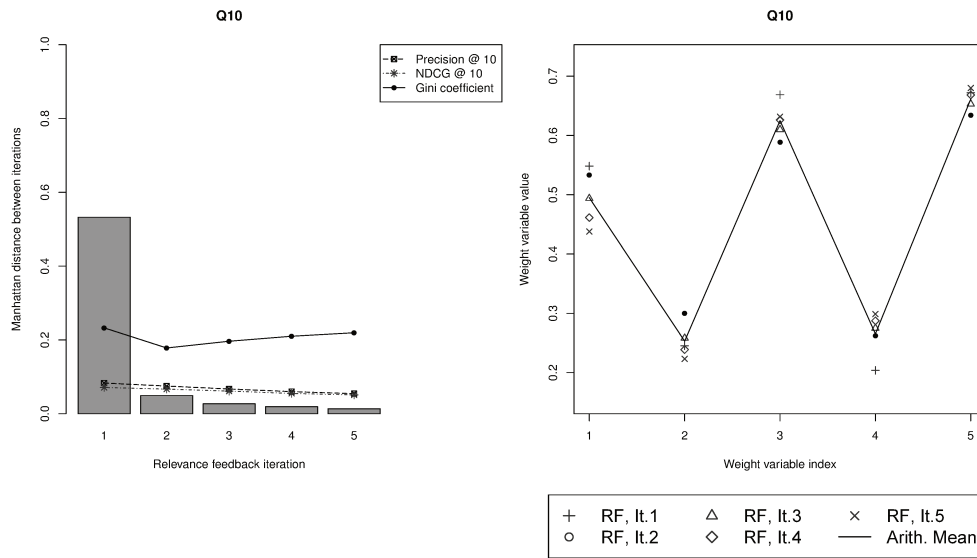


Figure B.44: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *MSRA-MM*

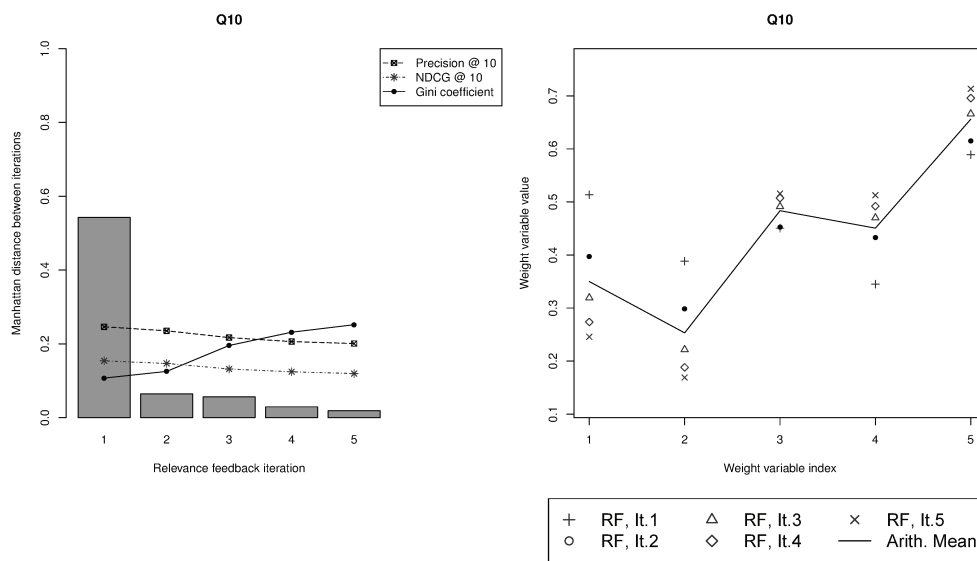


Figure B.45: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Pythia*

## B Evaluation Appendix

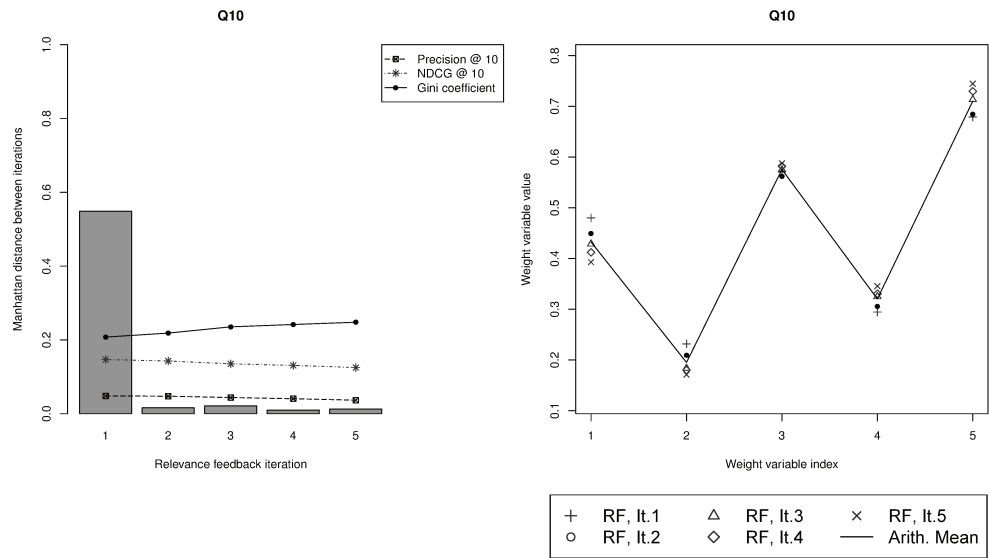


Figure B.46: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *UCID*

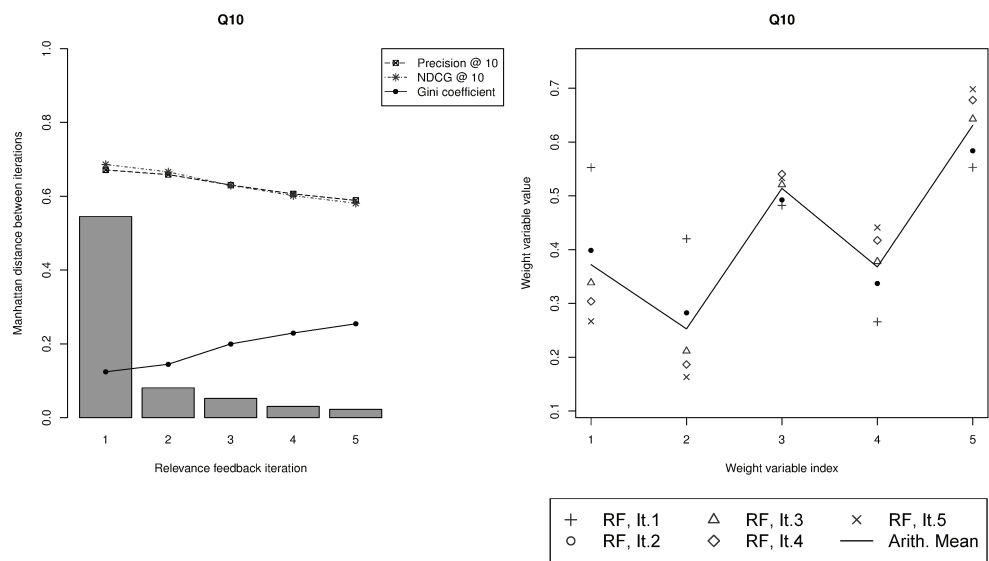


Figure B.47: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Wang*



### B.3 Weight Development of Different Matching Functions

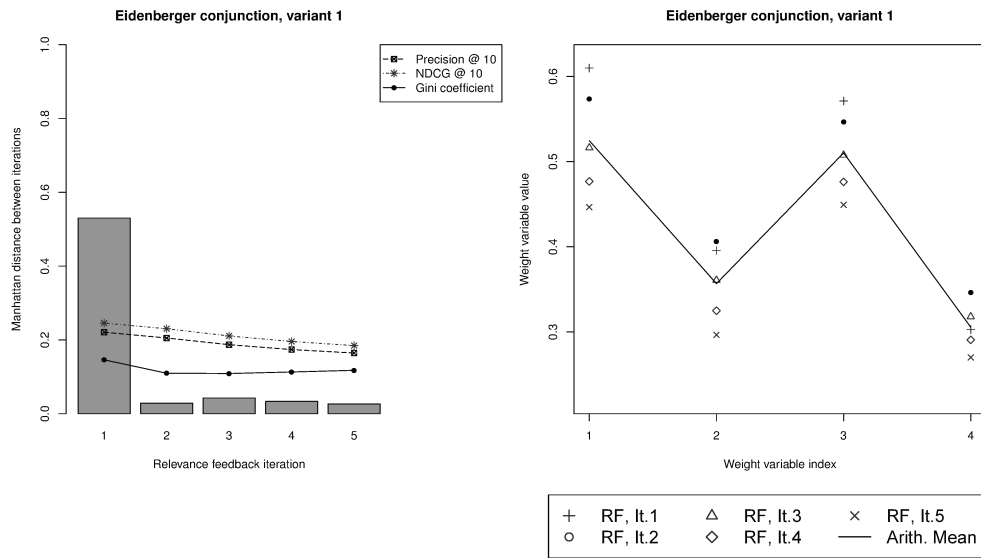


Figure B.48: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 101*

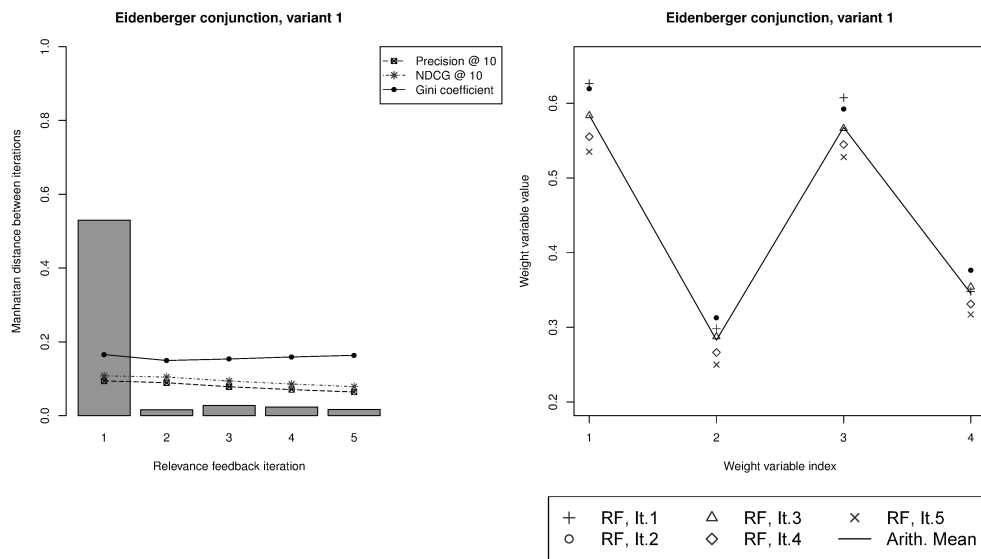


Figure B.49: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 256*

## B Evaluation Appendix

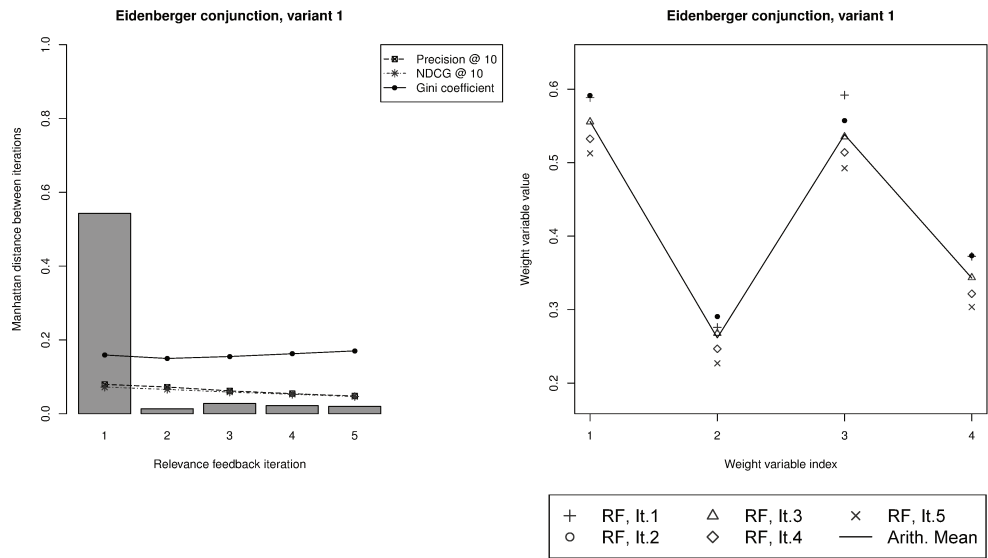


Figure B.50: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *MSRA-MM*

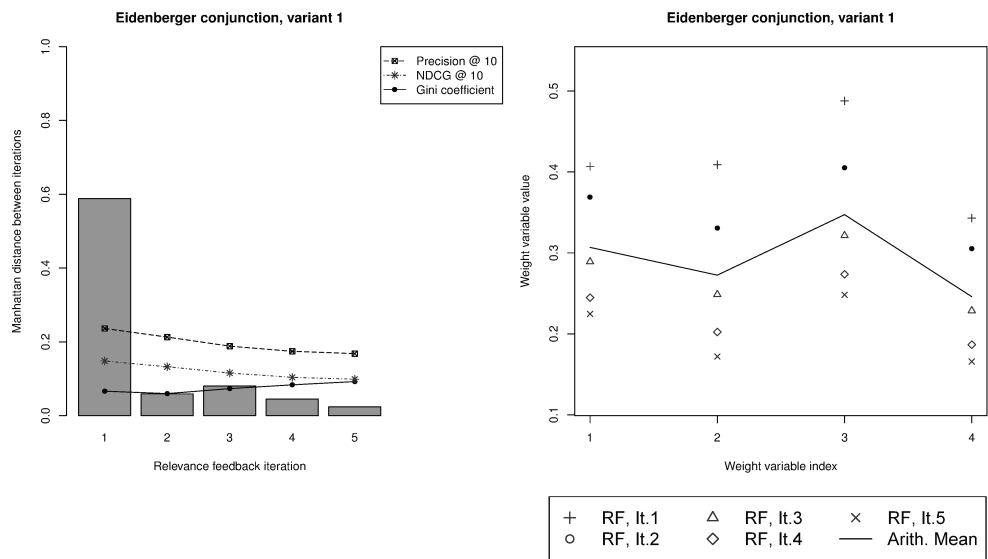


Figure B.51: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Pythia*

### B.3 Weight Development of Different Matching Functions

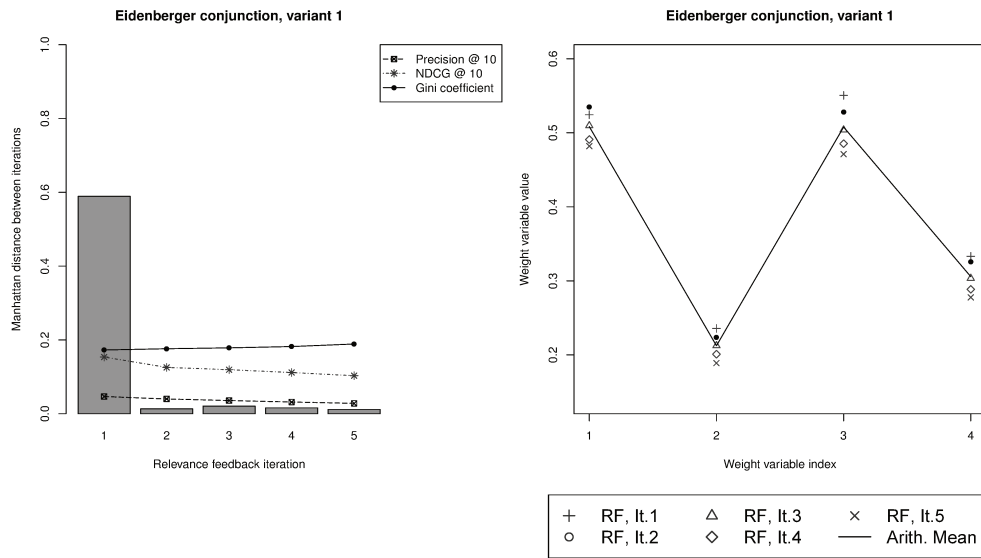


Figure B.52: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *UCID*

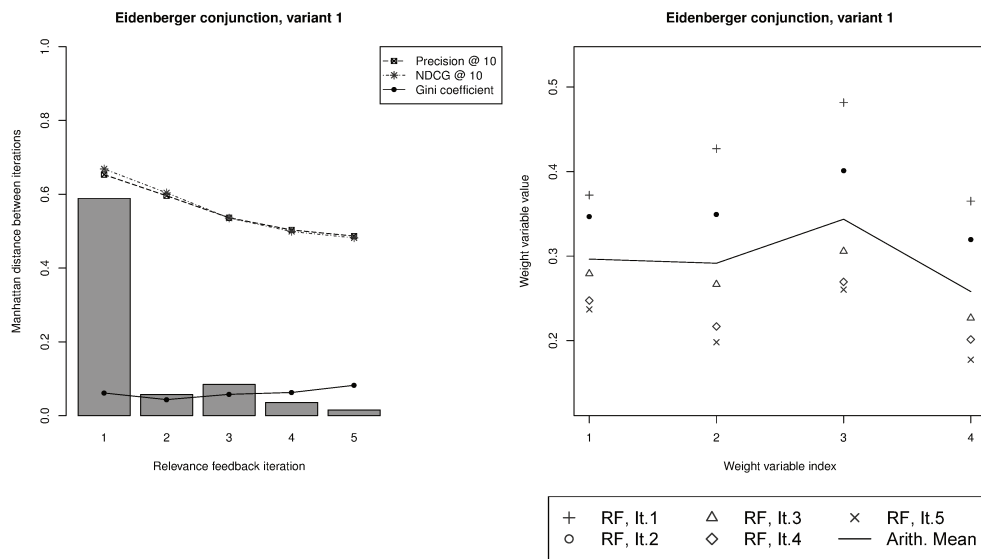


Figure B.53: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Wang*

## B Evaluation Appendix

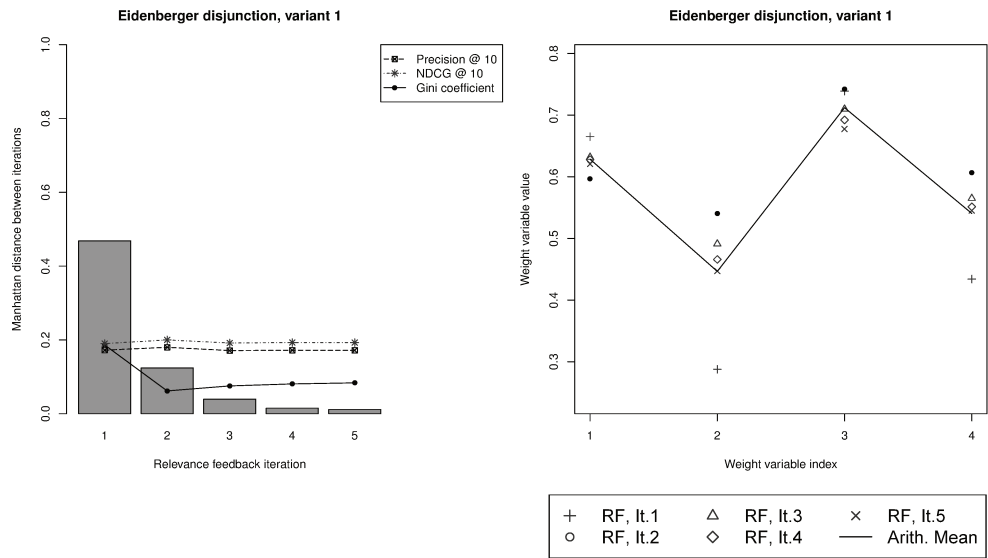


Figure B.54: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 101*

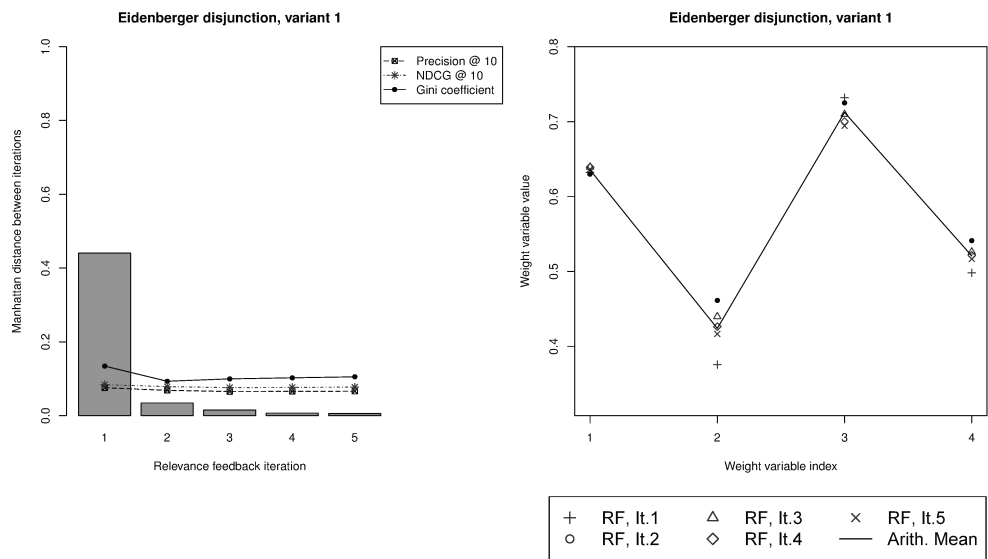


Figure B.55: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 256*

### B.3 Weight Development of Different Matching Functions

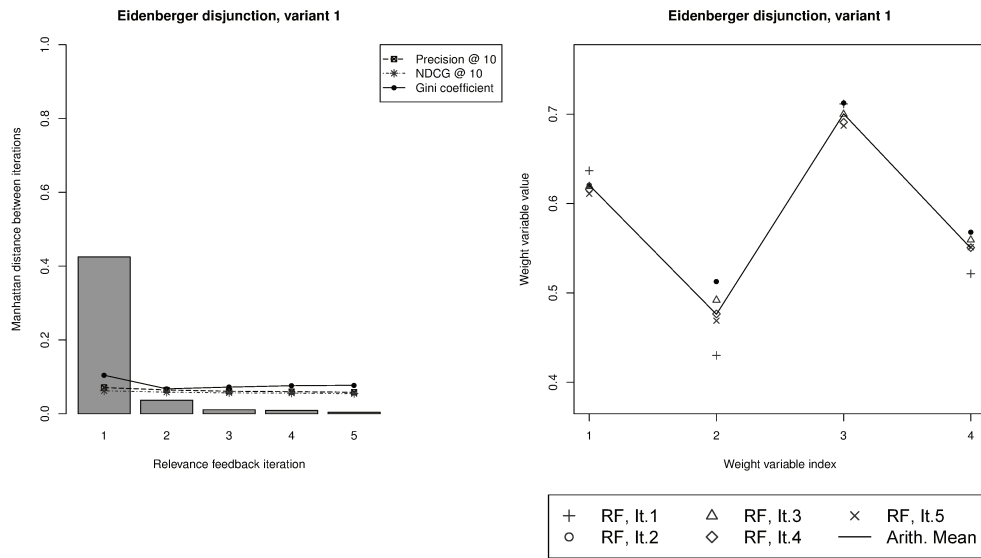


Figure B.56: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *MSRA-MM*

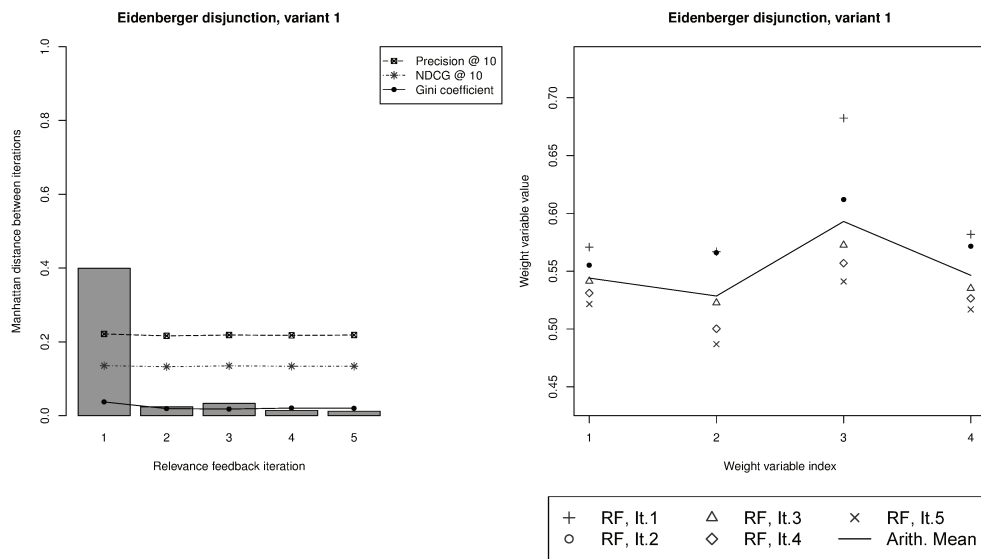


Figure B.57: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Pythia*

## B Evaluation Appendix

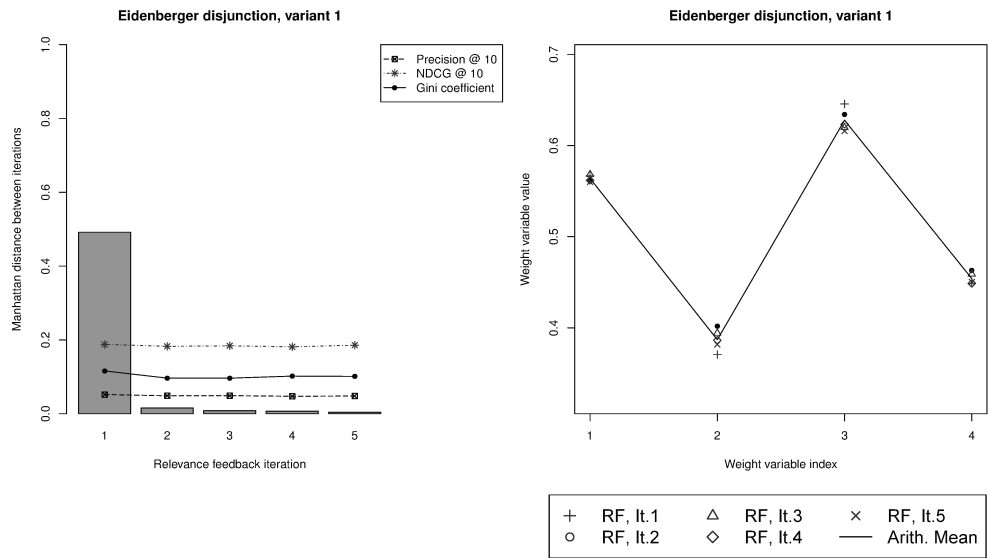


Figure B.58: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *UCID*

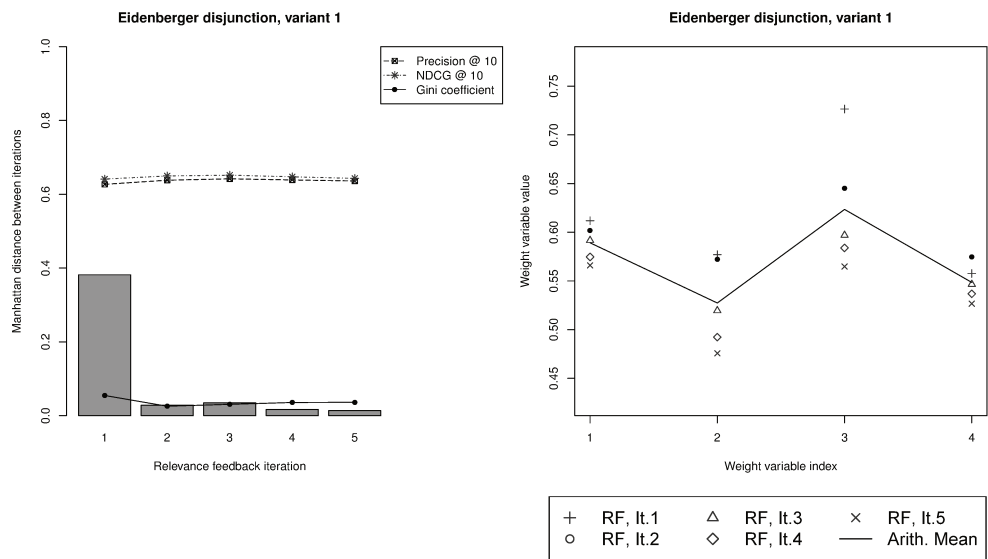


Figure B.59: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Wang*

### B.3 Weight Development of Different Matching Functions

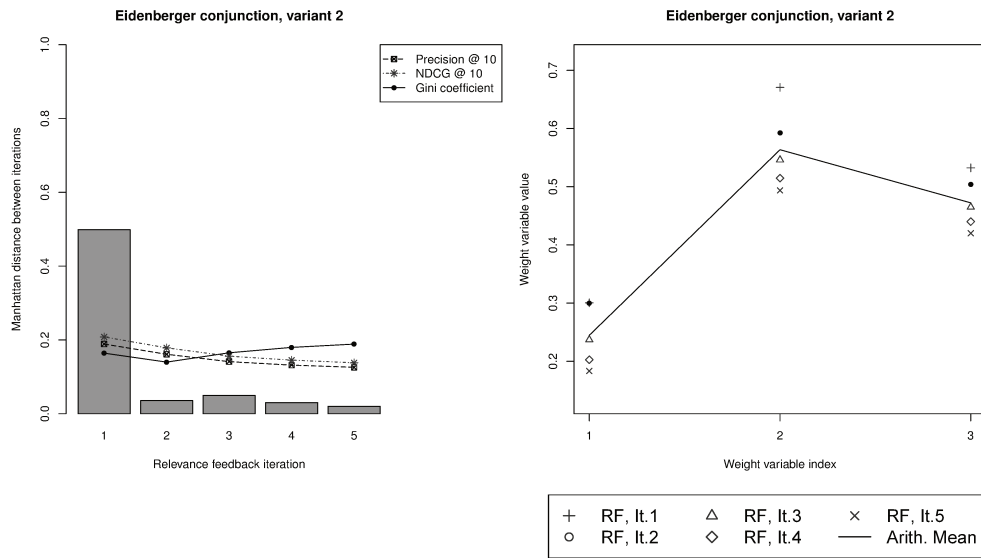


Figure B.60: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 101*

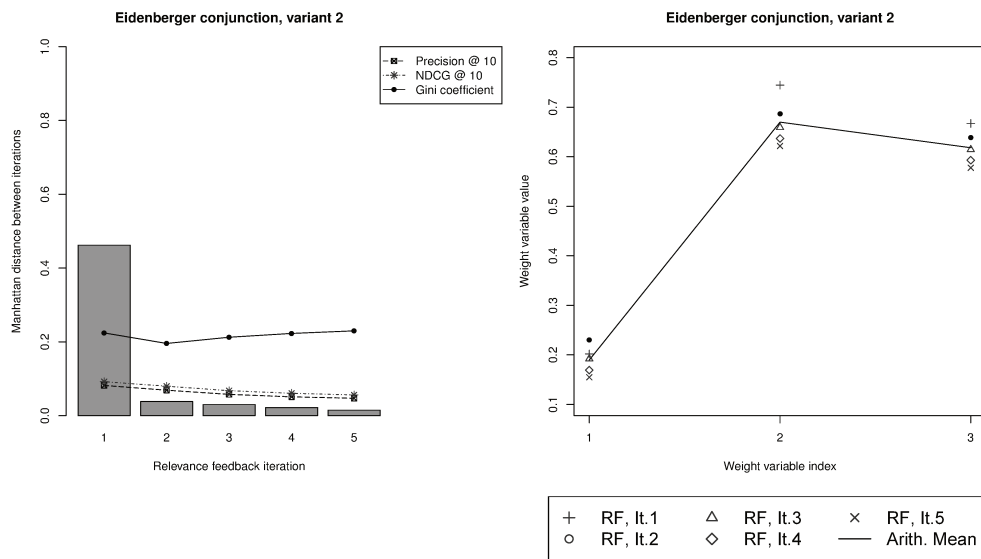


Figure B.61: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 256*

## B Evaluation Appendix

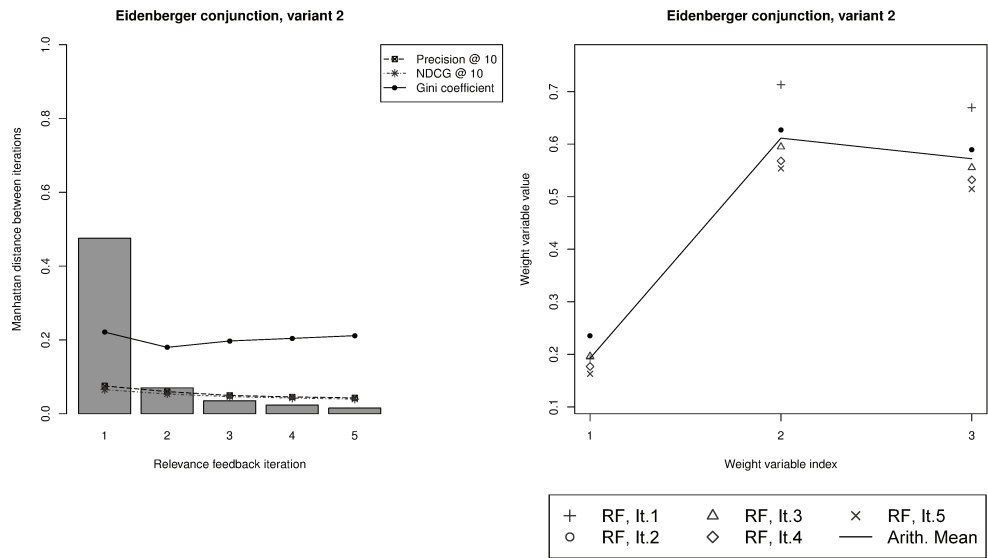


Figure B.62: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *MSRA-MM*

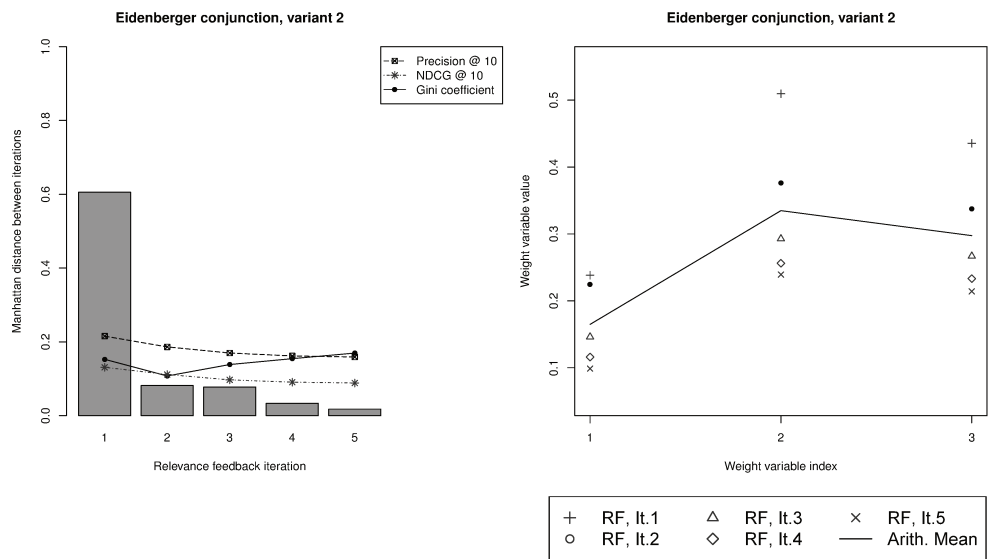


Figure B.63: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Pythia*



### B.3 Weight Development of Different Matching Functions

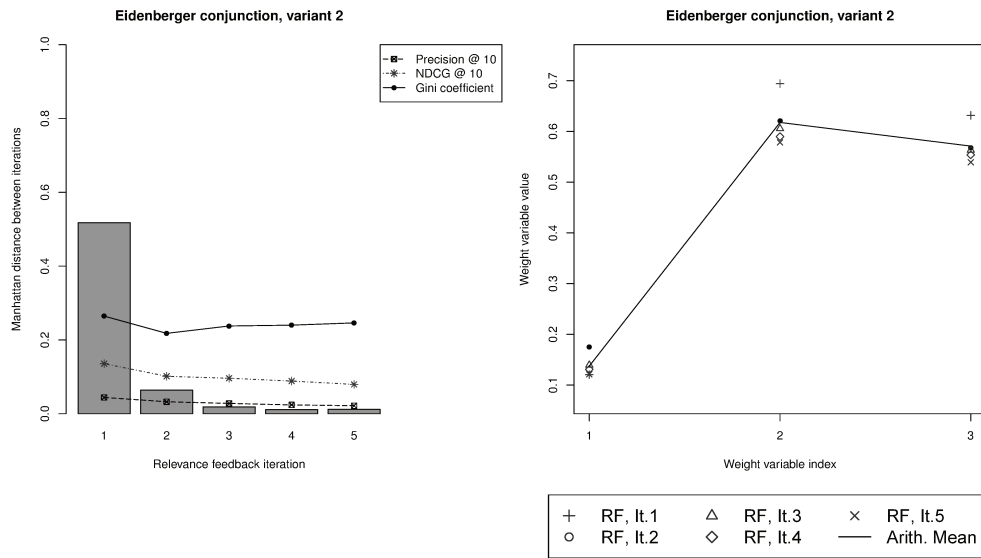


Figure B.64: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *UCID*

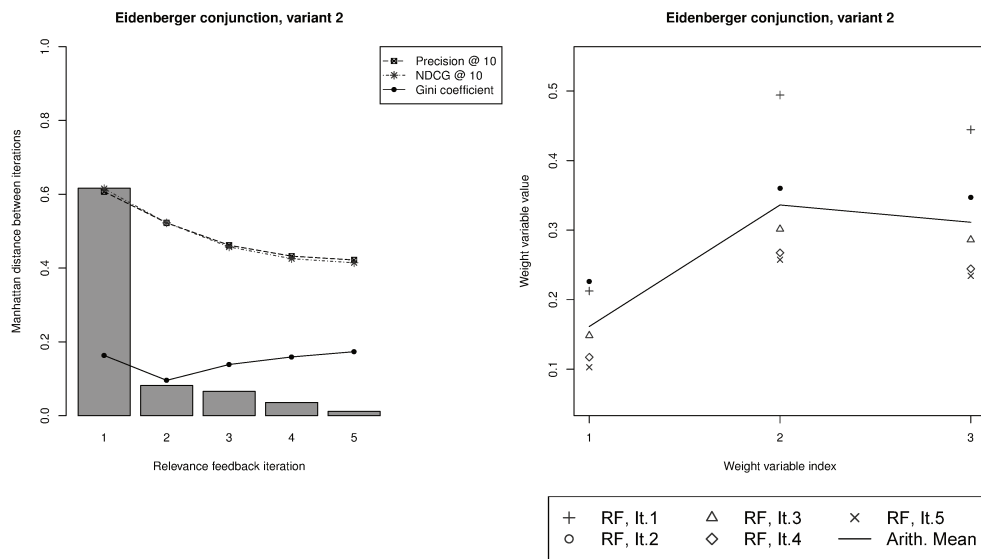


Figure B.65: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Wang*

## B Evaluation Appendix

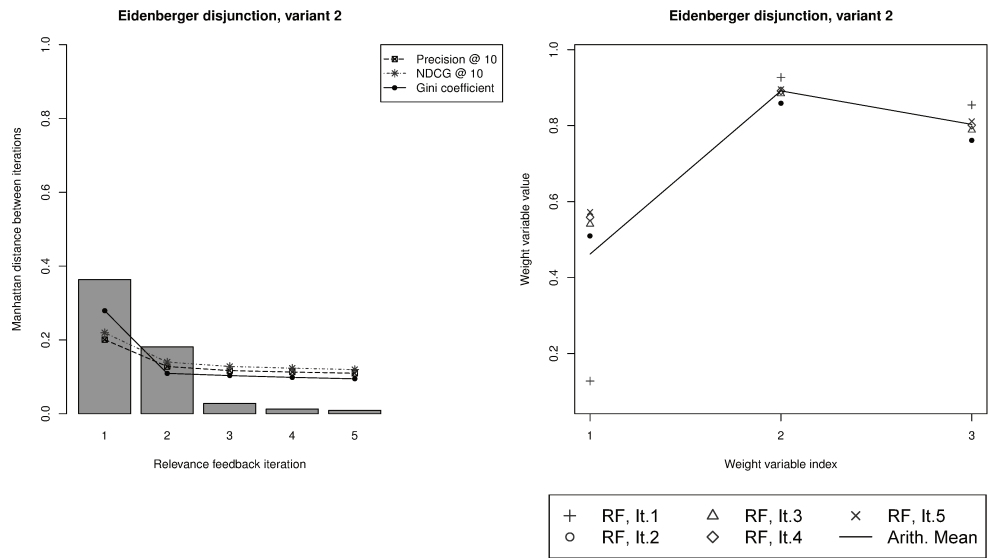


Figure B.66: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 101*

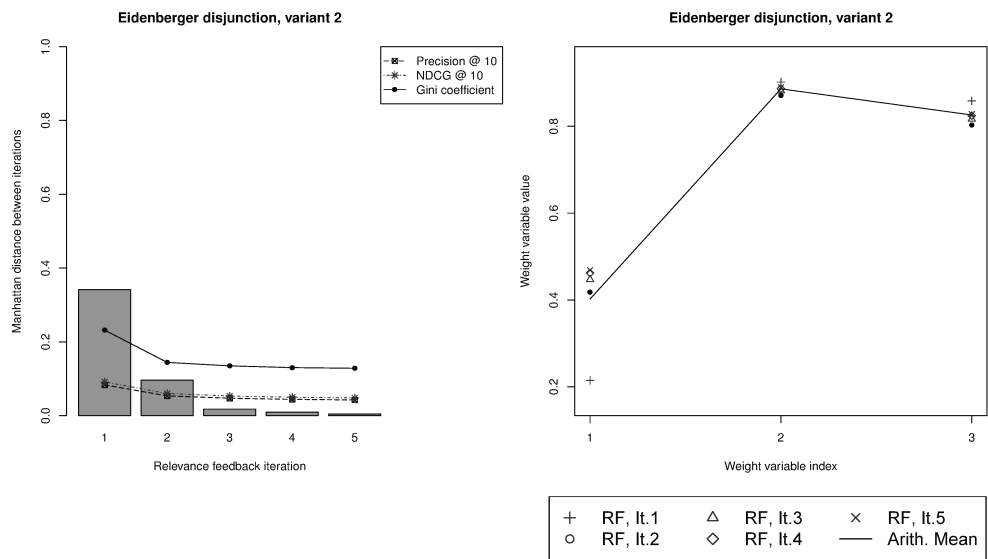


Figure B.67: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 256*

### B.3 Weight Development of Different Matching Functions

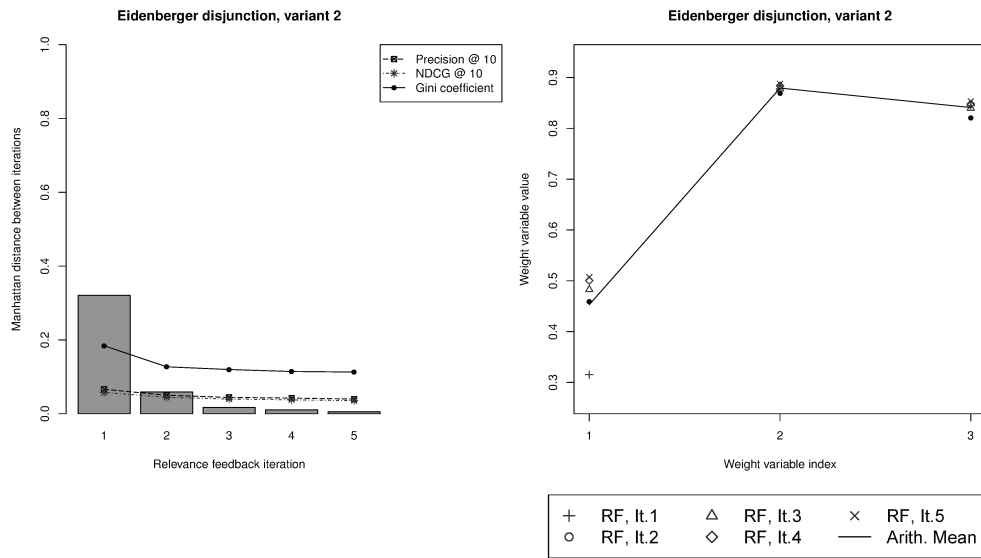


Figure B.68: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *MSRA-MM*

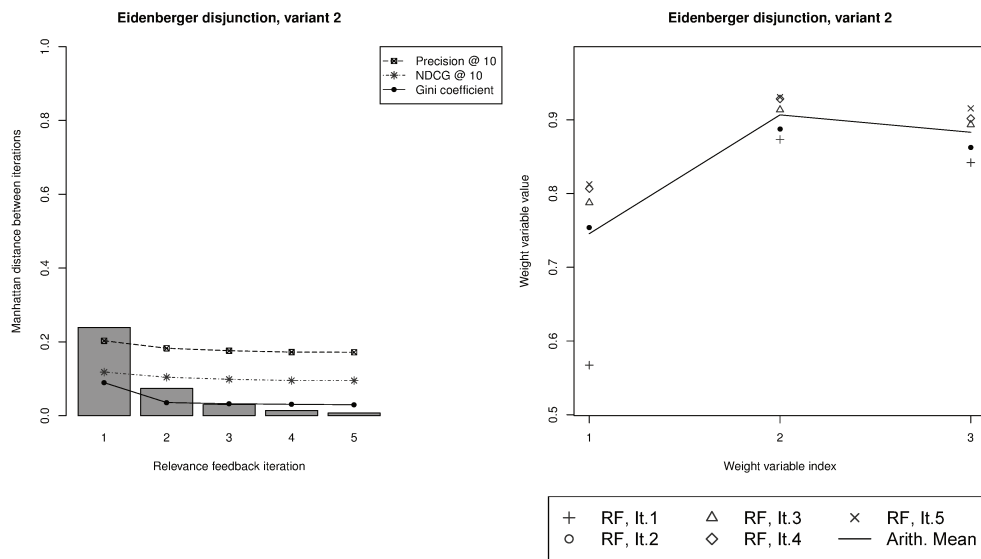


Figure B.69: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Pythia*

## B Evaluation Appendix

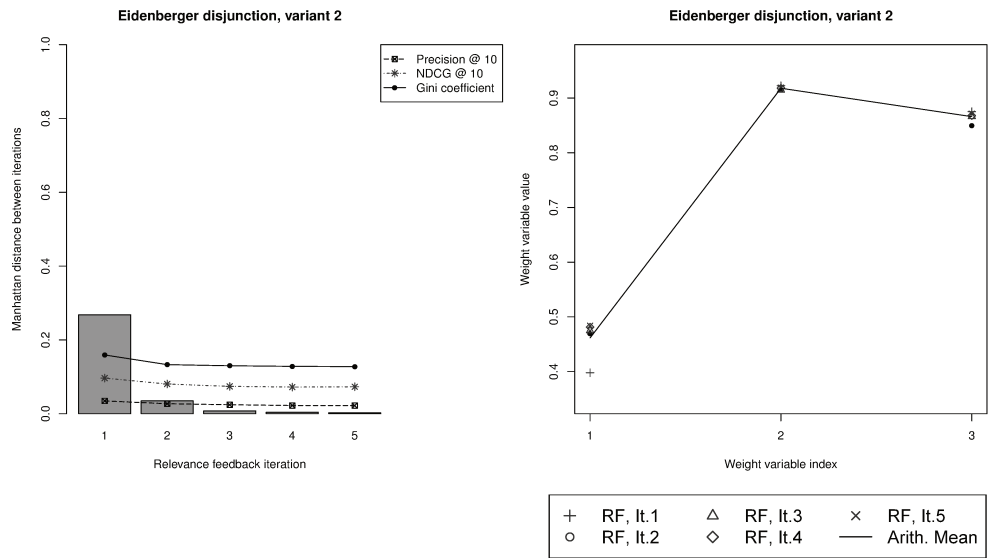


Figure B.70: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *UCID*

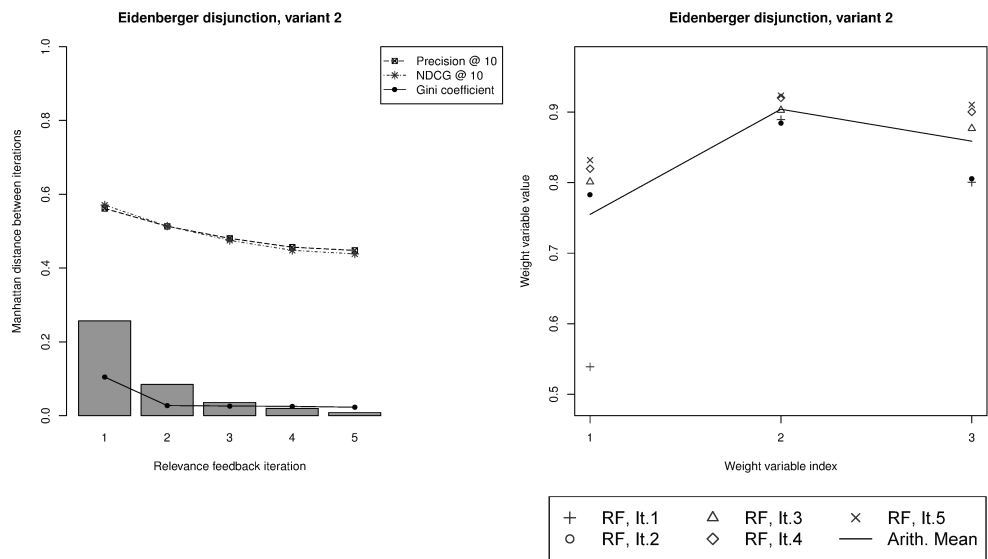


Figure B.71: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Wang*

### B.3 Weight Development of Different Matching Functions

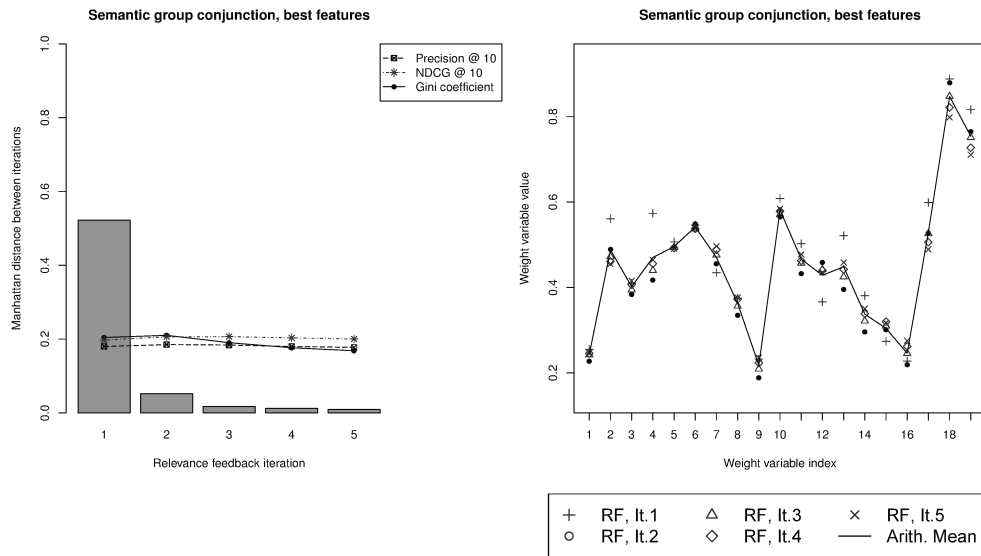


Figure B.72: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 101*

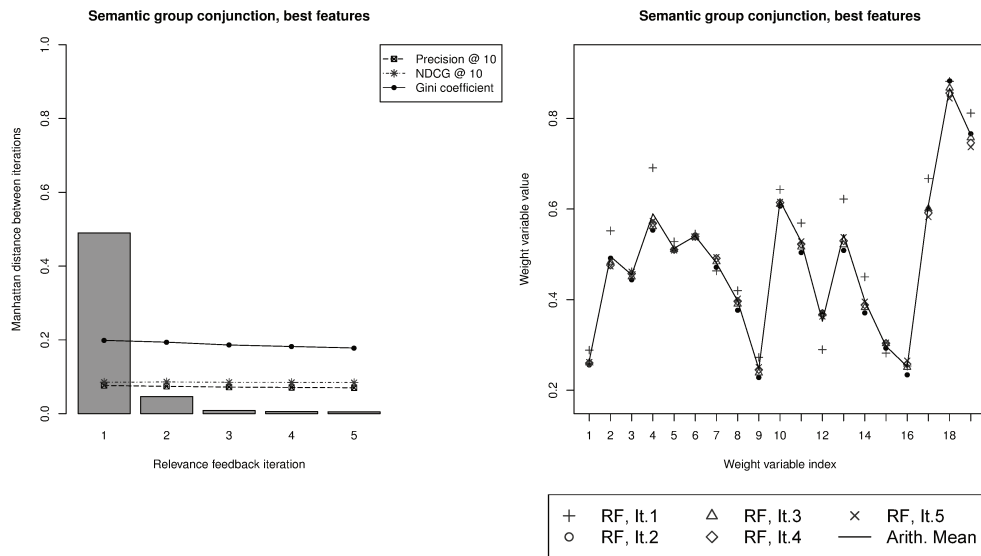


Figure B.73: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 256*

## B Evaluation Appendix

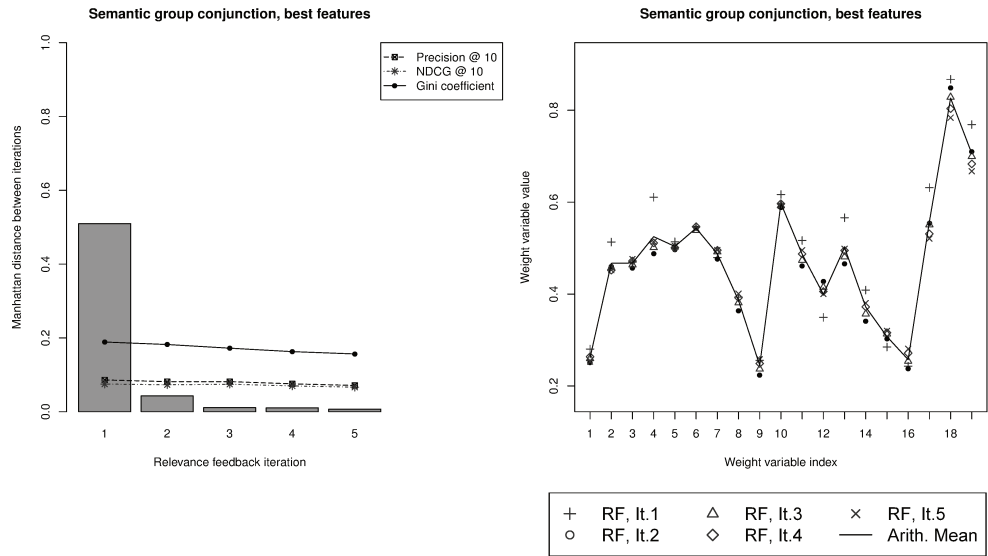


Figure B.74: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *MSRA-MM*

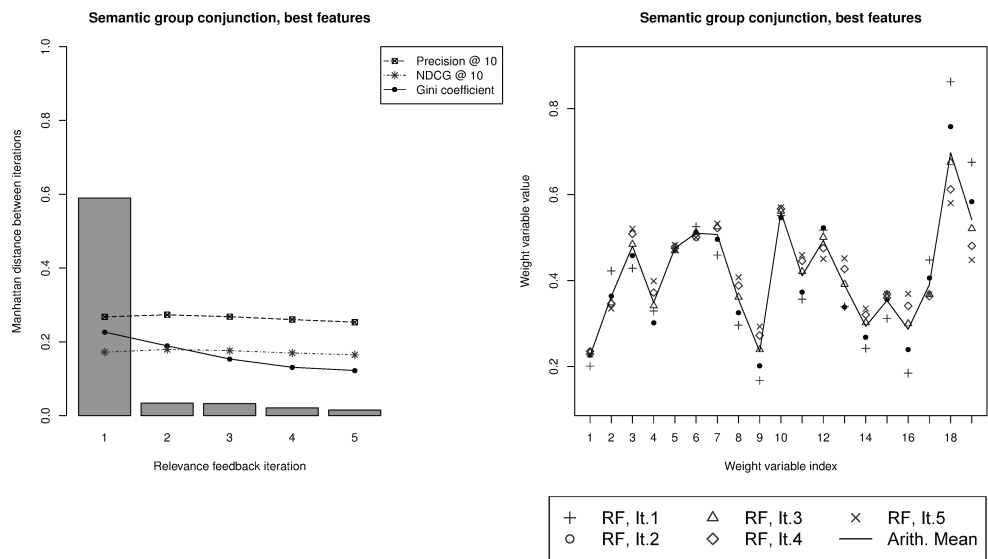


Figure B.75: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Pythia*

### B.3 Weight Development of Different Matching Functions

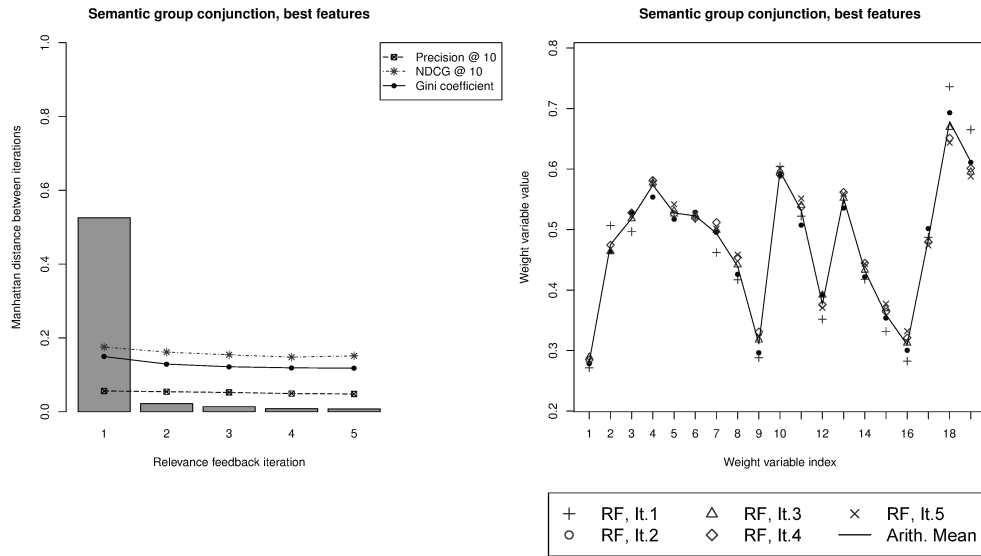


Figure B.76: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *UCID*

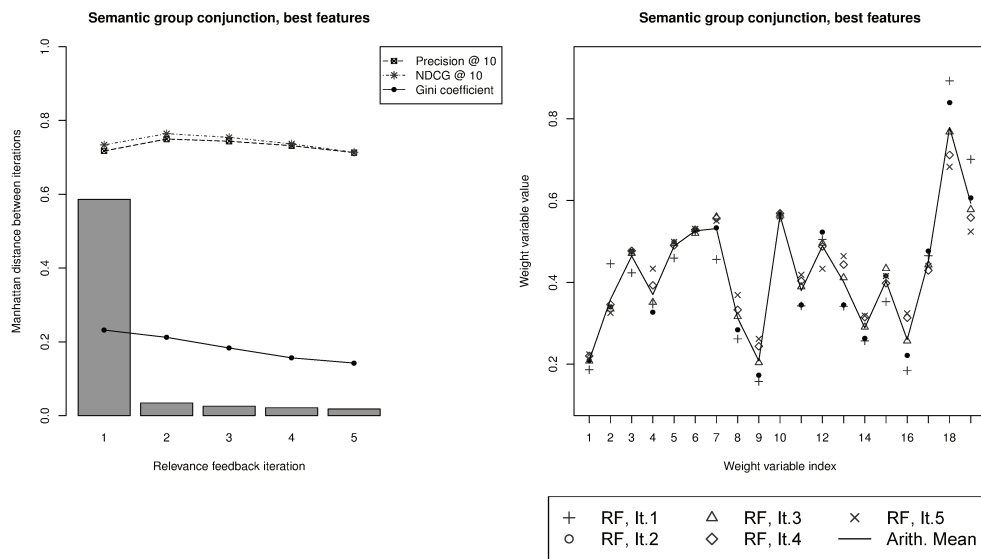


Figure B.77: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Wang*

## B Evaluation Appendix

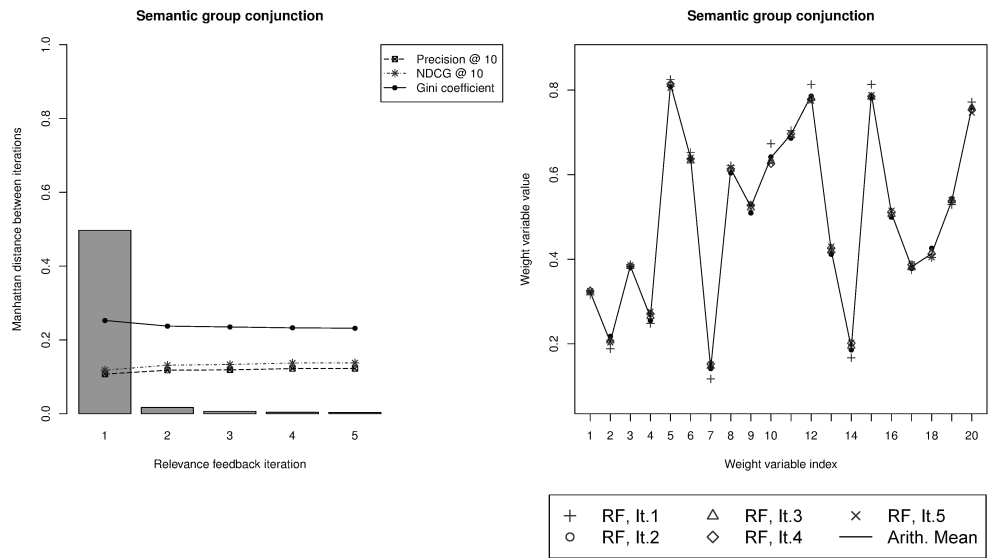


Figure B.78: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 101*

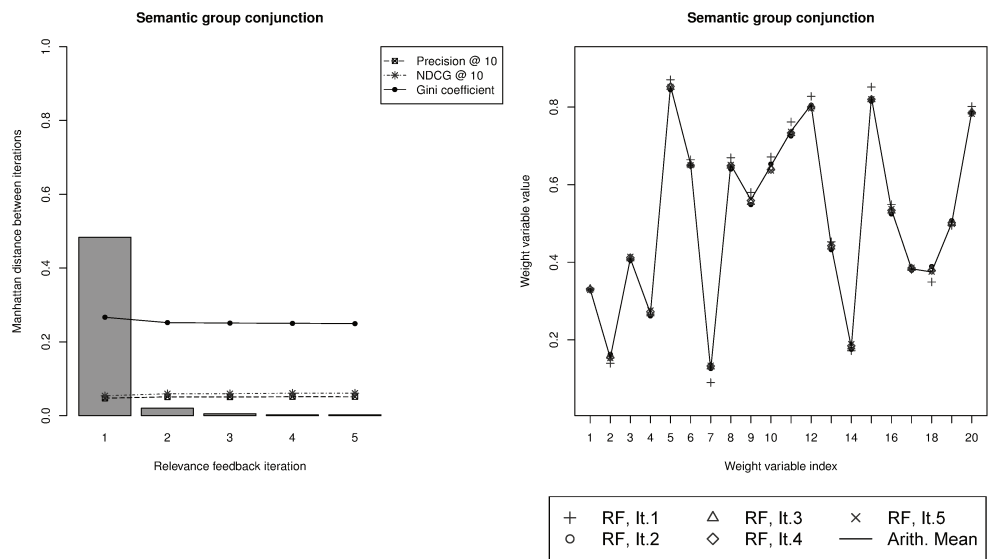


Figure B.79: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 256*



### B.3 Weight Development of Different Matching Functions

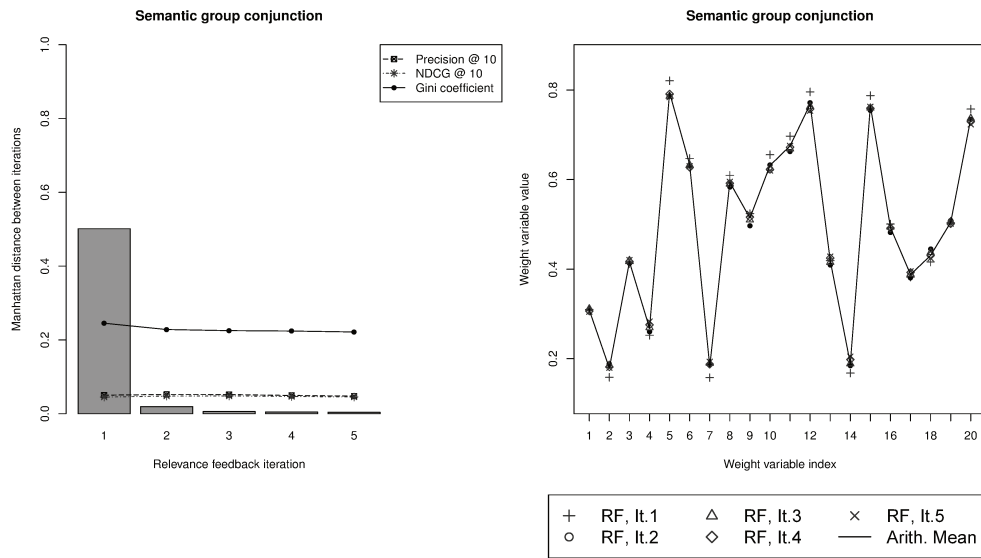


Figure B.80: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *MSRA-MM*

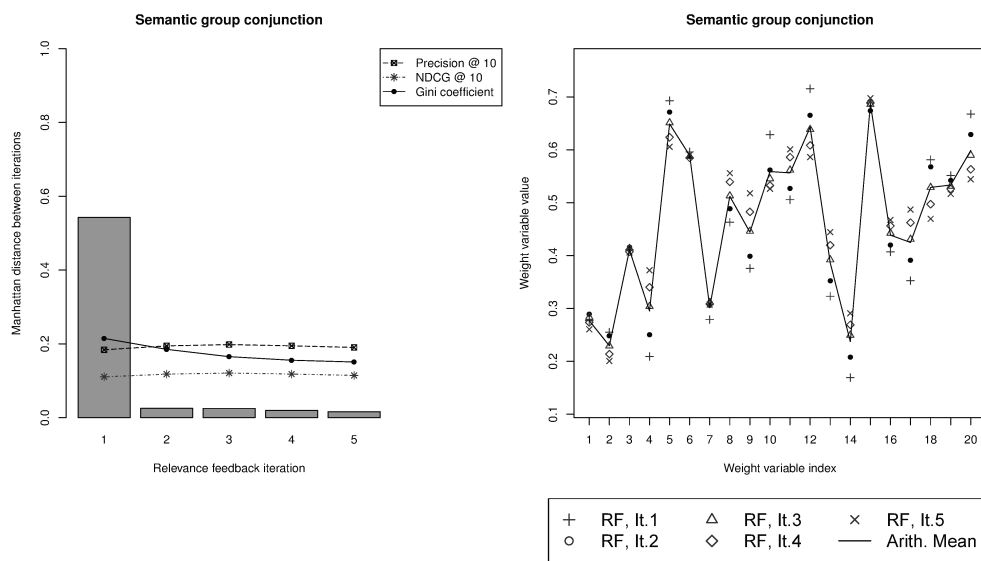


Figure B.81: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Pythia*

## B Evaluation Appendix

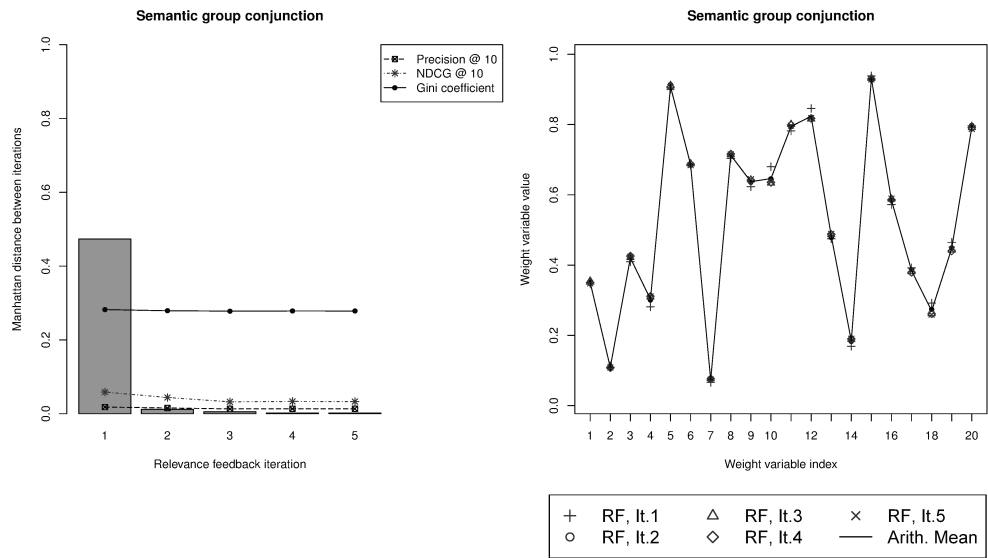


Figure B.82: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *UCID*

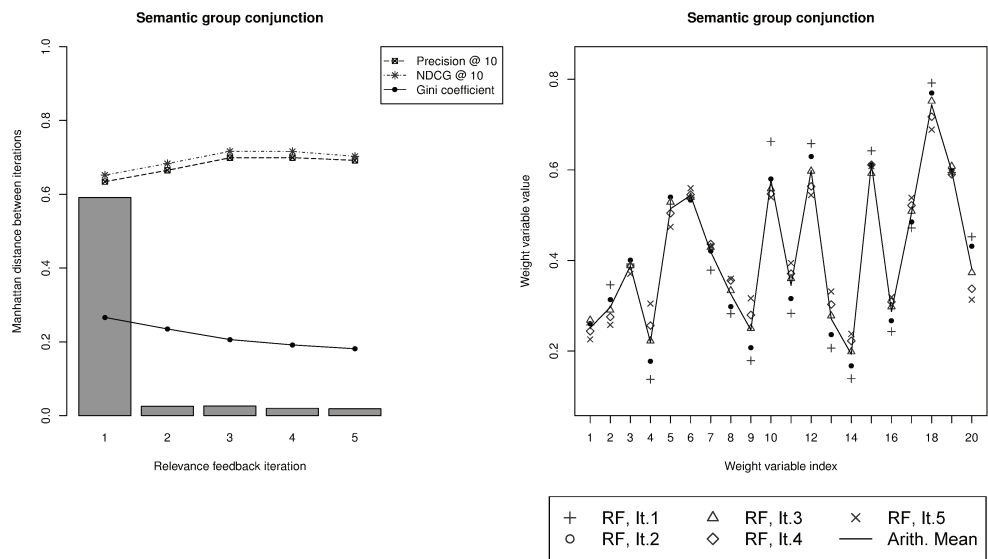


Figure B.83: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Wang*

### B.3 Weight Development of Different Matching Functions

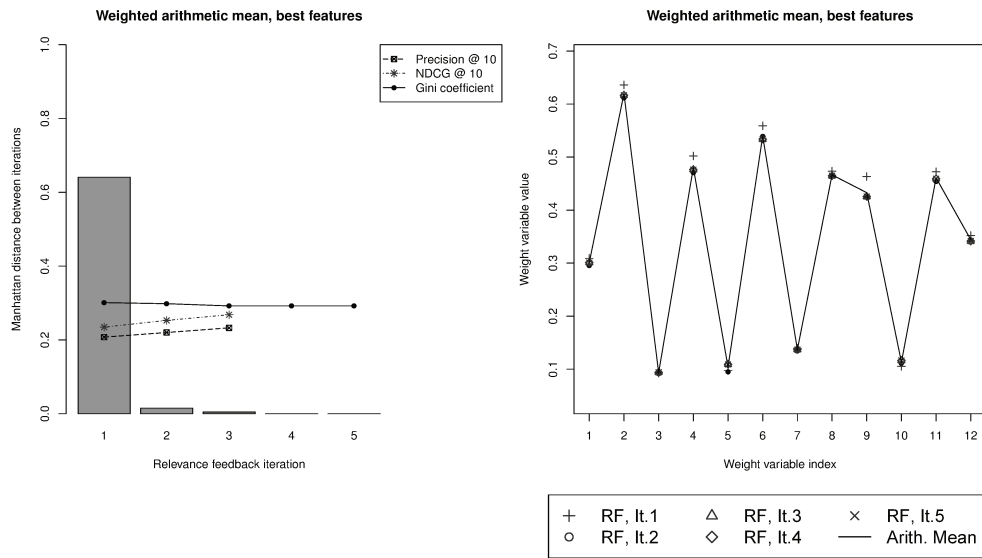


Figure B.84: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 101*

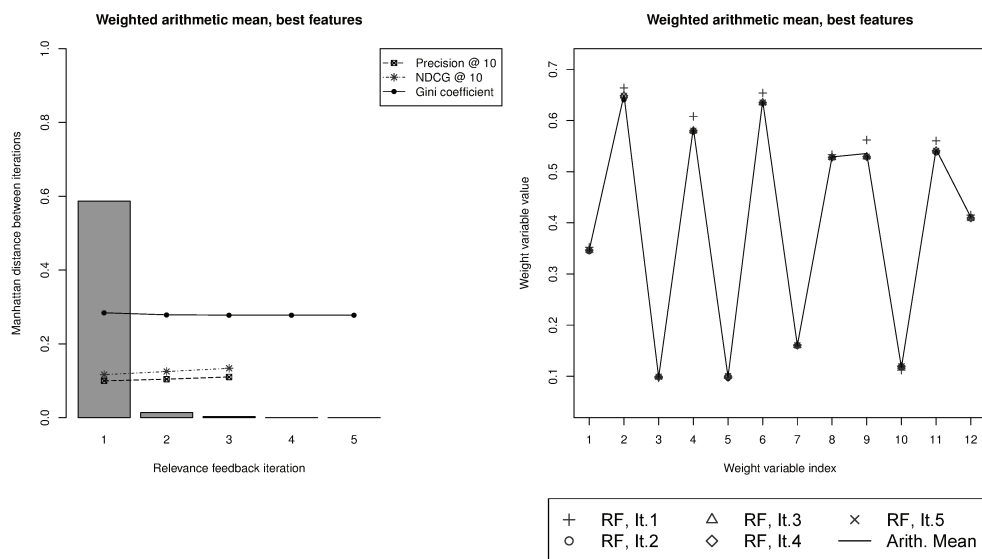


Figure B.85: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Caltech 256*

## B Evaluation Appendix

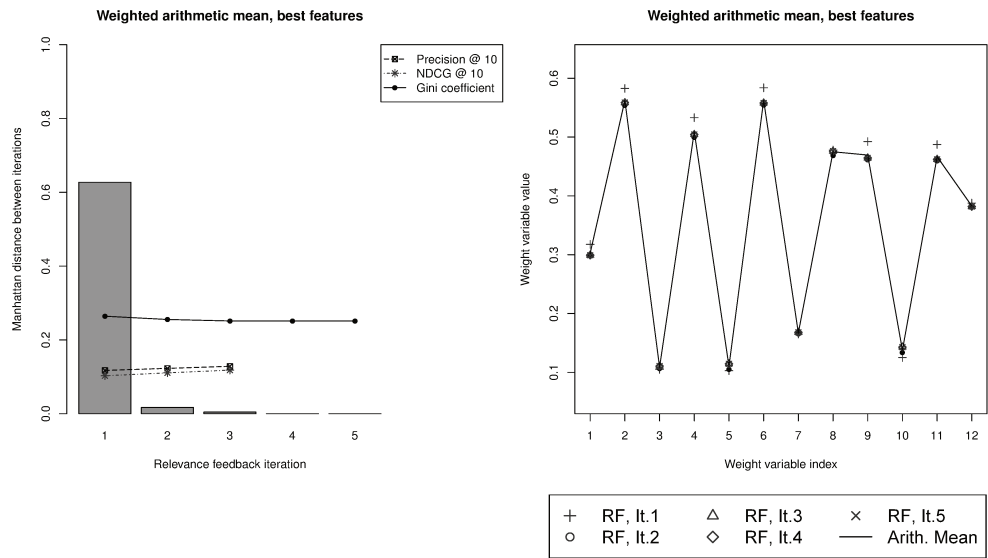


Figure B.86: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *MSRA-MM*

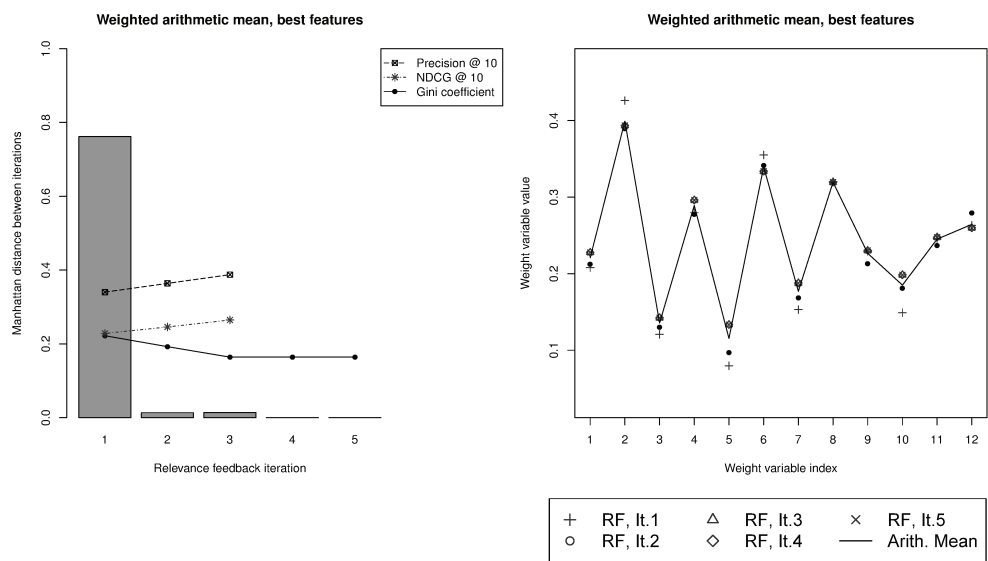


Figure B.87: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Pythia*

### B.3 Weight Development of Different Matching Functions

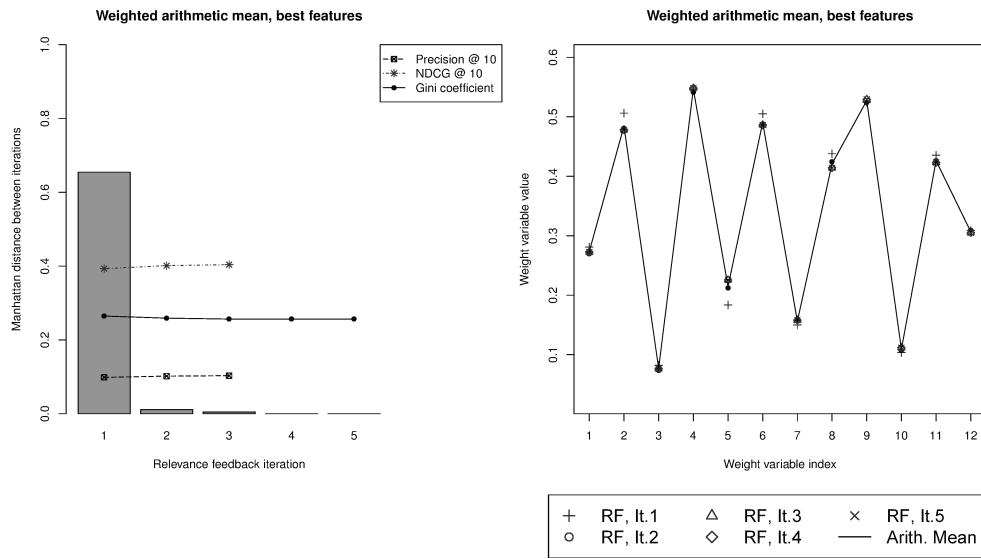


Figure B.88: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *UCID*

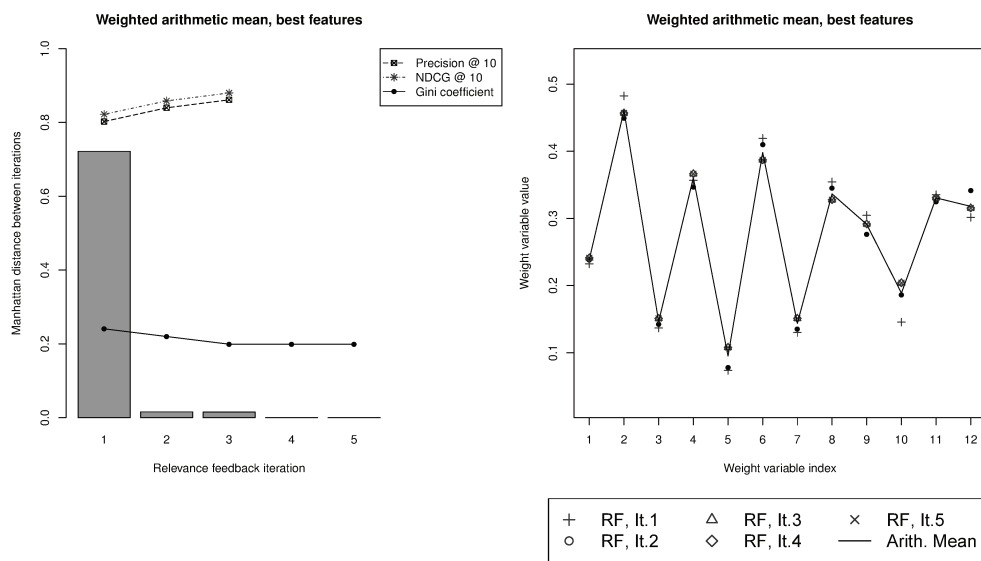
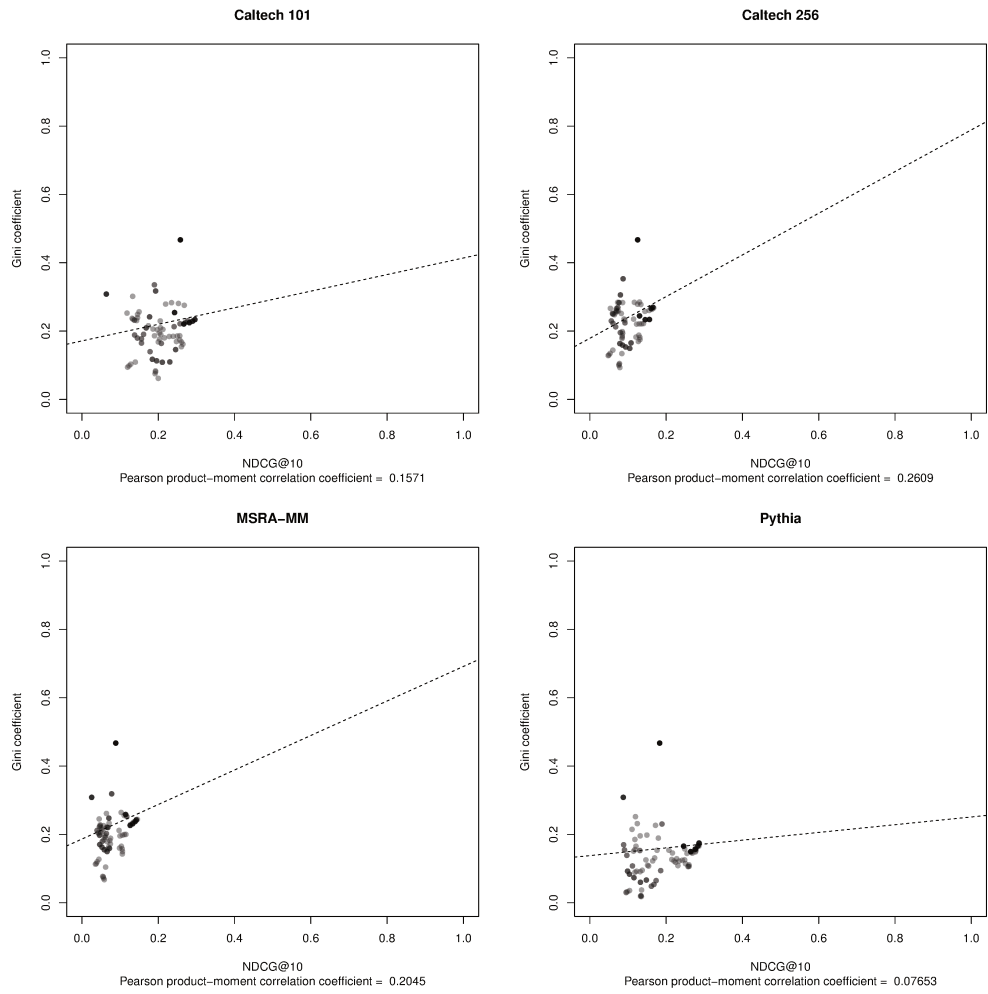


Figure B.89: Weight variable development and distribution over 5 relevance feedback iterations with respect to retrieval metrics (all weights are set to 1 initially) measured with collection *Wang*

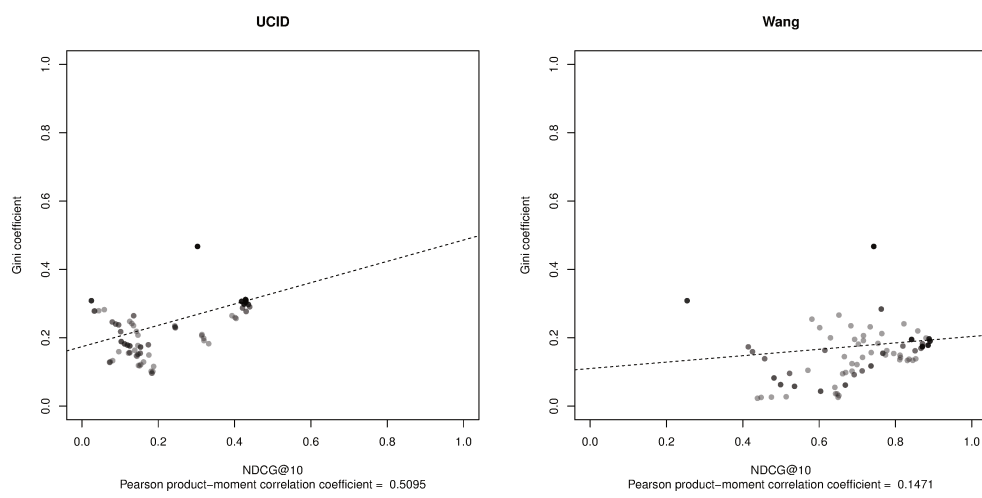
### B.3.3 Correlation Analyses



Dashed lines indicate the fitted linear model of the data. Darker points indicate higher densities.

Figure B.90: Correlation of Gini coefficient and nDCG at 10

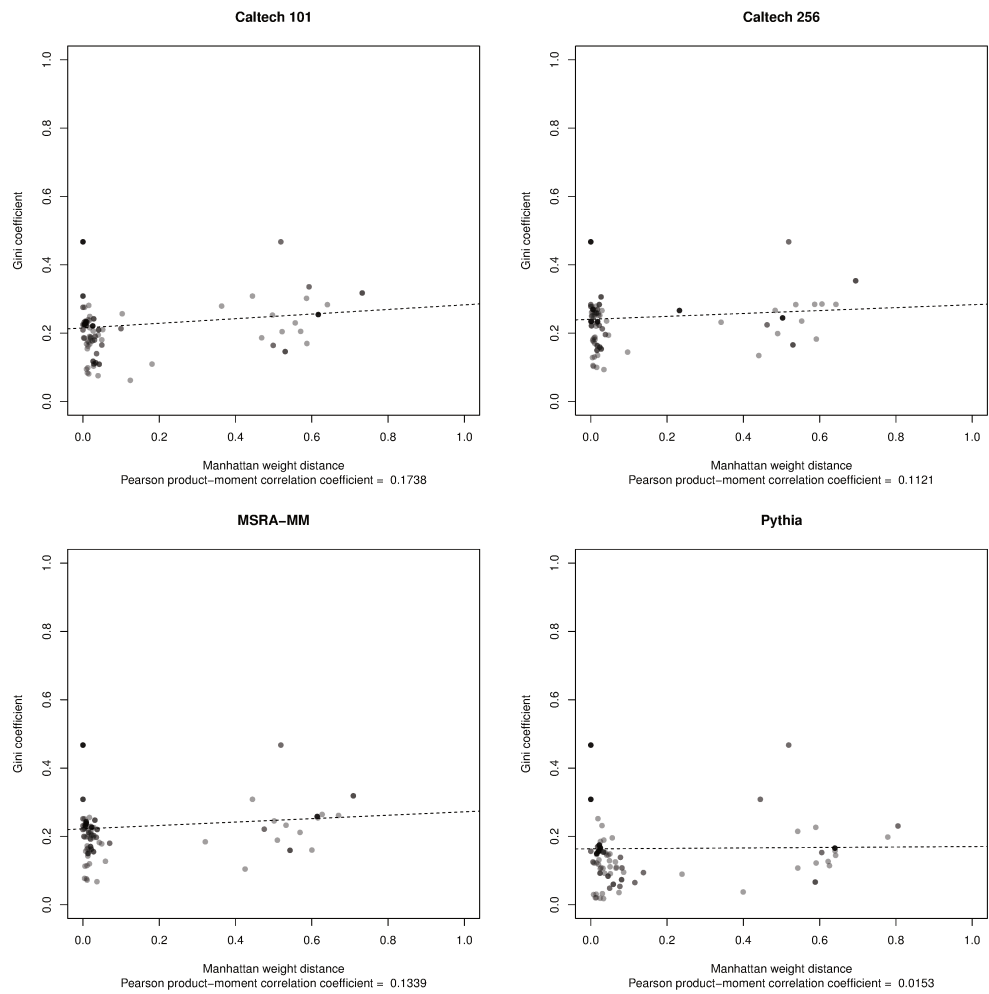
### B.3 Weight Development of Different Matching Functions



Dashed lines indicate the fitted linear model of the data. Darker points indicate higher densities.

Figure B.91: Correlation of Gini coefficient and nDCG at 10

## B Evaluation Appendix

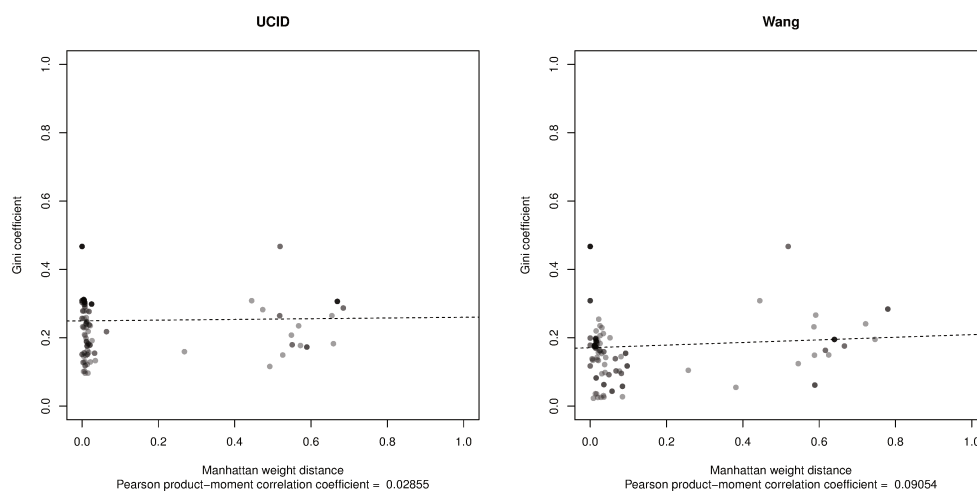


Dashed lines indicate the fitted linear model of the data. Darker points indicate higher densities.

Figure B.92: Correlation of Gini coefficient and weight distance between RF iterations



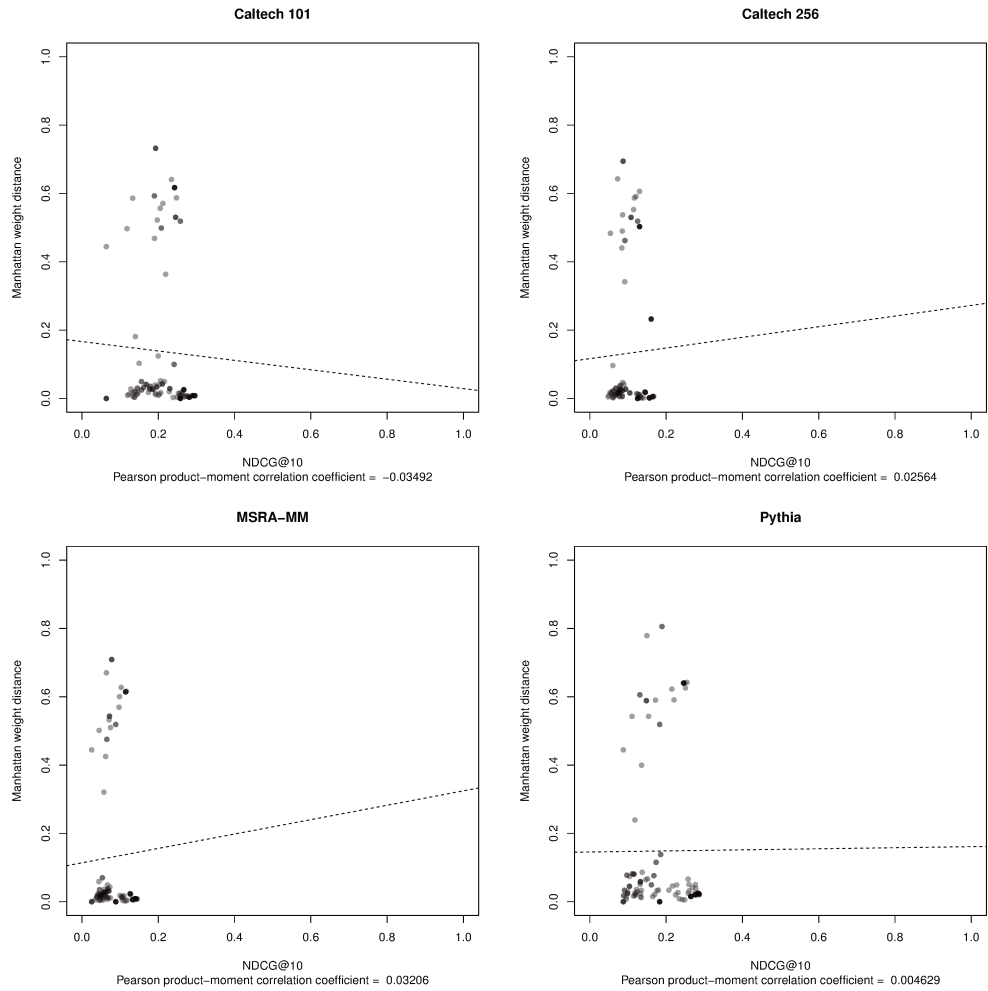
### B.3 Weight Development of Different Matching Functions



Dashed lines indicate the fitted linear model of the data. Darker points indicate higher densities.

Figure B.93: Correlation of Gini coefficient and weight distance between RF iterations

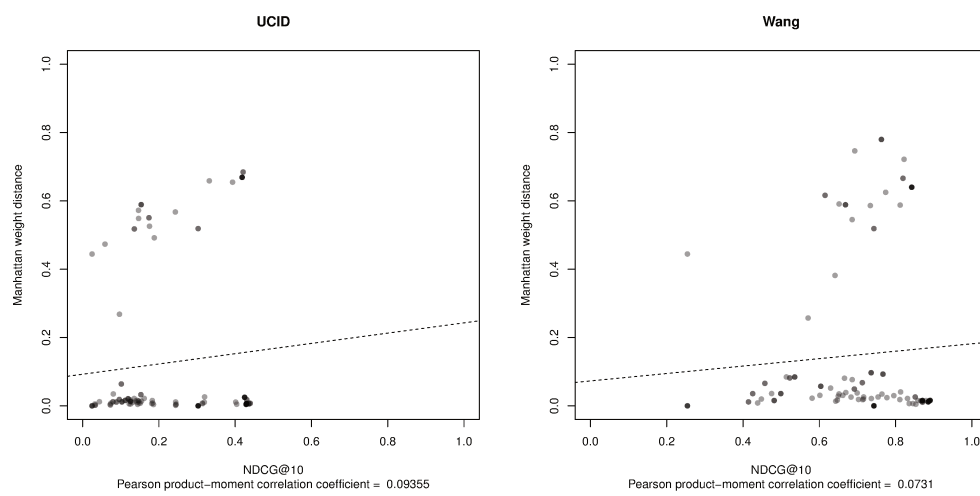
## B Evaluation Appendix



Dashed lines indicate the fitted linear model of the data. Darker points indicate higher densities.

Figure B.94: Correlation of Gini coefficient and nDCG at 10

### B.3 Weight Development of Different Matching Functions



Dashed lines indicate the fitted linear model of the data. Darker points indicate higher densities.

Figure B.95: Correlation of Gini coefficient and nDCG at 10

## B.4 Results of the Usability Study

The following questions were used during the usability study. For each item, the negative and positive statement is given. Each question is encoded as a 7-level Likert item that measures the level of agreement with the negative or positive statement.

1. **Suitability for the Task** In order to assess the suitability of the software for the task, users can express whether the software...
  - a) is complicated to use./is uncomplicated to use.
  - b) does not offer all functions needed to solve the task efficiently./does offer all functions needed to solve the task efficiently.
  - c) offers little support to automate reoccurring tasks./offers good support to automate reoccurring tasks.
  - d) requires superfluous input./requires no superfluous input.
  - e) does not meet the requirements of the task./meets the requirements of the task.
2. **Self-Descriptiveness** In order to assess the self-descriptiveness of the software, users can express whether the software...
  - a) provides a bad overview over its functions./provides a good overview over its functions.
  - b) uses hardly comprehensible terms, abbreviations, or symbols in its interface./uses comprehensible terms, abbreviations, or symbols in its interface.
  - c) offers insufficient information about valid and required input./offers sufficient information about valid and required input.
  - d) offers no context-specific explanations that are helpful on demand./offers context-specific explanations that are helpful on demand.
  - e) offers no automatic context-specific explanations that are helpful./offers no automatic context-specific explanations that are helpful.
3. **Controllability** In order to assess the controllability of the software, users can express whether the software...
  - a) offers no means to pause a task and to continue later on without losing the current progress./offers means to pause a task and to continue later on without losing the current progress.
  - b) forces a needless and fixed interaction process onto the user./does not force a needless and fixed interaction process onto the user.
  - c) allows only complicated switching between interface elements./allows uncomplicated switching between interface elements.

- d) is designed in a way that users cannot control how and what information is displayed on the screen./is designed in a way that users can control how and what information is displayed on the screen.
- e) forces needless interruptions of the work onto the user./does not force needless interruptions of the work onto the user.

4. **Conformity with User Expectations** In order to assess the conformity with the users' expectations of the software, users can express whether the software...

- a) makes orienting difficult due to its inconsistent design./facilitates orienting due to its consistent design.
- b) keeps users in doubt whether an input was successful or not./does not keep users in doubt whether an input was successful or not.
- c) informs the user about its current state insufficiently./informs the user about its current state sufficiently.
- d) reacts with hardly predictable processing times./reacts with predictable processing times.
- e) cannot be operated following consistent principles./can be operated following consistent principles.

5. **Error Tolerance** In order to assess the error tolerance of the software, users can express whether the software...

- a) is designed in a way that minor mistakes will lead to severe consequences./is designed in a way that minor mistakes will not lead to severe consequences.
- b) informs too late about invalid input./informs at once about invalid input.
- c) provides hardly comprehensible error messages./provides comprehensible error messages.
- d) requires a tremendous effort to correct mistakes on the whole./requires little effort to correct mistakes on the whole.
- e) does not provide hints how to solve errors./does provide hints how to solve errors.

6. **Suitability for Individualization** In order to assess the suitability for individualization of the software, users can express whether the software...

- a) is hard to extend by the user when new tasks have to be solved./is easy to extend by the user when new tasks have to be solved.
- b) can hardly be adapted by the user to the individual way of solving a task./can be adapted by the user to the individual way of solving a task.
- c) is not likewise suitable for both beginners and experts because it cannot be adjusted to the user's state of knowledge easily./is likewise suitable for both beginners and experts because it cannot be adjusted to the user's state of knowledge easily.

## B Evaluation Appendix

- d) cannot be adjusted to different work tasks with respect to its core functionality./can be adjusted to different work tasks with respect to its core functionality.
- e) is designed in a way that users cannot adjust the screen layout to their individual needs./is designed in a way that users can adjust the screen layout to their individual needs.

7. **Suitability for Learning** In order to assess the suitability for learning of the software, users can express whether the software...

- a) takes a lot of time to learn./takes little time to learn.
- b) does not encourage to test new functions./encourages to test new functions.
- c) requires users to keep many details in mind./does not require users to keep many details in mind.
- d) is designed in a way that learned actions are not easy to keep in mind./is designed in a way that learned actions are easy to keep in mind.
- e) cannot be learned with help by others or a manual./is easy to learn without help by others or a manual.

The following sections will present the levels of agreement for each question per GUI variant. This data is also available in a technical report [Zellhöfer 2012d].

### B.4.1 Results for the T2 GUI Variant

Table B.2: Suitability for the task (T2)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
1a) is complicated to use.	<table border="1"> <caption>Data for Bar Chart: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Choice (Level)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>12</td></tr> <tr><td>3</td><td>16</td></tr> <tr><td>4</td><td>6</td></tr> <tr><td>5</td><td>11</td></tr> <tr><td>6</td><td>7</td></tr> <tr><td>7</td><td>6</td></tr> </tbody> </table>	Agreement Choice (Level)	Frequency	1	1	2	12	3	16	4	6	5	11	6	7	7	6	is uncomplicated to use.
Agreement Choice (Level)	Frequency																	
1	1																	
2	12																	
3	16																	
4	6																	
5	11																	
6	7																	
7	6																	

## B.4 Results of the Usability Study

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>1b)</b> does not offer all functions needed to solve the task efficiently.</p>	<table border="1"> <caption>Data for Item 1b</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>6</td></tr> <tr><td>2</td><td>14</td></tr> <tr><td>3</td><td>15</td></tr> <tr><td>4</td><td>8</td></tr> <tr><td>5</td><td>7</td></tr> <tr><td>6</td><td>5</td></tr> <tr><td>7</td><td>4</td></tr> </tbody> </table>	Level	Frequency	1	6	2	14	3	15	4	8	5	7	6	5	7	4	<p>does offer all functions needed to solve the task efficiently.</p>
Level	Frequency																	
1	6																	
2	14																	
3	15																	
4	8																	
5	7																	
6	5																	
7	4																	
<p><b>1c)</b> offers little support to automate reoccurring tasks.</p>	<table border="1"> <caption>Data for Item 1c</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>3</td></tr> <tr><td>3</td><td>11</td></tr> <tr><td>4</td><td>27</td></tr> <tr><td>5</td><td>9</td></tr> <tr><td>6</td><td>7</td></tr> <tr><td>7</td><td>1</td></tr> </tbody> </table>	Level	Frequency	1	1	2	3	3	11	4	27	5	9	6	7	7	1	<p>offers good support to automate reoccurring tasks.</p>
Level	Frequency																	
1	1																	
2	3																	
3	11																	
4	27																	
5	9																	
6	7																	
7	1																	
<p><b>1d)</b> requires superfluous input.</p>	<table border="1"> <caption>Data for Item 1d</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>2</td></tr> <tr><td>2</td><td>4</td></tr> <tr><td>3</td><td>8</td></tr> <tr><td>4</td><td>15</td></tr> <tr><td>5</td><td>10</td></tr> <tr><td>6</td><td>15</td></tr> <tr><td>7</td><td>5</td></tr> </tbody> </table>	Level	Frequency	1	2	2	4	3	8	4	15	5	10	6	15	7	5	<p>requires no superfluous input.</p>
Level	Frequency																	
1	2																	
2	4																	
3	8																	
4	15																	
5	10																	
6	15																	
7	5																	
<p><b>1e)</b> does not meet the requirements of the task.</p>	<table border="1"> <caption>Data for Item 1e</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>5</td></tr> <tr><td>2</td><td>12</td></tr> <tr><td>3</td><td>11</td></tr> <tr><td>4</td><td>13</td></tr> <tr><td>5</td><td>9</td></tr> <tr><td>6</td><td>7</td></tr> <tr><td>7</td><td>2</td></tr> </tbody> </table>	Level	Frequency	1	5	2	12	3	11	4	13	5	9	6	7	7	2	<p>meets the requirements of the task.</p>
Level	Frequency																	
1	5																	
2	12																	
3	11																	
4	13																	
5	9																	
6	7																	
7	2																	

Table B.3: Self-descriptiveness (T2)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>2a)</b> provides a bad overview over its functions.</p>	<table border="1"> <caption>Data for Item 2a Bar Chart</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>2</td></tr> <tr><td>2</td><td>12</td></tr> <tr><td>3</td><td>16</td></tr> <tr><td>4</td><td>7</td></tr> <tr><td>5</td><td>9</td></tr> <tr><td>6</td><td>8</td></tr> <tr><td>7</td><td>5</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	2	2	12	3	16	4	7	5	9	6	8	7	5	<p>provides a good overview over its functions.</p>
Agreement Choice	Frequency																	
1	2																	
2	12																	
3	16																	
4	7																	
5	9																	
6	8																	
7	5																	
<p><b>2b)</b> uses hardly comprehensible terms, abbreviations, or symbols in its interface.</p>	<table border="1"> <caption>Data for Item 2b Bar Chart</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>14</td></tr> <tr><td>2</td><td>17</td></tr> <tr><td>3</td><td>12</td></tr> <tr><td>4</td><td>7</td></tr> <tr><td>5</td><td>5</td></tr> <tr><td>6</td><td>3</td></tr> <tr><td>7</td><td>1</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	14	2	17	3	12	4	7	5	5	6	3	7	1	<p>uses comprehensible terms, abbreviations, or symbols in its interface.</p>
Agreement Choice	Frequency																	
1	14																	
2	17																	
3	12																	
4	7																	
5	5																	
6	3																	
7	1																	
<p><b>2c)</b> offers insufficient information about valid and required input.</p>	<table border="1"> <caption>Data for Item 2c Bar Chart</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>4</td></tr> <tr><td>2</td><td>15</td></tr> <tr><td>3</td><td>12</td></tr> <tr><td>4</td><td>13</td></tr> <tr><td>5</td><td>10</td></tr> <tr><td>6</td><td>4</td></tr> <tr><td>7</td><td>1</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	4	2	15	3	12	4	13	5	10	6	4	7	1	<p>offers sufficient information about valid and required input.</p>
Agreement Choice	Frequency																	
1	4																	
2	15																	
3	12																	
4	13																	
5	10																	
6	4																	
7	1																	



## B.4 Results of the Usability Study

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>2d)</b> offers no context-specific explanations that are helpful on demand.</p>	<table border="1"> <caption>Data for Item 2d: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Choice (1-7)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>3</td></tr> <tr><td>2</td><td>16</td></tr> <tr><td>3</td><td>12</td></tr> <tr><td>4</td><td>22</td></tr> <tr><td>5</td><td>5</td></tr> <tr><td>6</td><td>1</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice (1-7)	Frequency	1	3	2	16	3	12	4	22	5	5	6	1	7	0	<p>offers context-specific explanations that are helpful on demand.</p>
Agreement Choice (1-7)	Frequency																	
1	3																	
2	16																	
3	12																	
4	22																	
5	5																	
6	1																	
7	0																	
<p><b>2e)</b> offers no automatic context-specific explanations that are helpful.</p>	<table border="1"> <caption>Data for Item 2e: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Choice (1-7)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>4</td></tr> <tr><td>2</td><td>23</td></tr> <tr><td>3</td><td>11</td></tr> <tr><td>4</td><td>17</td></tr> <tr><td>5</td><td>3</td></tr> <tr><td>6</td><td>0</td></tr> <tr><td>7</td><td>1</td></tr> </tbody> </table>	Agreement Choice (1-7)	Frequency	1	4	2	23	3	11	4	17	5	3	6	0	7	1	<p>offers no automatic context-specific explanations that are helpful.</p>
Agreement Choice (1-7)	Frequency																	
1	4																	
2	23																	
3	11																	
4	17																	
5	3																	
6	0																	
7	1																	

Table B.4: Controllability (T2)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>3a)</b> offers no means to pause a task and to continue later on without losing the current progress.</p>	<table border="1"> <caption>Data for Item 3a: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Choice (1-7)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>4</td></tr> <tr><td>3</td><td>5</td></tr> <tr><td>4</td><td>29</td></tr> <tr><td>5</td><td>7</td></tr> <tr><td>6</td><td>12</td></tr> <tr><td>7</td><td>2</td></tr> </tbody> </table>	Agreement Choice (1-7)	Frequency	1	0	2	4	3	5	4	29	5	7	6	12	7	2	<p>offers means to pause a task and to continue later on without losing the current progress.</p>
Agreement Choice (1-7)	Frequency																	
1	0																	
2	4																	
3	5																	
4	29																	
5	7																	
6	12																	
7	2																	

## B Evaluation Appendix

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>3b)</b> forces a needless and fixed interaction process onto the user.</p>	<table border="1"> <caption>Data for Item 3b</caption> <thead> <tr> <th>Likert Point</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>4</td></tr> <tr><td>3</td><td>7</td></tr> <tr><td>4</td><td>12</td></tr> <tr><td>5</td><td>18</td></tr> <tr><td>6</td><td>15</td></tr> <tr><td>7</td><td>2</td></tr> </tbody> </table>	Likert Point	Frequency	1	0	2	4	3	7	4	12	5	18	6	15	7	2	<p>does not force a needless and fixed interaction process onto the user.</p>
Likert Point	Frequency																	
1	0																	
2	4																	
3	7																	
4	12																	
5	18																	
6	15																	
7	2																	
<p><b>3c)</b> allows only complicated switching between interface elements.</p>	<table border="1"> <caption>Data for Item 3c</caption> <thead> <tr> <th>Likert Point</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>1</td></tr> <tr><td>3</td><td>4</td></tr> <tr><td>4</td><td>13</td></tr> <tr><td>5</td><td>14</td></tr> <tr><td>6</td><td>20</td></tr> <tr><td>7</td><td>7</td></tr> </tbody> </table>	Likert Point	Frequency	1	0	2	1	3	4	4	13	5	14	6	20	7	7	<p>allows uncomplicated switching between interface elements.</p>
Likert Point	Frequency																	
1	0																	
2	1																	
3	4																	
4	13																	
5	14																	
6	20																	
7	7																	
<p><b>3d)</b> is designed in a way that users cannot control how and what information is displayed on the screen.</p>	<table border="1"> <caption>Data for Item 3d</caption> <thead> <tr> <th>Likert Point</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>3</td></tr> <tr><td>2</td><td>1</td></tr> <tr><td>3</td><td>9</td></tr> <tr><td>4</td><td>15</td></tr> <tr><td>5</td><td>14</td></tr> <tr><td>6</td><td>14</td></tr> <tr><td>7</td><td>3</td></tr> </tbody> </table>	Likert Point	Frequency	1	3	2	1	3	9	4	15	5	14	6	14	7	3	<p>is designed in a way that users can control how and what information is displayed on the screen.</p>
Likert Point	Frequency																	
1	3																	
2	1																	
3	9																	
4	15																	
5	14																	
6	14																	
7	3																	
<p><b>3e)</b> forces needless interruptions of the work onto the user.</p>	<table border="1"> <caption>Data for Item 3e</caption> <thead> <tr> <th>Likert Point</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>1</td></tr> <tr><td>3</td><td>2</td></tr> <tr><td>4</td><td>21</td></tr> <tr><td>5</td><td>8</td></tr> <tr><td>6</td><td>19</td></tr> <tr><td>7</td><td>7</td></tr> </tbody> </table>	Likert Point	Frequency	1	0	2	1	3	2	4	21	5	8	6	19	7	7	<p>does not force needless interruptions of the work onto the user.</p>
Likert Point	Frequency																	
1	0																	
2	1																	
3	2																	
4	21																	
5	8																	
6	19																	
7	7																	

Table B.5: Conformity with user expectations (T2)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>4a)</b> makes orienting difficult due to its inconsistent design.</p>	<table border="1"> <caption>Data for Item 4a: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>1</td></tr> <tr><td>3</td><td>5</td></tr> <tr><td>4</td><td>6</td></tr> <tr><td>5</td><td>8</td></tr> <tr><td>6</td><td>31</td></tr> <tr><td>7</td><td>7</td></tr> </tbody> </table>	Agreement Level	Frequency	1	1	2	1	3	5	4	6	5	8	6	31	7	7	<p>facilitates orienting due to its consistent design.</p>
Agreement Level	Frequency																	
1	1																	
2	1																	
3	5																	
4	6																	
5	8																	
6	31																	
7	7																	
<p><b>4b)</b> keeps users in doubt whether an input was successful or not.</p>	<table border="1"> <caption>Data for Item 4b: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>3</td></tr> <tr><td>2</td><td>4</td></tr> <tr><td>3</td><td>7</td></tr> <tr><td>4</td><td>10</td></tr> <tr><td>5</td><td>12</td></tr> <tr><td>6</td><td>17</td></tr> <tr><td>7</td><td>6</td></tr> </tbody> </table>	Agreement Level	Frequency	1	3	2	4	3	7	4	10	5	12	6	17	7	6	<p>does not keep users in doubt whether an input was successful or not.</p>
Agreement Level	Frequency																	
1	3																	
2	4																	
3	7																	
4	10																	
5	12																	
6	17																	
7	6																	
<p><b>4c)</b> informs the user about its current state insufficiently.</p>	<table border="1"> <caption>Data for Item 4c: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>3</td></tr> <tr><td>2</td><td>5</td></tr> <tr><td>3</td><td>13</td></tr> <tr><td>4</td><td>13</td></tr> <tr><td>5</td><td>14</td></tr> <tr><td>6</td><td>9</td></tr> <tr><td>7</td><td>2</td></tr> </tbody> </table>	Agreement Level	Frequency	1	3	2	5	3	13	4	13	5	14	6	9	7	2	<p>informs the user about its current state sufficiently.</p>
Agreement Level	Frequency																	
1	3																	
2	5																	
3	13																	
4	13																	
5	14																	
6	9																	
7	2																	

## B Evaluation Appendix

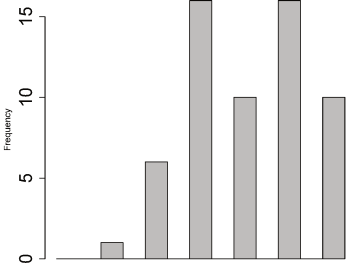
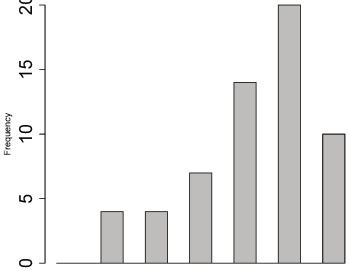
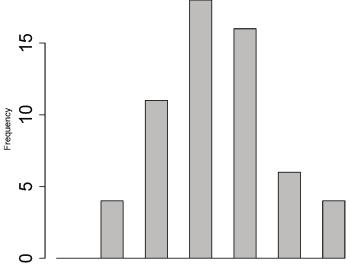
The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>4d)</b> reacts with hardly predictable processing times.</p>	 <table border="1"> <caption>Data for Item 4d: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Choice (1-7)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>6</td></tr> <tr><td>3</td><td>16</td></tr> <tr><td>4</td><td>10</td></tr> <tr><td>5</td><td>16</td></tr> <tr><td>6</td><td>10</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice (1-7)	Frequency	1	1	2	6	3	16	4	10	5	16	6	10	7	0	<p>reacts with predictable processing times.</p>
Agreement Choice (1-7)	Frequency																	
1	1																	
2	6																	
3	16																	
4	10																	
5	16																	
6	10																	
7	0																	
<p><b>4e)</b> cannot be operated following consistent principles.</p>	 <table border="1"> <caption>Data for Item 4e: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Choice (1-7)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>4</td></tr> <tr><td>2</td><td>4</td></tr> <tr><td>3</td><td>7</td></tr> <tr><td>4</td><td>14</td></tr> <tr><td>5</td><td>20</td></tr> <tr><td>6</td><td>10</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice (1-7)	Frequency	1	4	2	4	3	7	4	14	5	20	6	10	7	0	<p>can be operated following consistent principles.</p>
Agreement Choice (1-7)	Frequency																	
1	4																	
2	4																	
3	7																	
4	14																	
5	20																	
6	10																	
7	0																	

Table B.6: Error tolerance (T2)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>5a)</b> is designed in a way that minor mistakes will lead to severe consequences.</p>	 <table border="1"> <caption>Data for Item 5a: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Choice (1-7)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>4</td></tr> <tr><td>2</td><td>11</td></tr> <tr><td>3</td><td>17</td></tr> <tr><td>4</td><td>16</td></tr> <tr><td>5</td><td>6</td></tr> <tr><td>6</td><td>4</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice (1-7)	Frequency	1	4	2	11	3	17	4	16	5	6	6	4	7	0	<p>is designed in a way that minor mistakes will not lead to severe consequences.</p>
Agreement Choice (1-7)	Frequency																	
1	4																	
2	11																	
3	17																	
4	16																	
5	6																	
6	4																	
7	0																	

## B.4 Results of the Usability Study

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>5b)</b> informs too late about invalid input.</p>	<table border="1"> <caption>Frequency of Agreement Choices for Item 5b</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>1</td></tr> <tr><td>3</td><td>4</td></tr> <tr><td>4</td><td>41</td></tr> <tr><td>5</td><td>7</td></tr> <tr><td>6</td><td>3</td></tr> <tr><td>7</td><td>2</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	1	2	1	3	4	4	41	5	7	6	3	7	2	<p>informs at once about invalid input.</p>
Agreement Choice	Frequency																	
1	1																	
2	1																	
3	4																	
4	41																	
5	7																	
6	3																	
7	2																	
<p><b>5c)</b> provides hardly comprehensible error messages.</p>	<table border="1"> <caption>Frequency of Agreement Choices for Item 5c</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>1</td></tr> <tr><td>3</td><td>2</td></tr> <tr><td>4</td><td>39</td></tr> <tr><td>5</td><td>7</td></tr> <tr><td>6</td><td>6</td></tr> <tr><td>7</td><td>4</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	0	2	1	3	2	4	39	5	7	6	6	7	4	<p>provides comprehensible error messages.</p>
Agreement Choice	Frequency																	
1	0																	
2	1																	
3	2																	
4	39																	
5	7																	
6	6																	
7	4																	
<p><b>5d)</b> requires a tremendous effort to correct mistakes on the whole.</p>	<table border="1"> <caption>Frequency of Agreement Choices for Item 5d</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>1</td></tr> <tr><td>3</td><td>11</td></tr> <tr><td>4</td><td>25</td></tr> <tr><td>5</td><td>12</td></tr> <tr><td>6</td><td>6</td></tr> <tr><td>7</td><td>5</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	0	2	1	3	11	4	25	5	12	6	6	7	5	<p>requires little effort to correct mistakes on the whole.</p>
Agreement Choice	Frequency																	
1	0																	
2	1																	
3	11																	
4	25																	
5	12																	
6	6																	
7	5																	
<p><b>5e)</b> does not provide hints how to solve errors.</p>	<table border="1"> <caption>Frequency of Agreement Choices for Item 5e</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>2</td></tr> <tr><td>2</td><td>5</td></tr> <tr><td>3</td><td>10</td></tr> <tr><td>4</td><td>33</td></tr> <tr><td>5</td><td>4</td></tr> <tr><td>6</td><td>3</td></tr> <tr><td>7</td><td>2</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	2	2	5	3	10	4	33	5	4	6	3	7	2	<p>does provide hints how to solve errors.</p>
Agreement Choice	Frequency																	
1	2																	
2	5																	
3	10																	
4	33																	
5	4																	
6	3																	
7	2																	

B Evaluation Appendix

Table B.7: Suitability for individualization (T2)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>6a)</b> is hard to extend by the user when new tasks have to be solved.</p>	<table border="1"> <caption>Data for Item 6a Bar Chart</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>10</td></tr> <tr><td>3</td><td>6</td></tr> <tr><td>4</td><td>34</td></tr> <tr><td>5</td><td>6</td></tr> <tr><td>6</td><td>2</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Level	Frequency	1	0	2	10	3	6	4	34	5	6	6	2	7	0	<p>is easy to extend by the user when new tasks have to be solved.</p>
Level	Frequency																	
1	0																	
2	10																	
3	6																	
4	34																	
5	6																	
6	2																	
7	0																	
<p><b>6b)</b> can hardly be adapted by the user to the individual way of solving a task.</p>	<table border="1"> <caption>Data for Item 6b Bar Chart</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>3</td></tr> <tr><td>2</td><td>7</td></tr> <tr><td>3</td><td>10</td></tr> <tr><td>4</td><td>23</td></tr> <tr><td>5</td><td>11</td></tr> <tr><td>6</td><td>4</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Level	Frequency	1	3	2	7	3	10	4	23	5	11	6	4	7	0	<p>can be adapted by the user to the individual way of solving a task.</p>
Level	Frequency																	
1	3																	
2	7																	
3	10																	
4	23																	
5	11																	
6	4																	
7	0																	
<p><b>6c)</b> is not likewise suitable for both beginners and experts because it cannot be adjusted to the user's state of knowledge easily.</p>	<table border="1"> <caption>Data for Item 6c Bar Chart</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>7</td></tr> <tr><td>2</td><td>16</td></tr> <tr><td>3</td><td>11</td></tr> <tr><td>4</td><td>10</td></tr> <tr><td>5</td><td>6</td></tr> <tr><td>6</td><td>7</td></tr> <tr><td>7</td><td>2</td></tr> </tbody> </table>	Level	Frequency	1	7	2	16	3	11	4	10	5	6	6	7	7	2	<p>is likewise suitable for both beginners and experts because it cannot be adjusted to the user's state of knowledge easily.</p>
Level	Frequency																	
1	7																	
2	16																	
3	11																	
4	10																	
5	6																	
6	7																	
7	2																	

B.4 Results of the Usability Study

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>6d)</b> cannot be adjusted to different work tasks with respect to its core functionality.</p>	<table border="1"> <caption>Data for Item 6d Bar Chart</caption> <thead> <tr> <th>Agreement Choice (1-7)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>6</td></tr> <tr><td>3</td><td>14</td></tr> <tr><td>4</td><td>22</td></tr> <tr><td>5</td><td>9</td></tr> <tr><td>6</td><td>6</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice (1-7)	Frequency	1	1	2	6	3	14	4	22	5	9	6	6	7	0	<p>can be adjusted to different work tasks with respect to its core functionality.</p>
Agreement Choice (1-7)	Frequency																	
1	1																	
2	6																	
3	14																	
4	22																	
5	9																	
6	6																	
7	0																	
<p><b>6e)</b> is designed in a way that users cannot adjust the screen layout to their individual needs.</p>	<table border="1"> <caption>Data for Item 6e Bar Chart</caption> <thead> <tr> <th>Agreement Choice (1-7)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>3</td></tr> <tr><td>2</td><td>7</td></tr> <tr><td>3</td><td>10</td></tr> <tr><td>4</td><td>11</td></tr> <tr><td>5</td><td>13</td></tr> <tr><td>6</td><td>10</td></tr> <tr><td>7</td><td>5</td></tr> </tbody> </table>	Agreement Choice (1-7)	Frequency	1	3	2	7	3	10	4	11	5	13	6	10	7	5	<p>is designed in a way that users can adjust the screen layout to their individual needs.</p>
Agreement Choice (1-7)	Frequency																	
1	3																	
2	7																	
3	10																	
4	11																	
5	13																	
6	10																	
7	5																	

Table B.8: Suitability for learning (T2)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>7a)</b> takes a lot of time to learn.</p>	<table border="1"> <caption>Data for Item 7a Bar Chart</caption> <thead> <tr> <th>Agreement Choice (1-7)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>2</td></tr> <tr><td>2</td><td>8</td></tr> <tr><td>3</td><td>13</td></tr> <tr><td>4</td><td>6</td></tr> <tr><td>5</td><td>7</td></tr> <tr><td>6</td><td>14</td></tr> <tr><td>7</td><td>9</td></tr> </tbody> </table>	Agreement Choice (1-7)	Frequency	1	2	2	8	3	13	4	6	5	7	6	14	7	9	<p>takes little time to learn.</p>
Agreement Choice (1-7)	Frequency																	
1	2																	
2	8																	
3	13																	
4	6																	
5	7																	
6	14																	
7	9																	

## B Evaluation Appendix

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>7b)</b> does not encourage to test new functions.</p>	<table border="1"> <caption>Data for 7b)</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>2</td></tr> <tr><td>2</td><td>6</td></tr> <tr><td>3</td><td>11</td></tr> <tr><td>4</td><td>7</td></tr> <tr><td>5</td><td>17</td></tr> <tr><td>6</td><td>12</td></tr> <tr><td>7</td><td>4</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	2	2	6	3	11	4	7	5	17	6	12	7	4	<p>encourages to test new functions</p>
Agreement Choice	Frequency																	
1	2																	
2	6																	
3	11																	
4	7																	
5	17																	
6	12																	
7	4																	
<p><b>7c)</b> requires users to keep many details in mind.</p>	<table border="1"> <caption>Data for 7c)</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>8</td></tr> <tr><td>3</td><td>22</td></tr> <tr><td>4</td><td>2</td></tr> <tr><td>5</td><td>7</td></tr> <tr><td>6</td><td>12</td></tr> <tr><td>7</td><td>8</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	0	2	8	3	22	4	2	5	7	6	12	7	8	<p>does not require users to keep many details in mind.</p>
Agreement Choice	Frequency																	
1	0																	
2	8																	
3	22																	
4	2																	
5	7																	
6	12																	
7	8																	
<p><b>7d)</b> is designed in a way that learned actions are not easy to keep in mind.</p>	<table border="1"> <caption>Data for 7d)</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>2</td></tr> <tr><td>2</td><td>6</td></tr> <tr><td>3</td><td>12</td></tr> <tr><td>4</td><td>17</td></tr> <tr><td>5</td><td>11</td></tr> <tr><td>6</td><td>11</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	2	2	6	3	12	4	17	5	11	6	11	7	0	<p>is designed in a way that learned actions are easy to keep in mind.</p>
Agreement Choice	Frequency																	
1	2																	
2	6																	
3	12																	
4	17																	
5	11																	
6	11																	
7	0																	
<p><b>7e)</b> cannot be learned with help by others or a manual.</p>	<table border="1"> <caption>Data for 7e)</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>8</td></tr> <tr><td>2</td><td>10</td></tr> <tr><td>3</td><td>8</td></tr> <tr><td>4</td><td>12</td></tr> <tr><td>5</td><td>9</td></tr> <tr><td>6</td><td>8</td></tr> <tr><td>7</td><td>4</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	8	2	10	3	8	4	12	5	9	6	8	7	4	<p>is easy to learn without help by others or a manual.</p>
Agreement Choice	Frequency																	
1	8																	
2	10																	
3	8																	
4	12																	
5	9																	
6	8																	
7	4																	



**B.4.2 Results for the T3 GUI Variant**

Table B.9: Suitability for the task (T3)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>1a)</b> is complicated to use.</p>	<table border="1"> <caption>Data for 1a) is complicated to use.</caption> <thead> <tr> <th>Agreement Choice (Level)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>9</td></tr> <tr><td>3</td><td>5</td></tr> <tr><td>4</td><td>12</td></tr> <tr><td>5</td><td>22</td></tr> <tr><td>6</td><td>11</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice (Level)	Frequency	1	0	2	9	3	5	4	12	5	22	6	11	7	0	<p>is uncomplicated to use.</p>
Agreement Choice (Level)	Frequency																	
1	0																	
2	9																	
3	5																	
4	12																	
5	22																	
6	11																	
7	0																	
<p><b>1b)</b> does not offer all functions needed to solve the task efficiently.</p>	<table border="1"> <caption>Data for 1b) does not offer all functions needed to solve the task efficiently.</caption> <thead> <tr> <th>Agreement Choice (Level)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>3</td></tr> <tr><td>2</td><td>10</td></tr> <tr><td>3</td><td>12</td></tr> <tr><td>4</td><td>9</td></tr> <tr><td>5</td><td>12</td></tr> <tr><td>6</td><td>11</td></tr> <tr><td>7</td><td>2</td></tr> </tbody> </table>	Agreement Choice (Level)	Frequency	1	3	2	10	3	12	4	9	5	12	6	11	7	2	<p>does offer all functions needed to solve the task efficiently.</p>
Agreement Choice (Level)	Frequency																	
1	3																	
2	10																	
3	12																	
4	9																	
5	12																	
6	11																	
7	2																	
<p><b>1c)</b> offers little support to automate reoccurring tasks.</p>	<table border="1"> <caption>Data for 1c) offers little support to automate reoccurring tasks.</caption> <thead> <tr> <th>Agreement Choice (Level)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>2</td></tr> <tr><td>2</td><td>14</td></tr> <tr><td>3</td><td>16</td></tr> <tr><td>4</td><td>13</td></tr> <tr><td>5</td><td>12</td></tr> <tr><td>6</td><td>2</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice (Level)	Frequency	1	2	2	14	3	16	4	13	5	12	6	2	7	0	<p>offers good support to automate reoccurring tasks.</p>
Agreement Choice (Level)	Frequency																	
1	2																	
2	14																	
3	16																	
4	13																	
5	12																	
6	2																	
7	0																	

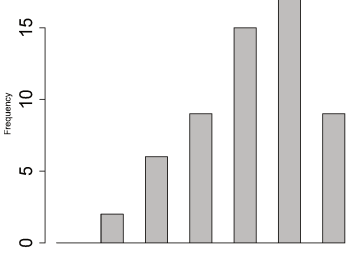
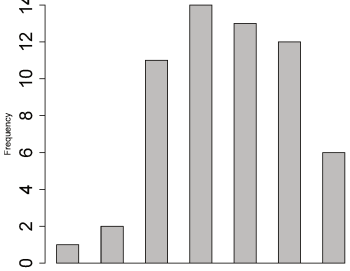
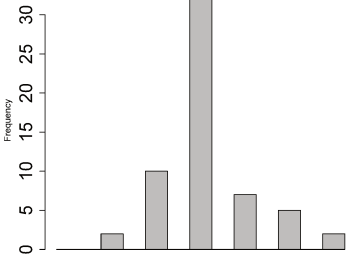
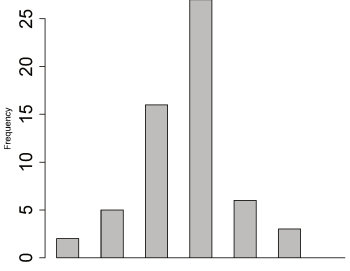
## B Evaluation Appendix

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)
1d) requires superfluous input.	<p>A bar chart with a y-axis labeled 'Frequency' ranging from 0 to 15. The x-axis represents a 7-level Likert scale. The bars show the following frequencies: 1 (1), 2 (5), 3 (13), 4 (17), 5 (16), 6 (7), and 7 (0).</p>	requires no superfluous input.
1e) does not meet the requirements of the task.	<p>A bar chart with a y-axis labeled 'Frequency' ranging from 0 to 15. The x-axis represents a 7-level Likert scale. The bars show the following frequencies: 1 (2), 2 (4), 3 (6), 4 (13), 5 (17), 6 (15), and 7 (2).</p>	meets the requirements of the task.

Table B.10: Self-descriptiveness (T3)

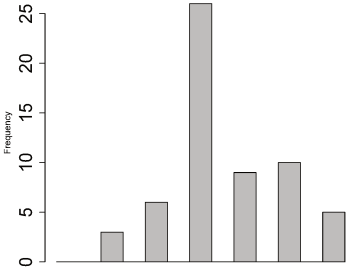
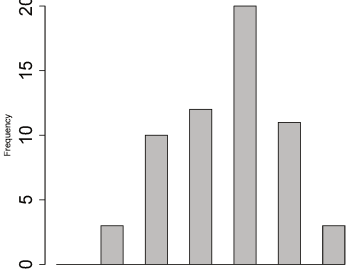
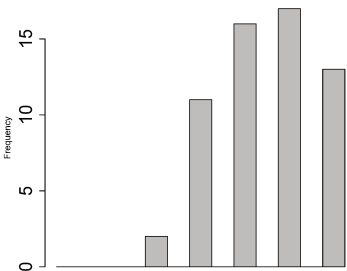
The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)
2a) provides a bad overview over its functions.	<p>A bar chart with a y-axis labeled 'Frequency' ranging from 0 to 20. The x-axis represents a 7-level Likert scale. The bars show the following frequencies: 1 (2), 2 (10), 3 (4), 4 (22), 5 (14), 6 (7), and 7 (0).</p>	provides a good overview over its functions.

## B.4 Results of the Usability Study

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)
<p><b>2b)</b> uses hardly comprehensible terms, abbreviations, or symbols in its interface.</p>		<p>uses comprehensible terms, abbreviations, or symbols in its interface.</p>
<p><b>2c)</b> offers insufficient information about valid and required input.</p>		<p>offers sufficient information about valid and required input.</p>
<p><b>2d)</b> offers no context-specific explanations that are helpful on demand.</p>		<p>offers context-specific explanations that are helpful on demand.</p>
<p><b>2e)</b> offers no automatic context-specific explanations that are helpful.</p>		<p>offers no automatic context-specific explanations that are helpful.</p>

B Evaluation Appendix

Table B.11: Controllability (T3)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)
<p><b>3a)</b> offers no means to pause a task and to continue later on without losing the current progress.</p>		<p>offers means to pause a task and to continue later on without losing the current progress.</p>
<p><b>3b)</b> forces a needless and fixed interaction process onto the user.</p>		<p>does not force a needless and fixed interaction process onto the user.</p>
<p><b>3c)</b> allows only complicated switching between interface elements.</p>		<p>allows uncomplicated switching between interface elements.</p>

## B.4 Results of the Usability Study

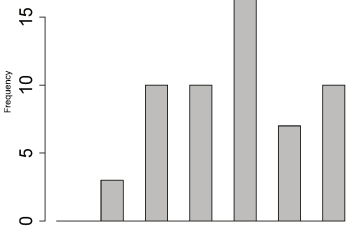
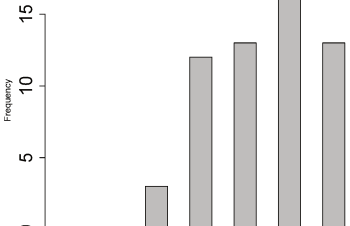
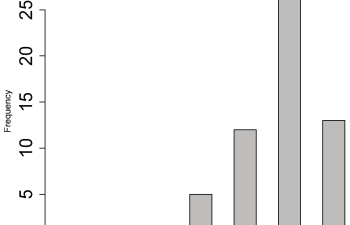
The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)														
<p><b>3d)</b> is designed in a way that users cannot control how and what information is displayed on the screen.</p>	 <table border="1"> <caption>Data for Item 3d: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>3</td></tr> <tr><td>2</td><td>10</td></tr> <tr><td>3</td><td>10</td></tr> <tr><td>4</td><td>18</td></tr> <tr><td>5</td><td>7</td></tr> <tr><td>6</td><td>10</td></tr> </tbody> </table>	Agreement Level	Frequency	1	3	2	10	3	10	4	18	5	7	6	10	<p>is designed in a way that users can control how and what information is displayed on the screen.</p>
Agreement Level	Frequency															
1	3															
2	10															
3	10															
4	18															
5	7															
6	10															
<p><b>3e)</b> forces needless interruptions of the work onto the user.</p>	 <table border="1"> <caption>Data for Item 3e: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>2</td><td>3</td></tr> <tr><td>3</td><td>12</td></tr> <tr><td>4</td><td>13</td></tr> <tr><td>5</td><td>18</td></tr> <tr><td>6</td><td>13</td></tr> </tbody> </table>	Agreement Level	Frequency	2	3	3	12	4	13	5	18	6	13	<p>does not force needless interruptions of the work onto the user.</p>		
Agreement Level	Frequency															
2	3															
3	12															
4	13															
5	18															
6	13															

Table B.12: Conformity with user expectations (T3)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)												
<p><b>4a)</b> makes orienting difficult due to its inconsistent design.</p>	 <table border="1"> <caption>Data for Item 4a: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>5</td></tr> <tr><td>3</td><td>12</td></tr> <tr><td>4</td><td>28</td></tr> <tr><td>5</td><td>13</td></tr> </tbody> </table>	Agreement Level	Frequency	1	1	2	5	3	12	4	28	5	13	<p>facilitates orienting due to its consistent design.</p>
Agreement Level	Frequency													
1	1													
2	5													
3	12													
4	28													
5	13													

## B Evaluation Appendix

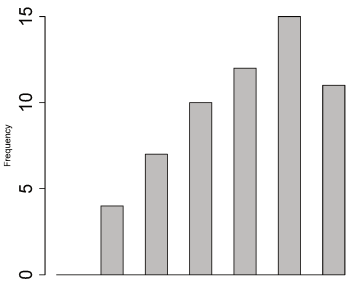
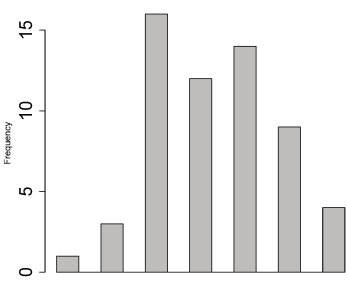
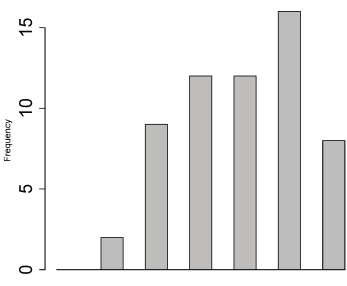
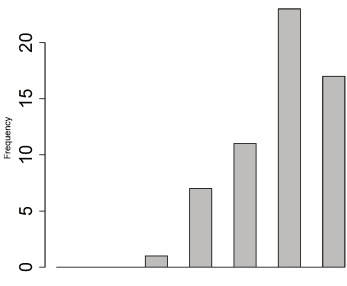
The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)
<p><b>4b)</b> keeps users in doubt whether an input was successful or not.</p>		<p>does not keep users in doubt whether an input was successful or not.</p>
<p><b>4c)</b> informs the user about its current state insufficiently.</p>		<p>informs the user about its current state sufficiently.</p>
<p><b>4d)</b> reacts with hardly predictable processing times.</p>		<p>reacts with predictable processing times.</p>
<p><b>4e)</b> cannot be operated following consistent principles.</p>		<p>can be operated following consistent principles.</p>

Table B.13: Error tolerance (T3)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>5a)</b> is designed in a way that minor mistakes will lead to severe consequences.</p>	<table border="1"> <caption>Data for 5a Bar Chart</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>4</td></tr> <tr><td>3</td><td>10</td></tr> <tr><td>4</td><td>16</td></tr> <tr><td>5</td><td>11</td></tr> <tr><td>6</td><td>10</td></tr> <tr><td>7</td><td>7</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	1	2	4	3	10	4	16	5	11	6	10	7	7	<p>is designed in a way that minor mistakes will not lead to severe consequences.</p>
Agreement Choice	Frequency																	
1	1																	
2	4																	
3	10																	
4	16																	
5	11																	
6	10																	
7	7																	
<p><b>5b)</b> informs too late about invalid input.</p>	<table border="1"> <caption>Data for 5b Bar Chart</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>3</td></tr> <tr><td>3</td><td>2</td></tr> <tr><td>4</td><td>32</td></tr> <tr><td>5</td><td>9</td></tr> <tr><td>6</td><td>8</td></tr> <tr><td>7</td><td>5</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	0	2	3	3	2	4	32	5	9	6	8	7	5	<p>informs at once about invalid input.</p>
Agreement Choice	Frequency																	
1	0																	
2	3																	
3	2																	
4	32																	
5	9																	
6	8																	
7	5																	
<p><b>5c)</b> provides hardly comprehensible error messages.</p>	<table border="1"> <caption>Data for 5c Bar Chart</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>6</td></tr> <tr><td>3</td><td>0</td></tr> <tr><td>4</td><td>31</td></tr> <tr><td>5</td><td>7</td></tr> <tr><td>6</td><td>11</td></tr> <tr><td>7</td><td>3</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	1	2	6	3	0	4	31	5	7	6	11	7	3	<p>provides comprehensible error messages.</p>
Agreement Choice	Frequency																	
1	1																	
2	6																	
3	0																	
4	31																	
5	7																	
6	11																	
7	3																	

## B Evaluation Appendix

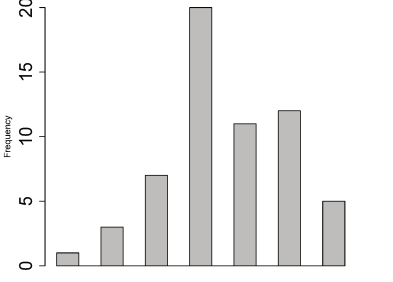
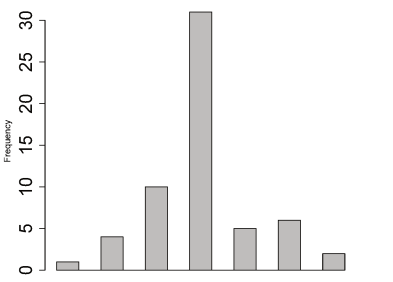
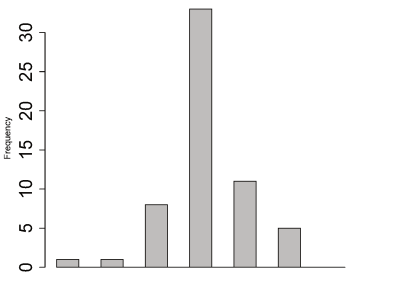
The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>5d)</b> requires a tremendous effort to correct mistakes on the whole.</p>	 <table border="1" data-bbox="678 324 1072 604"> <caption>Data for 5d)</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>3</td></tr> <tr><td>3</td><td>7</td></tr> <tr><td>4</td><td>20</td></tr> <tr><td>5</td><td>11</td></tr> <tr><td>6</td><td>12</td></tr> <tr><td>7</td><td>5</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	1	2	3	3	7	4	20	5	11	6	12	7	5	<p>requires little effort to correct mistakes on the whole.</p>
Agreement Choice	Frequency																	
1	1																	
2	3																	
3	7																	
4	20																	
5	11																	
6	12																	
7	5																	
<p><b>5e)</b> does not provide hints how to solve errors.</p>	 <table border="1" data-bbox="678 698 1072 978"> <caption>Data for 5e)</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>4</td></tr> <tr><td>3</td><td>10</td></tr> <tr><td>4</td><td>31</td></tr> <tr><td>5</td><td>5</td></tr> <tr><td>6</td><td>6</td></tr> <tr><td>7</td><td>2</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	1	2	4	3	10	4	31	5	5	6	6	7	2	<p>does provide hints how to solve errors.</p>
Agreement Choice	Frequency																	
1	1																	
2	4																	
3	10																	
4	31																	
5	5																	
6	6																	
7	2																	

Table B.14: Suitability for individualization (T3)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)														
<p><b>6a)</b> is hard to extend by the user when new tasks have to be solved.</p>	 <table border="1" data-bbox="678 1249 1072 1529"> <caption>Data for 6a)</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>1</td></tr> <tr><td>3</td><td>8</td></tr> <tr><td>4</td><td>33</td></tr> <tr><td>5</td><td>11</td></tr> <tr><td>6</td><td>5</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	1	2	1	3	8	4	33	5	11	6	5	<p>is easy to extend by the user when new tasks have to be solved.</p>
Agreement Choice	Frequency															
1	1															
2	1															
3	8															
4	33															
5	11															
6	5															

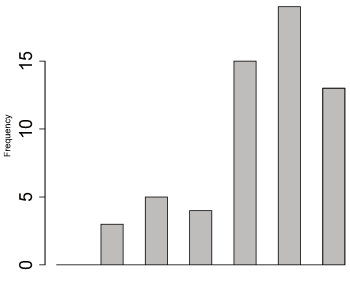
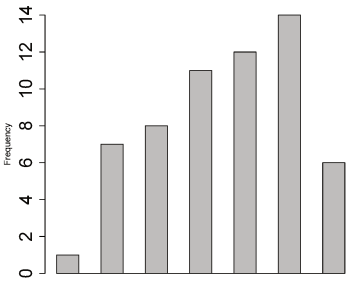
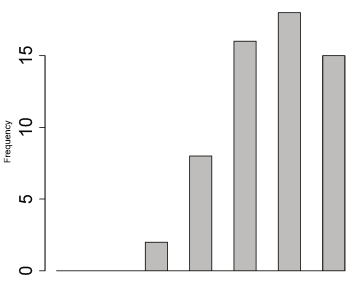


## B.4 Results of the Usability Study

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>6b)</b> can hardly be adapted by the user to the individual way of solving a task.</p>	<table border="1"> <caption>Frequency of Agreement Choices for Item 6b</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>2</td></tr> <tr><td>2</td><td>3</td></tr> <tr><td>3</td><td>12</td></tr> <tr><td>4</td><td>21</td></tr> <tr><td>5</td><td>8</td></tr> <tr><td>6</td><td>10</td></tr> <tr><td>7</td><td>3</td></tr> </tbody> </table>	Level	Frequency	1	2	2	3	3	12	4	21	5	8	6	10	7	3	<p>can be adapted by the user to the individual way of solving a task.</p>
Level	Frequency																	
1	2																	
2	3																	
3	12																	
4	21																	
5	8																	
6	10																	
7	3																	
<p><b>6c)</b> is not likewise suitable for both beginners and experts because it cannot be adjusted to the user's state of knowledge easily.</p>	<table border="1"> <caption>Frequency of Agreement Choices for Item 6c</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>3</td></tr> <tr><td>2</td><td>4</td></tr> <tr><td>3</td><td>8</td></tr> <tr><td>4</td><td>10</td></tr> <tr><td>5</td><td>17</td></tr> <tr><td>6</td><td>9</td></tr> <tr><td>7</td><td>8</td></tr> </tbody> </table>	Level	Frequency	1	3	2	4	3	8	4	10	5	17	6	9	7	8	<p>is likewise suitable for both beginners and experts because it cannot be adjusted to the user's state of knowledge easily.</p>
Level	Frequency																	
1	3																	
2	4																	
3	8																	
4	10																	
5	17																	
6	9																	
7	8																	
<p><b>6d)</b> cannot be adjusted to different work tasks with respect to its core functionality.</p>	<table border="1"> <caption>Frequency of Agreement Choices for Item 6d</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>2</td></tr> <tr><td>3</td><td>12</td></tr> <tr><td>4</td><td>20</td></tr> <tr><td>5</td><td>13</td></tr> <tr><td>6</td><td>11</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Level	Frequency	1	1	2	2	3	12	4	20	5	13	6	11	7	0	<p>can be adjusted to different work tasks with respect to its core functionality.</p>
Level	Frequency																	
1	1																	
2	2																	
3	12																	
4	20																	
5	13																	
6	11																	
7	0																	
<p><b>6e)</b> is designed in a way that users cannot adjust the screen layout to their individual needs.</p>	<table border="1"> <caption>Frequency of Agreement Choices for Item 6e</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>2</td></tr> <tr><td>3</td><td>6</td></tr> <tr><td>4</td><td>12</td></tr> <tr><td>5</td><td>17</td></tr> <tr><td>6</td><td>14</td></tr> <tr><td>7</td><td>7</td></tr> </tbody> </table>	Level	Frequency	1	0	2	2	3	6	4	12	5	17	6	14	7	7	<p>is designed in a way that users can adjust the screen layout to their individual needs.</p>
Level	Frequency																	
1	0																	
2	2																	
3	6																	
4	12																	
5	17																	
6	14																	
7	7																	

B Evaluation Appendix

Table B.15: Suitability for learning (T3)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>7a)</b> takes a lot of time to learn.</p>	 <table border="1" data-bbox="678 380 1029 660"> <caption>Data for 7a Bar Chart</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>3</td></tr> <tr><td>3</td><td>5</td></tr> <tr><td>4</td><td>4</td></tr> <tr><td>5</td><td>15</td></tr> <tr><td>6</td><td>18</td></tr> <tr><td>7</td><td>13</td></tr> </tbody> </table>	Level	Frequency	1	0	2	3	3	5	4	4	5	15	6	18	7	13	<p>takes little time to learn.</p>
Level	Frequency																	
1	0																	
2	3																	
3	5																	
4	4																	
5	15																	
6	18																	
7	13																	
<p><b>7b)</b> does not encourage to test new functions.</p>	 <table border="1" data-bbox="678 750 1029 1030"> <caption>Data for 7b Bar Chart</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>7</td></tr> <tr><td>3</td><td>8</td></tr> <tr><td>4</td><td>11</td></tr> <tr><td>5</td><td>12</td></tr> <tr><td>6</td><td>14</td></tr> <tr><td>7</td><td>6</td></tr> </tbody> </table>	Level	Frequency	1	1	2	7	3	8	4	11	5	12	6	14	7	6	<p>encourages to test new functions</p>
Level	Frequency																	
1	1																	
2	7																	
3	8																	
4	11																	
5	12																	
6	14																	
7	6																	
<p><b>7c)</b> requires users to keep many details in mind.</p>	 <table border="1" data-bbox="678 1131 1029 1411"> <caption>Data for 7c Bar Chart</caption> <thead> <tr> <th>Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>2</td></tr> <tr><td>3</td><td>8</td></tr> <tr><td>4</td><td>16</td></tr> <tr><td>5</td><td>18</td></tr> <tr><td>6</td><td>15</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Level	Frequency	1	0	2	2	3	8	4	16	5	18	6	15	7	0	<p>does not require users to keep many details in mind.</p>
Level	Frequency																	
1	0																	
2	2																	
3	8																	
4	16																	
5	18																	
6	15																	
7	0																	

## B.4 Results of the Usability Study

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
7d) is designed in a way that learned actions are not easy to keep in mind.	<table border="1"> <caption>Data for 7d) Bar Chart</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>1</td></tr> <tr><td>3</td><td>6</td></tr> <tr><td>4</td><td>16</td></tr> <tr><td>5</td><td>23</td></tr> <tr><td>6</td><td>13</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	0	2	1	3	6	4	16	5	23	6	13	7	0	is designed in a way that learned actions are easy to keep in mind.
Agreement Choice	Frequency																	
1	0																	
2	1																	
3	6																	
4	16																	
5	23																	
6	13																	
7	0																	
7e) cannot be learned with help by others or a manual.	<table border="1"> <caption>Data for 7e) Bar Chart</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>4</td></tr> <tr><td>3</td><td>6</td></tr> <tr><td>4</td><td>13</td></tr> <tr><td>5</td><td>12</td></tr> <tr><td>6</td><td>13</td></tr> <tr><td>7</td><td>10</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	1	2	4	3	6	4	13	5	12	6	13	7	10	is easy to learn without help by others or a manual.
Agreement Choice	Frequency																	
1	1																	
2	4																	
3	6																	
4	13																	
5	12																	
6	13																	
7	10																	

### B.4.3 Results for the T4 GUI Variant

Table B.16: Suitability for the task (T4)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
1a) is complicated to use.	<table border="1"> <caption>Data for 1a) Bar Chart</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>2</td></tr> <tr><td>2</td><td>7</td></tr> <tr><td>3</td><td>15</td></tr> <tr><td>4</td><td>28</td></tr> <tr><td>5</td><td>7</td></tr> <tr><td>6</td><td>0</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	2	2	7	3	15	4	28	5	7	6	0	7	0	is uncomplicated to use.
Agreement Choice	Frequency																	
1	2																	
2	7																	
3	15																	
4	28																	
5	7																	
6	0																	
7	0																	

## B Evaluation Appendix

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>1b)</b> does not offer all functions needed to solve the task efficiently.</p>	<table border="1"> <caption>Data for Item 1b</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>3</td></tr> <tr><td>2</td><td>5</td></tr> <tr><td>3</td><td>9</td></tr> <tr><td>4</td><td>17</td></tr> <tr><td>5</td><td>16</td></tr> <tr><td>6</td><td>16</td></tr> <tr><td>7</td><td>9</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	3	2	5	3	9	4	17	5	16	6	16	7	9	<p>does offer all functions needed to solve the task efficiently.</p>
Agreement Choice	Frequency																	
1	3																	
2	5																	
3	9																	
4	17																	
5	16																	
6	16																	
7	9																	
<p><b>1c)</b> offers little support to automate reoccurring tasks.</p>	<table border="1"> <caption>Data for Item 1c</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>5</td></tr> <tr><td>2</td><td>10</td></tr> <tr><td>3</td><td>17</td></tr> <tr><td>4</td><td>16</td></tr> <tr><td>5</td><td>8</td></tr> <tr><td>6</td><td>3</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	5	2	10	3	17	4	16	5	8	6	3	7	0	<p>offers good support to automate reoccurring tasks.</p>
Agreement Choice	Frequency																	
1	5																	
2	10																	
3	17																	
4	16																	
5	8																	
6	3																	
7	0																	
<p><b>1d)</b> requires superfluous input.</p>	<table border="1"> <caption>Data for Item 1d</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>2</td></tr> <tr><td>3</td><td>8</td></tr> <tr><td>4</td><td>8</td></tr> <tr><td>5</td><td>15</td></tr> <tr><td>6</td><td>21</td></tr> <tr><td>7</td><td>4</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	1	2	2	3	8	4	8	5	15	6	21	7	4	<p>requires no superfluous input.</p>
Agreement Choice	Frequency																	
1	1																	
2	2																	
3	8																	
4	8																	
5	15																	
6	21																	
7	4																	
<p><b>1e)</b> does not meet the requirements of the task.</p>	<table border="1"> <caption>Data for Item 1e</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>3</td></tr> <tr><td>2</td><td>2</td></tr> <tr><td>3</td><td>12</td></tr> <tr><td>4</td><td>15</td></tr> <tr><td>5</td><td>22</td></tr> <tr><td>6</td><td>5</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	3	2	2	3	12	4	15	5	22	6	5	7	0	<p>meets the requirements of the task.</p>
Agreement Choice	Frequency																	
1	3																	
2	2																	
3	12																	
4	15																	
5	22																	
6	5																	
7	0																	

Table B.17: Self-descriptiveness (T4)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>2a)</b> provides a bad overview over its functions.</p>	<table border="1"> <caption>Data for Statement 2a</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>1</td></tr> <tr><td>3</td><td>10</td></tr> <tr><td>4</td><td>6</td></tr> <tr><td>5</td><td>22</td></tr> <tr><td>6</td><td>16</td></tr> <tr><td>7</td><td>4</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	0	2	1	3	10	4	6	5	22	6	16	7	4	<p>provides a good overview over its functions.</p>
Agreement Choice	Frequency																	
1	0																	
2	1																	
3	10																	
4	6																	
5	22																	
6	16																	
7	4																	
<p><b>2b)</b> uses hardly comprehensible terms, abbreviations, or symbols in its interface.</p>	<table border="1"> <caption>Data for Statement 2b</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>5</td></tr> <tr><td>3</td><td>15</td></tr> <tr><td>4</td><td>7</td></tr> <tr><td>5</td><td>18</td></tr> <tr><td>6</td><td>9</td></tr> <tr><td>7</td><td>4</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	1	2	5	3	15	4	7	5	18	6	9	7	4	<p>uses comprehensible terms, abbreviations, or symbols in its interface.</p>
Agreement Choice	Frequency																	
1	1																	
2	5																	
3	15																	
4	7																	
5	18																	
6	9																	
7	4																	
<p><b>2c)</b> offers insufficient information about valid and required input.</p>	<table border="1"> <caption>Data for Statement 2c</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>2</td></tr> <tr><td>2</td><td>2</td></tr> <tr><td>3</td><td>13</td></tr> <tr><td>4</td><td>13</td></tr> <tr><td>5</td><td>14</td></tr> <tr><td>6</td><td>12</td></tr> <tr><td>7</td><td>3</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	2	2	2	3	13	4	13	5	14	6	12	7	3	<p>offers sufficient information about valid and required input.</p>
Agreement Choice	Frequency																	
1	2																	
2	2																	
3	13																	
4	13																	
5	14																	
6	12																	
7	3																	

## B Evaluation Appendix

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>2d)</b> offers no context-specific explanations that are helpful on demand.</p>	<table border="1"> <caption>Data for Item 2d: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Choice (1-7)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>4</td></tr> <tr><td>3</td><td>13</td></tr> <tr><td>4</td><td>27</td></tr> <tr><td>5</td><td>11</td></tr> <tr><td>6</td><td>3</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice (1-7)	Frequency	1	1	2	4	3	13	4	27	5	11	6	3	7	0	<p>offers context-specific explanations that are helpful on demand.</p>
Agreement Choice (1-7)	Frequency																	
1	1																	
2	4																	
3	13																	
4	27																	
5	11																	
6	3																	
7	0																	
<p><b>2e)</b> offers no automatic context-specific explanations that are helpful.</p>	<table border="1"> <caption>Data for Item 2e: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Choice (1-7)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>7</td></tr> <tr><td>3</td><td>14</td></tr> <tr><td>4</td><td>23</td></tr> <tr><td>5</td><td>7</td></tr> <tr><td>6</td><td>5</td></tr> <tr><td>7</td><td>1</td></tr> </tbody> </table>	Agreement Choice (1-7)	Frequency	1	1	2	7	3	14	4	23	5	7	6	5	7	1	<p>offers no automatic context-specific explanations that are helpful.</p>
Agreement Choice (1-7)	Frequency																	
1	1																	
2	7																	
3	14																	
4	23																	
5	7																	
6	5																	
7	1																	

Table B.18: Controllability (T4)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>3a)</b> offers no means to pause a task and to continue later on without losing the current progress.</p>	<table border="1"> <caption>Data for Item 3a: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Choice (1-7)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>3</td></tr> <tr><td>3</td><td>5</td></tr> <tr><td>4</td><td>23</td></tr> <tr><td>5</td><td>8</td></tr> <tr><td>6</td><td>14</td></tr> <tr><td>7</td><td>5</td></tr> </tbody> </table>	Agreement Choice (1-7)	Frequency	1	0	2	3	3	5	4	23	5	8	6	14	7	5	<p>offers means to pause a task and to continue later on without losing the current progress.</p>
Agreement Choice (1-7)	Frequency																	
1	0																	
2	3																	
3	5																	
4	23																	
5	8																	
6	14																	
7	5																	

## B.4 Results of the Usability Study

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>3b)</b> forces a needless and fixed interaction process onto the user.</p>	<table border="1"> <caption>Data for Item 3b</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>2</td></tr> <tr><td>2</td><td>6</td></tr> <tr><td>3</td><td>11</td></tr> <tr><td>4</td><td>16</td></tr> <tr><td>5</td><td>20</td></tr> <tr><td>6</td><td>18</td></tr> <tr><td>7</td><td>4</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	2	2	6	3	11	4	16	5	20	6	18	7	4	<p>does not force a needless and fixed interaction process onto the user.</p>
Agreement Choice	Frequency																	
1	2																	
2	6																	
3	11																	
4	16																	
5	20																	
6	18																	
7	4																	
<p><b>3c)</b> allows only complicated switching between interface elements.</p>	<table border="1"> <caption>Data for Item 3c</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>1</td></tr> <tr><td>3</td><td>2</td></tr> <tr><td>4</td><td>7</td></tr> <tr><td>5</td><td>12</td></tr> <tr><td>6</td><td>23</td></tr> <tr><td>7</td><td>13</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	1	2	1	3	2	4	7	5	12	6	23	7	13	<p>allows uncomplicated switching between interface elements.</p>
Agreement Choice	Frequency																	
1	1																	
2	1																	
3	2																	
4	7																	
5	12																	
6	23																	
7	13																	
<p><b>3d)</b> is designed in a way that users cannot control how and what information is displayed on the screen.</p>	<table border="1"> <caption>Data for Item 3d</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>1</td></tr> <tr><td>3</td><td>5</td></tr> <tr><td>4</td><td>5</td></tr> <tr><td>5</td><td>19</td></tr> <tr><td>6</td><td>22</td></tr> <tr><td>7</td><td>6</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	1	2	1	3	5	4	5	5	19	6	22	7	6	<p>is designed in a way that users can control how and what information is displayed on the screen.</p>
Agreement Choice	Frequency																	
1	1																	
2	1																	
3	5																	
4	5																	
5	19																	
6	22																	
7	6																	
<p><b>3e)</b> forces needless interruptions of the work onto the user.</p>	<table border="1"> <caption>Data for Item 3e</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>1</td></tr> <tr><td>3</td><td>14</td></tr> <tr><td>4</td><td>12</td></tr> <tr><td>5</td><td>24</td></tr> <tr><td>6</td><td>8</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	1	2	1	3	14	4	12	5	24	6	8	<p>does not force needless interruptions of the work onto the user.</p>		
Agreement Choice	Frequency																	
1	1																	
2	1																	
3	14																	
4	12																	
5	24																	
6	8																	

B Evaluation Appendix

Table B.19: Conformity with user expectations (T4)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>4a)</b> makes orienting difficult due to its inconsistent design.</p>	<table border="1"> <caption>Data for Item 4a Bar Chart</caption> <thead> <tr> <th>Agreement Choice (Level)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>2</td></tr> <tr><td>2</td><td>3</td></tr> <tr><td>3</td><td>3</td></tr> <tr><td>4</td><td>10</td></tr> <tr><td>5</td><td>30</td></tr> <tr><td>6</td><td>11</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice (Level)	Frequency	1	2	2	3	3	3	4	10	5	30	6	11	7	0	<p>facilitates orienting due to its consistent design.</p>
Agreement Choice (Level)	Frequency																	
1	2																	
2	3																	
3	3																	
4	10																	
5	30																	
6	11																	
7	0																	
<p><b>4b)</b> keeps users in doubt whether an input was successful or not.</p>	<table border="1"> <caption>Data for Item 4b Bar Chart</caption> <thead> <tr> <th>Agreement Choice (Level)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>8</td></tr> <tr><td>3</td><td>15</td></tr> <tr><td>4</td><td>16</td></tr> <tr><td>5</td><td>13</td></tr> <tr><td>6</td><td>6</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice (Level)	Frequency	1	1	2	8	3	15	4	16	5	13	6	6	7	0	<p>does not keep users in doubt whether an input was successful or not.</p>
Agreement Choice (Level)	Frequency																	
1	1																	
2	8																	
3	15																	
4	16																	
5	13																	
6	6																	
7	0																	
<p><b>4c)</b> informs the user about its current state insufficiently.</p>	<table border="1"> <caption>Data for Item 4c Bar Chart</caption> <thead> <tr> <th>Agreement Choice (Level)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>4</td></tr> <tr><td>2</td><td>11</td></tr> <tr><td>3</td><td>14</td></tr> <tr><td>4</td><td>15</td></tr> <tr><td>5</td><td>13</td></tr> <tr><td>6</td><td>2</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice (Level)	Frequency	1	4	2	11	3	14	4	15	5	13	6	2	7	0	<p>informs the user about its current state sufficiently.</p>
Agreement Choice (Level)	Frequency																	
1	4																	
2	11																	
3	14																	
4	15																	
5	13																	
6	2																	
7	0																	



B.4 Results of the Usability Study

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)
<p><b>4d)</b> reacts with hardly predictable processing times.</p>		<p>reacts with predictable processing times.</p>
<p><b>4e)</b> cannot be operated following consistent principles.</p>		<p>can be operated following consistent principles.</p>

Table B.20: Error tolerance (T4)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)
<p><b>5a)</b> is designed in a way that minor mistakes will lead to severe consequences.</p>		<p>is designed in a way that minor mistakes will not lead to severe consequences.</p>

## B Evaluation Appendix

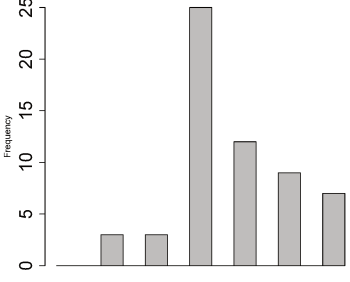
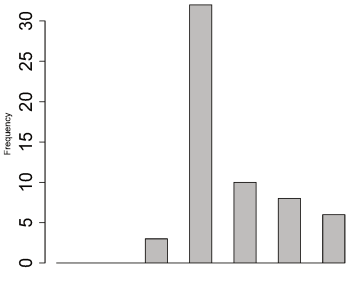
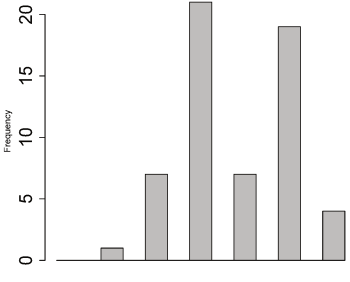
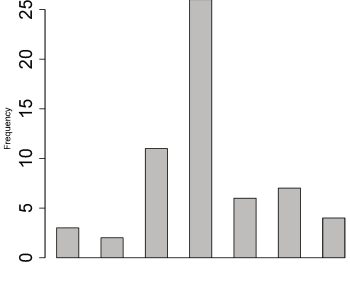
The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)
<p><b>5b)</b> informs too late about invalid input.</p>		<p>informs at once about invalid input.</p>
<p><b>5c)</b> provides hardly comprehensible error messages.</p>		<p>provides comprehensible error messages.</p>
<p><b>5d)</b> requires a tremendous effort to correct mistakes on the whole.</p>		<p>requires little effort to correct mistakes on the whole.</p>
<p><b>5e)</b> does not provide hints how to solve errors.</p>		<p>does provide hints how to solve errors.</p>

Table B.21: Suitability for individualization (T4)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>6a)</b> is hard to extend by the user when new tasks have to be solved.</p>	<table border="1"> <caption>Data for Item 6a: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>4</td></tr> <tr><td>3</td><td>36</td></tr> <tr><td>4</td><td>12</td></tr> <tr><td>5</td><td>5</td></tr> <tr><td>6</td><td>2</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Level	Frequency	1	0	2	4	3	36	4	12	5	5	6	2	7	0	<p>is easy to extend by the user when new tasks have to be solved.</p>
Agreement Level	Frequency																	
1	0																	
2	4																	
3	36																	
4	12																	
5	5																	
6	2																	
7	0																	
<p><b>6b)</b> can hardly be adapted by the user to the individual way of solving a task.</p>	<table border="1"> <caption>Data for Item 6b: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>5</td></tr> <tr><td>3</td><td>6</td></tr> <tr><td>4</td><td>18</td></tr> <tr><td>5</td><td>16</td></tr> <tr><td>6</td><td>11</td></tr> <tr><td>7</td><td>1</td></tr> </tbody> </table>	Agreement Level	Frequency	1	1	2	5	3	6	4	18	5	16	6	11	7	1	<p>can be adapted by the user to the individual way of solving a task.</p>
Agreement Level	Frequency																	
1	1																	
2	5																	
3	6																	
4	18																	
5	16																	
6	11																	
7	1																	
<p><b>6c)</b> is not likewise suitable for both beginners and experts because it cannot be adjusted to the user's state of knowledge easily.</p>	<table border="1"> <caption>Data for Item 6c: Frequency of Agreement Choices</caption> <thead> <tr> <th>Agreement Level</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>4</td></tr> <tr><td>3</td><td>15</td></tr> <tr><td>4</td><td>7</td></tr> <tr><td>5</td><td>8</td></tr> <tr><td>6</td><td>17</td></tr> <tr><td>7</td><td>7</td></tr> </tbody> </table>	Agreement Level	Frequency	1	1	2	4	3	15	4	7	5	8	6	17	7	7	<p>is likewise suitable for both beginners and experts because it cannot be adjusted to the user's state of knowledge easily.</p>
Agreement Level	Frequency																	
1	1																	
2	4																	
3	15																	
4	7																	
5	8																	
6	17																	
7	7																	

B Evaluation Appendix

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)														
<p><b>6d)</b> cannot be adjusted to different work tasks with respect to its core functionality.</p>	<table border="1"> <caption>Data for Item 6d Bar Chart</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>2</td></tr> <tr><td>2</td><td>10</td></tr> <tr><td>3</td><td>18</td></tr> <tr><td>4</td><td>16</td></tr> <tr><td>5</td><td>13</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	2	2	10	3	18	4	16	5	13	<p>can be adjusted to different work tasks with respect to its core functionality.</p>		
Agreement Choice	Frequency															
1	2															
2	10															
3	18															
4	16															
5	13															
<p><b>6e)</b> is designed in a way that users cannot adjust the screen layout to their individual needs.</p>	<table border="1"> <caption>Data for Item 6e Bar Chart</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>2</td></tr> <tr><td>2</td><td>9</td></tr> <tr><td>3</td><td>12</td></tr> <tr><td>4</td><td>14</td></tr> <tr><td>5</td><td>11</td></tr> <tr><td>6</td><td>11</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	2	2	9	3	12	4	14	5	11	6	11	<p>is designed in a way that users can adjust the screen layout to their individual needs.</p>
Agreement Choice	Frequency															
1	2															
2	9															
3	12															
4	14															
5	11															
6	11															

Table B.22: Suitability for learning (T4)

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)														
<p><b>7a)</b> takes a lot of time to learn.</p>	<table border="1"> <caption>Data for Item 7a Bar Chart</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>6</td></tr> <tr><td>3</td><td>5</td></tr> <tr><td>4</td><td>20</td></tr> <tr><td>5</td><td>19</td></tr> <tr><td>6</td><td>8</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	1	2	6	3	5	4	20	5	19	6	8	<p>takes little time to learn.</p>
Agreement Choice	Frequency															
1	1															
2	6															
3	5															
4	20															
5	19															
6	8															

## B.4 Results of the Usability Study

The software... (bad)	Frequency of the Agreement Choices on a 7-Level Likert Scale	The software... (good)																
<p><b>7b)</b> does not encourage to test new functions.</p>	<table border="1"> <caption>Data for 7b)</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>6</td></tr> <tr><td>3</td><td>3</td></tr> <tr><td>4</td><td>4</td></tr> <tr><td>5</td><td>17</td></tr> <tr><td>6</td><td>20</td></tr> <tr><td>7</td><td>8</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	1	2	6	3	3	4	4	5	17	6	20	7	8	<p>encourages to test new functions</p>
Agreement Choice	Frequency																	
1	1																	
2	6																	
3	3																	
4	4																	
5	17																	
6	20																	
7	8																	
<p><b>7c)</b> requires users to keep many details in mind.</p>	<table border="1"> <caption>Data for 7c)</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>4</td></tr> <tr><td>3</td><td>5</td></tr> <tr><td>4</td><td>20</td></tr> <tr><td>5</td><td>22</td></tr> <tr><td>6</td><td>8</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	0	2	4	3	5	4	20	5	22	6	8	7	0	<p>does not require users to keep many details in mind.</p>
Agreement Choice	Frequency																	
1	0																	
2	4																	
3	5																	
4	20																	
5	22																	
6	8																	
7	0																	
<p><b>7d)</b> is designed in a way that learned actions are not easy to keep in mind.</p>	<table border="1"> <caption>Data for 7d)</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>0</td></tr> <tr><td>2</td><td>1</td></tr> <tr><td>3</td><td>3</td></tr> <tr><td>4</td><td>23</td></tr> <tr><td>5</td><td>22</td></tr> <tr><td>6</td><td>10</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	0	2	1	3	3	4	23	5	22	6	10	7	0	<p>is designed in a way that learned actions are easy to keep in mind.</p>
Agreement Choice	Frequency																	
1	0																	
2	1																	
3	3																	
4	23																	
5	22																	
6	10																	
7	0																	
<p><b>7e)</b> cannot be learned with help by others or a manual.</p>	<table border="1"> <caption>Data for 7e)</caption> <thead> <tr> <th>Agreement Choice</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>2</td></tr> <tr><td>2</td><td>10</td></tr> <tr><td>3</td><td>12</td></tr> <tr><td>4</td><td>12</td></tr> <tr><td>5</td><td>17</td></tr> <tr><td>6</td><td>6</td></tr> <tr><td>7</td><td>0</td></tr> </tbody> </table>	Agreement Choice	Frequency	1	2	2	10	3	12	4	12	5	17	6	6	7	0	<p>is easy to learn without help by others or a manual.</p>
Agreement Choice	Frequency																	
1	2																	
2	10																	
3	12																	
4	12																	
5	17																	
6	6																	
7	0																	

## B.5 Miscellaneous

### B.5.1 Distribution of the Random Number Generator

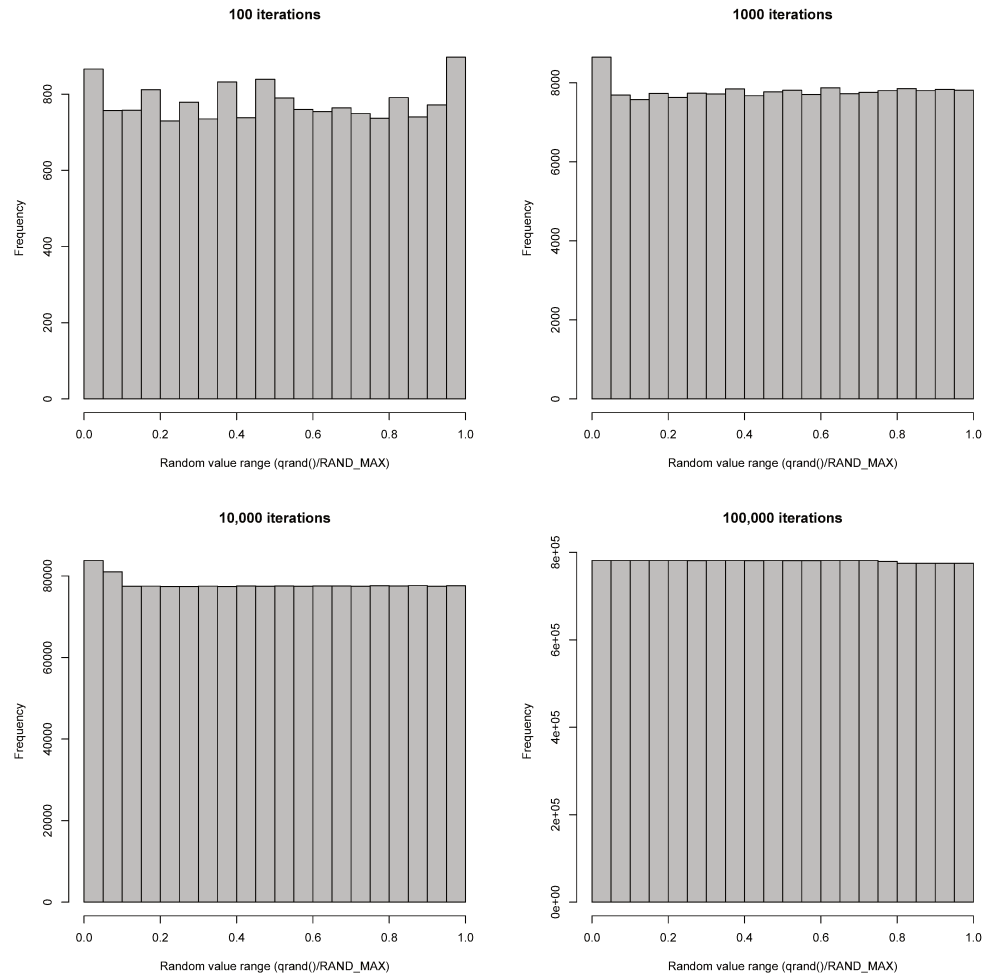


Figure B.96: Histograms of random values generated by *grand*, normalized to [0;1]

# C Questionnaires and Interviews

## C.1 Demographics and Usage Questionnaire

- Year of Birth
- Gender
- Job Type
  1. Pupil
  2. In job training
  3. Student
  4. Fully employed
  5. Part-time employed
  6. Not employed
  7. Retired
  8. Other

- Field of Study / Job Training

- Course Level

**Q0:** Have you visited one or more of the following lectures?

IR: Information Retrieval

MR: Multimedia Retrieval

**Q1:** Are you familiar with the principles of content-based information retrieval?

0. No
1. A little
2. I am an informed outsider
3. Very much
4. I am an expert

**Q2:** Are you colorblind?

0. I do not know.
1. No
2. Yes

**Q3:** How many minutes do you use the Internet per day?

0. Not at all
2. 1 - 30 minutes

## C Questionnaires and Interviews

3. 31 - 60 minutes
4. 61 - 90 minutes
5. 91 - 120 minutes
6. More than 120 minutes
7. More than 240 minutes

**Q4:** Do you know Web 2.0 services such as Flickr or Fotocommunity.de for sharing holiday, family or other photographs with friends?

0. Never heard of it
2. Know it by name
3. I have visited such websites
4. I do have an account

**Q5:** How often do you use such Web 2.0 services to share photographs with friends?

0. Never
1. Less than once a month
2. More than once a month
3. Weekly
4. Daily

**Q6:** Which of the following services do you use to upload and administrate holiday, family or other photographs? (Choose one or more.)

- None
- Facebook
- Flickr
- Fotocommunity.de
- Picasa
- Other

**Q7:** How often do you take photographs?

0. Seldom
1. Only at special events
2. Often
3. Virtually always

## C.2 Demographics of the Contributors and Assessors



Table C.1: Demographics of the Contributors; for the answer codes see Appendix C.1

AssessorID	Year of Birth	Gender	Job Type	Field of Study / Job Training	Course Level	MR	IR	Q1	Q2	Q3	Q4	Q5	Q7	Q6	Q6
actor0						0	0		0						
actor1	1983	f	4	Business Mathematics	M.Sc.	0	0	1	1	7	3	0	1	none	
actor2	1985	f	4	Hotel Business		0	0	0	1	5	4	1	2	Facebook	photocommunity
actor3	1985	f	4	Computer Science	M.Sc.	0	0	2	1	4	2	0	0	none	
actor4	1984	f	4	Business Mathematics	M.Sc.	0	0	3	1	7	3	0	1	none	
actor5	1985	m	4	Business Information Systems	M.Sc.	1	0	3	1	7	4	1	1	Facebook	other
actor6	1984	m	4	Information and Media Technology	M.Sc.	0	0	3	1	4	4	1	2	Picasa	
actor7	1985	m	4	Business Information Systems	M.Sc.	0	0	2	0	7	3	0	1	none	
actor8	1979	m	4	Computer Science	M.Sc.	1	1	4	1	7	4	2	1	Facebook	other
actor9	1944	m	8	Engineering		0	0	0	1	3	2	0	2	none	
actor10	1979	f	4	Media and Computing Sciences	M.Sc.	0	0	0	1	7	4	1	1	Facebook	photocommunity
actor11	1977	m	5	Media and Computing Sciences	M.Sc.	0	0	0	1	2	2	0	2	none	
actor12	1979	f	4	History of Art	M.A.	0	0	0	1	6	4	2	2	Facebook	
actor13	1982	f	4	Business Information Systems	M.Sc.	0	0	3	1	7	3	1	2	other	
actor14	1955	f	4	Travel Agency		0	0	0	1	4	2	0	0	none	
actor15	1959	m	4	Cybernetics	M.Sc.	0	0	0	1	7	3	1	1	Picasa	
actor16	1966	m	4	Mathematics	M.Sc.	0	0	2	1	4	0	0	1	none	
actor17	1983	f	5	Social Work	M.Sc.	0	0	0	1	2	2	0	1	none	
actor18	1983	m	4			0	0	0	1	7	4	2	2	other	
<b>Min</b>	1944	<b>Min</b>	4				<b>Min</b>	0	0	2	0	0	0		
<b>Max</b>	1985	<b>Max</b>	8				<b>Max</b>	4	1	7	4	2	2		
<b>Median</b>	1982,50	<b>Median</b>	4				<b>Median</b>	0,5	1	6,5	3	0,5	1		
<b>Mean</b>	1976,50	<b>Mean</b>	4,33				<b>Mean</b>	1,28	0,94	5,39	2,94	0,67	1,28		

Table C.2: Demographics of the Assessors; for the answer codes see Appendix C.1

AssessorID	Year of Birth	Gender	Job Type	Field of Study / Job Training	Course Level	MR	IR	Q1	Q2	Q3	Q4	Q5	Q7	Q6	Q6	Q6
assessor11	1988	m	3	Business Administration	M.Sc.	0	0	0	0	3	3	1	2	Facebook	flickr	
assessor18	1985	m	3	Business Administration	B.Sc.	0	0	0	1	6	1	1	2	Facebook		
assessor24	1990	f	3	Business Administration	B.Sc.	0	0	0	1	5	3	2	1	Facebook	Picasa	
assessor27	1987	f	3	Business Administration	B.Sc.	0	0	0	1	6	1	1	3	Facebook		
assessor48	1983	f	3	Business Administration	PhD	0	0	0	0	6	2	0	1			
assessor13	1988	m	3	Business Administration	M.Sc.	0	0	1	1	5	2	0	2			
assessor28	1988	f	3	Business Administration	M.Sc.	0	0	1	1	3	0	1	2	Facebook	Picasa	
assessor26	1987	f	3	Business Administration	M.Sc.	0	0	2	1	3	0	0	1	Picasa		
assessor25	1984	f	3	Business Administration	B.Sc.	0	0	3	1	3	0	0	2			
assessor10	1988	m	3	Business Administration & Engineering	M.Sc.	0	0	0	0	4	3	1	2	fotocommunity		
assessor12	1988	m	3	Business Administration & Engineering	B.Sc.	0	0	0	0	6	0	1	2	Facebook		
assessor16	1986	m	3	Business Administration & Engineering	M.Sc.	0	0	0	1	5	1	0	2	Facebook		
assessor17	1991	m	3	Business Administration & Engineering	B.Sc.	0	0	0	1	5	2	0	1			
assessor30	1988	m	3	Business Administration & Engineering	M.Sc.	0	0	0	1	5	3	0	3			
assessor21	1986	m	3	Business Administration & Engineering	M.Sc.	0	1	1	1	3	2	2	1			
assessor31	1989	m	3	Business Administration & Engineering	M.Sc.	0	0	1	1	5	3	1	2	Facebook	flickr	other
assessor33	1986	m	3	Business Administration & Engineering	M.Sc.	0	1	1	1	6	0	0	0			
assessor22	1988	f	3	Business Administration & Engineering	M.Sc.	0	0	1	1	5	1	1	1	Facebook		
assessor23	1987	f	3	Business Administration & Engineering	M.Sc.	0	0	2	1	2	1	0	1			
assessor20	1987	m	3	Computer Science	M.Sc.	0	0	0	1	6	2	0	0			
assessor51	1990	m	3	Computer Science	B.Sc.	0	0	0	1	5	2	0	0			
assessor37	1987	m	3	Computer Science	M.Sc.	1	1	2	1	6	0	0	2			
assessor36	1986	m	3	Computer Science	M.Sc.	1	1	3	1	3	2	0	1			
assessor41	1988	f	3	Computer Science	M.Sc.	1	1	3	1	6	1	0	1			
assessor42	1985	f	4	Computer Science		1	0	3	1	5	1	0	1			
assessor2	1979	m	4	Computer Science	PhD	1	1	4	1	6	3	1				
assessor44	1981	m	4	Computer Science		0	1	4	0	6	2	0	1			
assessor50	1985	m	4	eBusiness		0	0	0	1	6	2	1	1			
assessor32	1987	m	3	eBusiness	M.Sc.	1	1	1	1	4	2	0	1			
assessor39	1981	m	3	eBusiness	M.Sc.	0	0	1	0	6	1	0	2			
assessor29	1988	f	3	eBusiness	M.Sc.	0	0	1	1	5	1	0	0			
assessor38	1991	f	3	eBusiness	B.Sc.	0	0	1	1	3	1	0	1			
assessor46	1982	f	4	eBusiness	PhD	0	0	2	1	6	2	1	1	other		
assessor53	1988	m	3	eBusiness	M.Sc.	1	1	3	0	6	2	2	1	Facebook		
assessor15	1989	m	3	Information & Media Technology	M.Sc.	0	0	1	1	6	3	1	0	Facebook	flickr	
assessor19	1985	m	3	Information & Media Technology	M.Sc.	0	0	1	1	4	2	1	0	flickr		
assessor35	1986	m	3	Information & Media Technology	M.Sc.	0	1	2	1	5	1	0	1			
assessor45	1982	m	4	Information & Media Technology		1	1	3	1	5	1	0	1			
assessor43	1984	m	4	Information & Media Technology		1	0	4	1	3	3	1	1	Picasa		
assessor14	1987	m	3	Urban & Regional Planning	M.Sc.	0	0	1	1	5	3	1	1			
assessor49	1985	m	4			0	0	3	1	6	3	1	1	Facebook		
assessor47	1984	f	4			0	0	3	1	6	2	1	1	other		
<b>Min</b>	1979	<b>Min</b>	3				<b>Min</b>	0	0	2	0	0	0			
<b>Max</b>	1991	<b>Max</b>	4				<b>Max</b>	4	1	6	3	2	3			
<b>Median</b>	1987	<b>Median</b>	3				<b>Median</b>	1	1	5	2	0	1			
<b>Mean</b>	1986.29	<b>Mean</b>	3.21				<b>Mean</b>	1.40	0.83	4.88	1.67	0.55	1.22			
	<b>Male</b>	28			<b>Non-visited class</b>	33	31									
	<b>Female</b>	14			<b>Visited class</b>	9	11									

## C.3 Materials Used in the Usability Test

### C.3.1 Demographics and Usage Questionnaire

The following list contains the translated questions of the used questionnaire including their question and answer keys. The participants of the study had access to a German version of the questionnaire. Optional questions are indicated. The full results can be found as a supplement (see Appendix E).

1. **QDEM\_YEAR:** Year of Birth
2. **QDEM\_GENDER:** Gender
3. **QDEM\_JOBTYPE:** Job Type
  - A1 Pupil
  - A2 Student
  - A3 In job training
  - A4 Fully employed
  - A5 Part-time employed
  - A7 Not employed
  - A8 Retired
4. **QDEM\_LECTURE:** Did you visit one or both of the following lectures?  
 [This question is optional.]  
 SQ001 Information Retrieval  
 SQ002 Multimedia Retrieval
5. **QDEM\_CBIR:** Are you familiar with the principles of content-based information retrieval?
  - A0 No
  - A1 A little
  - A2 I am an informed outsider
  - A3 Very much
  - A4 I am an expert
6. **QDEM\_HOURS:** How many hours do you use the Internet per day?
  - A1 Less than 1 hour
  - A2 1-3 hours
  - A3 3-5 hours
  - A4 6-8 hours
  - A5 9-11 hours
  - A6 12-14 hours
  - A7 More than 14 hours
7. **QDEM\_MACOS:** Did you work with Mac OS X before?

## C Questionnaires and Interviews

8. **QDEM\_FAMILIARITY:** How often did you use the tested software before you did participate in this study?
9. **QDEM\_CBIRUSAGE:** Did you use image searching tools before? Please choose one or more.
  - SQ001 Google image search
  - SQ002 Microsoft image search
  - SQ003 Like.com
  - SQ004 Pixolution
  - SQ005 LIRE / Caliph & Emir
  - SQ006 Fire
  - SQ007 retrievr
  - SQ008 I have not used image searching tools.
10. **USE\_PREFERENCES:** How do you assess the utility of the preference elicitation in general? Please use the German grading scale (1= very good, 2= good, 3= satisfactory, 4= sufficient, 5= deficient). [A sample image was given. This question is optional.]
11. **USE\_FACETS:** How do you assess the utility of the faceted search in general? Please use the German grading scale (1= very good, 2= good, 3= satisfactory, 4= sufficient, 5= deficient). [A sample image was given. This question is optional.]
12. **USE\_MANUALWEIGHTS:** How do you assess the utility of the manual weight setup in general? Please use the German grading scale (1= very good, 2= good, 3= satisfactory, 4= sufficient, 5= deficient). [A sample image was given. This question is optional.]
13. **USE\_INSPECTOR:** How do you assess the utility of the query weight visualization in general? Please use the German grading scale (1= very good, 2= good, 3= satisfactory, 4= sufficient, 5= deficient). [A sample image was given. This question is optional.]
14. **USE\_MATRIX:** How do you assess the utility of the matrix result visualization in general? Please use the German grading scale (1= very good, 2= good, 3= satisfactory, 4= sufficient, 5= deficient). [A sample image was given. This question is optional.]
15. **USE\_CLUSTER:** How do you assess the utility of the cluster result visualization in general? Please use the German grading scale (1= very good, 2= good, 3= satisfactory, 4= sufficient, 5= deficient). [A sample image was given. This question is optional.]
16. **USE\_SOM:** How do you assess the utility of the SOM result visualization in general? Please use the German grading scale (1= very good, 2= good, 3= satisfactory, 4= sufficient, 5= deficient). [A sample image was given. This question is optional.]
17. **RANKING\_VARIANTS:** If you would have to rank the tested GUI variants according to your personal preferences, which order would you choose? [Sample images of the GUI variants were given.]

18. **GENERAL\_COMMENTS:** Please tell us your criticism and comments to enable us to revise the software in the future.

[This question is optional.]

### C.3.2 Usability Test Instructions

The instructions (in German) shown in Figure C.1 were visible throughout the full duration of each task and had to be read before a participant could start to solve a task.

In a separate dialog, the participants were informed not to judge the loading time of the software due to its prototypical state.

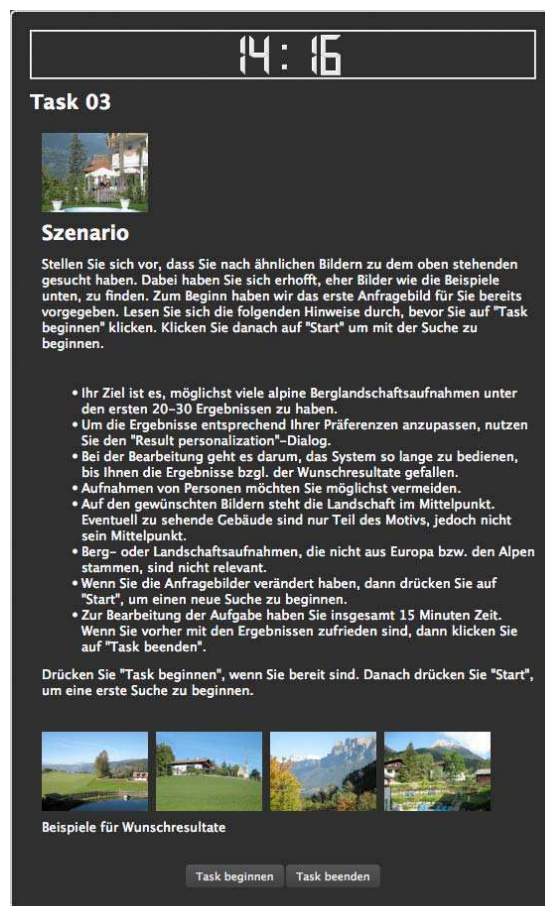


Figure C.1: Usability test instructions

**Translation of the Instructions** Imagine you have been searching similar images to the one shown above. You did expect to retrieve more images that resemble the sample images given below. To start with, the first query image has been pre-defined. Read the

## C Questionnaires and Interviews

following hints carefully before you click on “Task beginnen” [begin task]. Afterwards click “Start” to begin with the search.

- Your objective is to get as many as possible photographs of alpine mountain landscapes amongst the first 20-30 results.
- In order to adjust the results according to your preferences, use the “Result personalization” dialog.
- The objective of this task is to operate the system as long as you are not satisfied with the quality of the results regarding the given sample images.
- Photographs depicting persons should be avoided.
- The desired images focus on the landscape. Potentially visible buildings will only be part of the motif but are not its central point.
- Mountain or landscape photographs that are not taken in Europe or the Alps are not relevant.
- When you have altered the query images, please press “Start” to start a new search.
- To solve the task you are given 15 minutes. If you are satisfied with the results before, click “Task beenden” [end task].

Press “Task beginnen” whenever you are ready. Then press “Start” to start your first search.

## D Sample User Interaction with the Pythia MIR System

This following text and illustrations are taken from [Zellhöfer 2012a, Sec. 3.4]:

“To show the power of the proposed approach, we will discuss it for a simplified example: the search for documents about the Hamburg town hall. Although more complex scenarios are possible, we will use a simplified retrieval task as it illustrates the core concepts of our approach in a comprehensible way. For this example, we will rely on 14 visual modalities, one temporal (date of creation), one spatial (GPS), and one “textual” (the camera model). We regard these as cognitively different representations.

A combination of multi-lingual textual and visual modalities has been presented at ImageCLEF 2011 [Zellhöfer & Böttcher 2011] and shows the utility of our approach in a more sophisticated scenario. As we want to show the interaction with the discussed approach and its personalization mechanisms, we will skip a discussion of typical retrieval metrics or effectiveness in terms of usability, which are covered in prior works [Zellhöfer & Böttcher 2011; Zellhöfer & Schmitt 2011a].

To start with, the user provides a QBE document to the system without any additional keywords. Because a structured query and keywords are missing, the system assumes a weighted conjunction of all representations present in the QBE document. Fig. D.1 illustrates the first result documents with respect to the QBE document, which is depicted in the upper left corner throughout this example. Interested if there are more documents about the Hamburg town hall, the user decides to filter the results by choosing “Same as current” as a location facet from the faceted navigation. Internally, this increases the impact of the GPS representation in the weighted CQQL conjunction and lowers all others effectively retrieving only documents from the same location. Fig. D.8 shows the result of the operation.

Feeling not satisfied with the result, the user decides to specify a preference as depicted in Fig. D.3<sup>173</sup>. The preference input (a document showing the Hamburg town hall is more relevant than some night shot) starts the learning algorithm, which finds new weights for the conjunction modeling the current CO. Fig. D.4 shows the characteristics of the weighting scheme. For the sake of brevity, we will skip a discussion of the actual weights. Instead, we want to visualize the relative weighting trend after RF iteration #1. The same visualization is made available to the user for inspection, modification, and to communicate how the new result (Fig. D.5) has been determined.

Although the new results fits the user’s IN much better, she decides to input more precise preferences (Fig. D.6). Happy with the frontal pictures of the town hall,

<sup>173</sup>We agree that the user would be more likely to do an undo operation to simply go back in the search history but neglect this fact for the purpose of illustration.

## D Sample User Interaction with the Pythia MIR System

she places them at the second innermost ring because she cannot decide which one is better (or because it does not matter)<sup>174</sup>. Other motifs of the town hall are placed at ring 3 because they are still fairly relevant. Because she is familiar with Hamburg, she places some pictures that have been taken close to the town hall at ring 4. By carrying out this operation, she hopes to find some documents from the town hall's surroundings as well because her initial IN, the Hamburg town hall, has been slightly modified to the town hall and surroundings after exploring the collection. Finally, she saves the query incl. weights for later usage<sup>175</sup>."

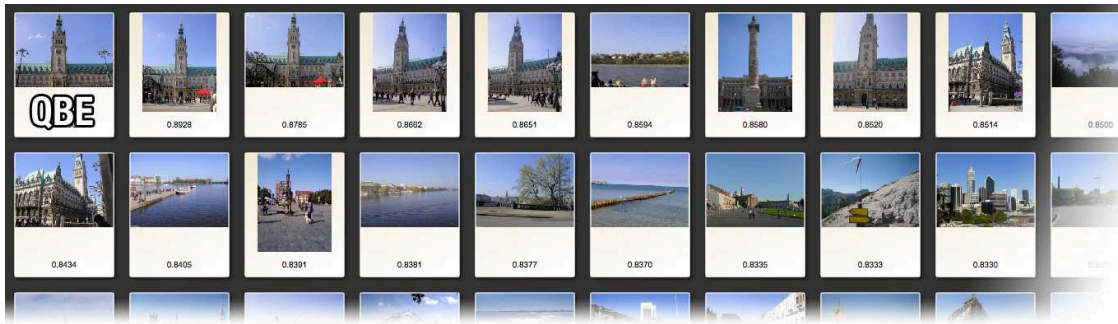


Figure D.1: Initial result set



Figure D.2: Result set after location facet "Same as current" has been chosen

<sup>174</sup>Note that this would be impossible with Liu et al.'s relevance feedback system [Liu et al. 2009].

<sup>175</sup>Note that this would be impossible for browsing approaches unless one would save the full browsing path and lock the data set in order to reconstruct the path. By using a query, we can retrieve the same results as long as the data set stays the same even if some intermediary browsing documents have been removed.



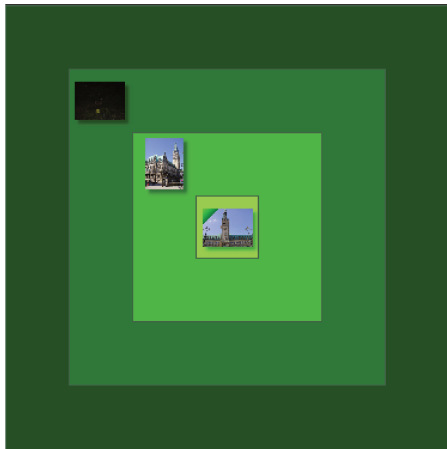


Figure D.3: Preferences at iteration #1

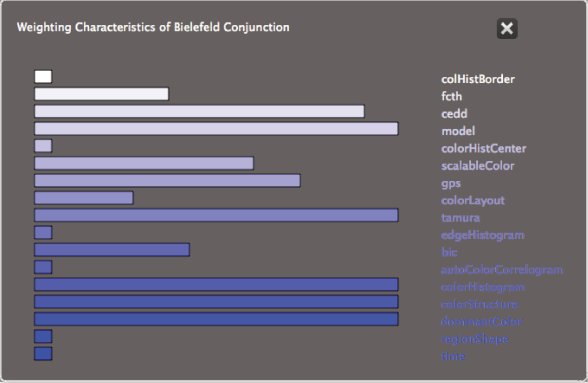


Figure D.4: Weighting characteristics of the query after iteration #1



Figure D.5: Result set after iteration #1 (see Figure D.3)

## D Sample User Interaction with the Pythia MIR System

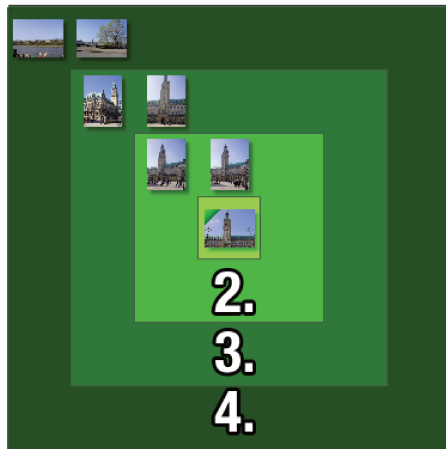


Figure D.6: Preferences at iteration #2

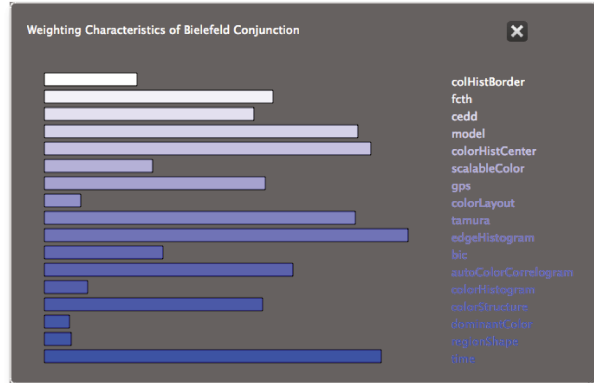


Figure D.7: Weighting characteristics of the query after iteration #2

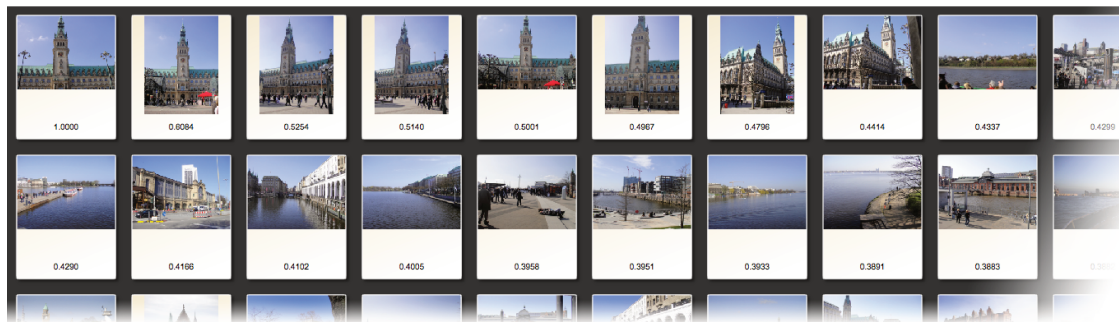


Figure D.8: Result set after iteration #2 (see Figure D.6)

## E Contents of the Enclosed DVD

Figure E.1 illustrates the file hierarchy of the enclosed DVD. To facilitate browsing, the file *content.html* contains a commented overview of the DVD's content. Furthermore, the full text of this dissertation is available in *fulltext.pdf*.

## E Contents of the Enclosed DVD

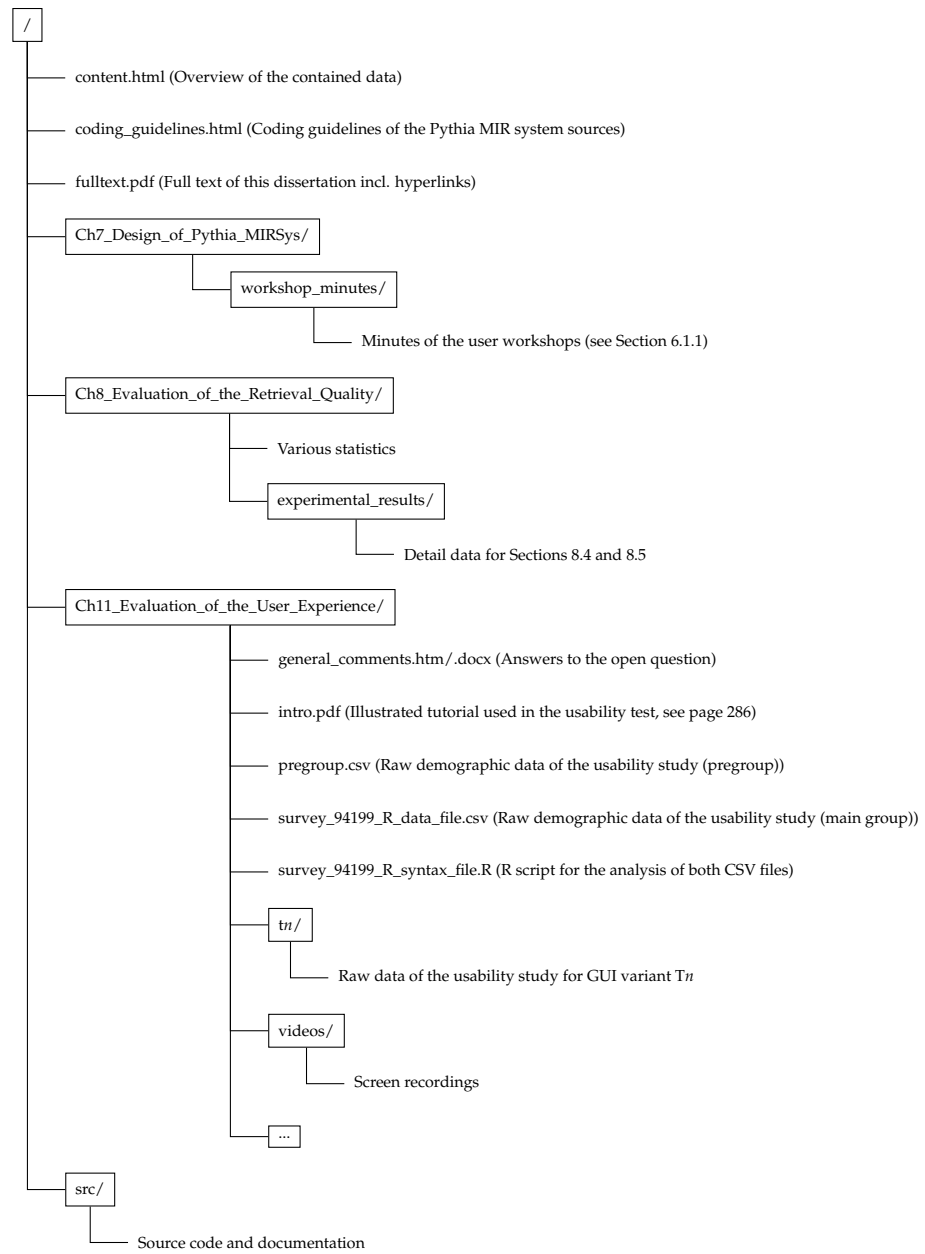


Figure E.1: Table of contents of the DVD

## F List of Publications

This section contains a list of selected publications in chronological order that are directly related to this dissertation.

Further publications by the author of this dissertation, such as monographs and essays addressing ethical problems of computer science, are available in the publications section of <http://www.tu-cottbus.de/fakultaet1/de/datenbank-informationssysteme/>.

### Journal Articles

1. ZELLHÖFER, David ; SCHMITT, Ingo: **A Preference-based Approach for Interactive Weight Learning: Learning Weights within a Logic-Based Query Language**. In: *Distributed and Parallel Databases* 27 (2010), Nr. 1, 31-51.
2. ZELLHÖFER, David ; SCHMITT, Ingo: **Interaktives Information-Retrieval auf der Basis der Polyrepräsentation: Diskussion und Experimentelle Auswertung anhand eines CBIR-Szenarios** (German). In: *Datenbank-Spektrum* 11 (2011), Nr. 3, S. 183–194

### Conference Proceedings

3. SCHMITT, Ingo ; ZELLHÖFER, David ; NÜRNBERGER, Andreas: **Towards Quantum Logic Based Multimedia Retrieval**. In: IEEE (Ed.): *Proceedings of the Fuzzy Information Processing Society (NAFIPS)*, IEEE, 2008, 1-6
4. SCHMITT, Ingo ; ZELLHÖFER, David: **Lernen nutzerspezifischer Gewichte innerhalb einer logikbasierten Anfragesprache** (German). In: FREYTAG, Christoph J. (Ed.) ; RUF, Thomas (Ed.) ; LEHNER, Wolfgang (Ed.) ; VOSSEN, Gottfried (Ed.): *Datenbanksysteme in Business, Technologie und Web (BTW 2009)*, 13. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme (DBIS), Proceedings, 2.-6. März 2009, Münster, Germany Bd. 144. GI, 2009.
5. ZELLHÖFER, David: **Inductive User Preference Manipulation for Multimedia Retrieval**. In: BÖSZÖRMENYI, Laszlo (Ed.) ; BURDESCU, Dumitru (Ed.) ; DAVIES, Philip (Ed.) ; NEWELL, David (Ed.): *Proc. of the Second International Conference on Advances in Multimedia (MMEDIA)*, IEEE, 2010.
6. ZELLHÖFER, David ; FROMMHOLZ, Ingo ; SCHMITT, Ingo ; LALMAS, Mounia ; RIJSBERGEN, Cornelis J.: **Towards Quantum-Based DB+IR Processing Based on the Principle of Polyrepresentation**. In: CLOUGH, P. (Ed.) ; FOLEY, C. (Ed.) ; GURRIN, C. (Ed.) ; JONES, G. (Ed.) ; KRAAIJ, W. (Ed.) ; LEE, H. (Ed.) ; MURDOCH,

## F List of Publications

- V. (Ed.): *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings* Bd. 6611. Springer, 2011.
7. SCHMITT, Ingo ; ZELLHÖFER, David: **Condition Learning from User Preferences**. In: *6th IEEE International Conference on Research Challenges in Information Science*. Valencia, Spain, 2012
  8. ZELLHÖFER, David: **An Extensible Personal Photograph Collection for Graded Relevance Assessments and User Simulation**. In: IP, H. H. S. (Ed.) ; RUI, Yong (Ed.): *ICMR '12: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ACM, 2012 (ICMR '12).
  9. ZELLHÖFER, David ; BERTRAM, Maria ; BÖTTCHER, Thomas ; SCHMIDT, Christoph ; TILLMANN, Claudius ; SCHMITT, Ingo: **PythiaSearch – A Multiple Search Strategy-supportive Multimedia Retrieval System**. In: *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ACM, 2012 (ICMR '12).
  10. ZELLHÖFER, David: **A Permeable Expert Search Strategy Approach to Multimodal Retrieval**. In: KAMPS, Jaap (Ed.) ; KRAAIJ, Wessel (Ed.) ; FUHR, Norbert (Ed.): *Proceedings of the 4th Information Interaction in Context Symposium*, ACM, 2012 (IIIX '12).
  11. CAPUTO, Barbara ; MÜLLER, Henning ; THOMEE, Bart ; VILLEGAS, Mauricio ; PAREDES, Roberto ; ZELLHÖFER, David ; GOEAU, Herve ; JOLY, Alexis ; BONNET, Pierre ; MARTINEZ GOMEZ, Jesus ; VAREA, I. G. ; CAZORLA, Miguel: **ImageCLEF 2013: The Vision, the Data and the Open Challenges**. In: FORNER, Pamela (Ed.) ; MÜLLER, Henning (Ed.) ; PAREDES, Roberto (Ed.) ; ROSSO, Paolo (Ed.) ; STEIN, Benno (Ed.): *Information Access Evaluation. Multilinguality, Multimodality, and Visualization* Bd. 8138. Springer Berlin Heidelberg, 2013.
  12. ZELLHÖFER, David ; BÖTTCHER, Thomas ; BERTRAM, Maria ; SCHMIDT, Christoph ; TILLMANN, Claudius ; UHLIG, Markus ; ZIERENBERG, Marcel ; SCHMITT, Ingo: **PythiaSearch - Interaktives, Multimodales Multimedia-Retrieval** (German). In: MARKL, Volker (Ed.) ; SAAKE, Gunter (Ed.) ; SATTLER, Kai-Uwe (Ed.) ; HACKENBROICH, Gregor (Ed.) ; MITSCHANG, Bernhard (Ed.) ; HÄRDER, Theo (Ed.) ; KÖPPEN, Veit (Ed.): *Datenbanksysteme für Business, Technologie und Web (BTW), 15. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), 11.-15.3.2013 in Magdeburg, Germany. Proceedings* Bd. 214. GI, 2013.

## Workshop Proceedings

13. ZELLHÖFER, David ; LEHRACK, Sebastian: **Nutzerzentriertes maschinenbasiertes Lernen von Gewichten beim Multimedia-Retrieval** (German). In: HÖPFNER, Hagen (Ed.) ; KLAN, Friederike (Ed.): *Proceedings of the 20. GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), Apolda, Thüringen, Germany, May 13-16, 2008* Bd. 01/2008. School of Information Technology, International University in Germany, 2008, S. 51–55
14. ZELLHÖFER, David ; SCHMITT, Ingo: **A Poset Based Approach for Condition Weighting**. In: *6th International Workshop on Adaptive Multimedia Retrieval*. 2008
15. ZELLHÖFER, David: **Eliciting Inductive User Preferences for Multimedia Information Retrieval**. In: BALKE, Wolf-Tilo (Ed.) ; LOFI, Christoph (Ed.): *Proceedings of the 22nd Workshop "Grundlagen von Datenbanken 2010"* Bd. 581, 2010
16. ZELLHÖFER, David ; SCHMITT, Ingo: **Ein Polyrepräsentatives Anfrageverfahren für das Multimedia Retrieval** (German). In: ATZMÜLLER, M. (Ed.) ; BENZ, D. (Ed.) ; HOTH, A. (Ed.) ; STUMME, G. (Ed.): *Proceedings of LWA2010 - Workshop-Woche: Lernen, Wissen & Adaptivität*, 2010
17. ZELLHÖFER, David ; BÖTTCHER, Thomas: **BTU DBIS' Multimodal Wikipedia Retrieval Runs at ImageCLEF 2011**. In: PETRAS, Vivien (Ed.) ; FORNER, Pamela; CLOUGH, Paul D. (Ed.): *CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands*. 2011.
18. ZELLHÖFER, David ; SCHMITT, Ingo: **A User Interaction Model based on the Principle of Polyrepresentation**. In: NICA, Anisoara (Ed.) ; SUCHANEK, F. M. (Ed.): *Proceedings of the 4th workshop on Ph.D. students in information and knowledge management, ACM, 2011 (PIKM '11)*
19. ZELLHÖFER, David ; SCHMITT, Ingo: **Approaching Multimedia Retrieval from a Polyrepresentative Perspective**. In: DETYNIECKI, Marcin (Ed.) ; KNEES, Peter (Ed.) ; NÜRNBERGER, Andreas (Ed.) ; SCHEDL, Markus (Ed.) ; STOBER, Sebastian (Ed.): *Adaptive Multimedia Retrieval. Context, Exploration, and Fusion - 8th International Workshop, AMR 2010, Linz, Austria, August 17-18, 2010, Revised Selected Papers* Bd. 6817. Springer, 2011.
20. BÖTTCHER, Thomas ; SCHMIDT, Christoph ; ZELLHÖFER, David ; SCHMITT, Ingo: **BTU DBIS' Plant Identification Runs at ImageCLEF 2012**. In: FORNER, Pamela (Ed.) ; KARLGREN, Jussi (Ed.) ; WOMSER-HACKER, Christa (Ed.): *CLEF 2012 Evaluation Labs and Workshop*, 2012.
21. ZELLHÖFER, David: **Overview of the Personal Photo Retrieval Pilot Task at ImageCLEF 2012**. In: FORNER, Pamela (Ed.) ; KARLGREN, Jussi (Ed.) ; WOMSER-HACKER, Christa (Ed.): *CLEF 2012 Evaluation Labs and Workshop*, 2012.

## F List of Publications

22. ZELLHÖFER, David: **Personas - The Missing Link between User Simulations and User-Centered Design?: Linking the Persona-based Design of Adaptive Multimedia Retrieval Systems with User Simulations.** In: *10th International Workshop on Adaptive Multimedia Retrieval*. 2012.
23. BÖTTCHER, Thomas ; ZELLHÖFER, David ; SCHMITT, Ingo: **BTU DBIS' Personal Photo Retrieval Runs at ImageCLEF 2013.** In: *CLEF 2013 Labs and Workshop, Notebook Papers, 23-26 September 2013, Valencia, Spain*. 2013
24. ZELLHÖFER, David: **Overview of the ImageCLEF 2013 Personal Photo Retrieval Subtask.** In: *CLEF 2013 Labs and Workshop, Notebook Papers, 23-26 September 2013, Valencia, Spain*. 2013

## Technical Reports

25. ZELLHÖFER, David ; BRANDENBURG UNIVERSITY OF TECHNOLOGY (Ed.): **An Evaluation of the Principle of Polyrepresentation in Multimodal and Content-based Image Retrieval.** Cottbus, 2012 (Computer Science Reports 8)
26. ZELLHÖFER, David ; BRANDENBURG UNIVERSITY OF TECHNOLOGY (Ed.): **On the Usability of PythiaSearch.** Cottbus, 2012 (Computer Science Reports 9)



# Bibliography

- [Abiteboul et al. 1996] ABITEBOUL, Serge ; HULL, Richard ; VIANU, Victor: *Foundations of Databases*. Reprinted with corr. Reading, Mass. : Addison-Wesley, 1996
- [Agrawal & Wimmers 2000] AGRAWAL, Rakesh ; WIMMERS, L. Edward: A Framework for Expressing and Combining Preferences. In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, ACM, 2000 (SIGMOD '00), 297–306
- [Aly & Demeester 2011] ALY, Robin ; DEMEESTER, Thomas: Towards a Better Understanding of the Relationship between Probabilistic Models in IR. In: *Proceedings of the Third International Conference on Advances in Information Retrieval Theory*, Springer-Verlag, 2011 (ICTIR'11), 164–175
- [Amarnath & Jain 2011] AMARNATH, Gupta ; JAIN, Ramesh: *Managing Event Information: Modeling, Retrieval, and Applications*. Morgan & Claypool Publishers, 2011 (Synthesis Lectures on Data Management)
- [Andon 1966] ANDON, F. I.: Algorithm for Simplifying DNF of Boolean Functions. In: *Cybernetics* 2 (1966), Nr. 6, 9-11
- [Arampatzis et al. 2011] ARAMPATZIS, Avi ; ZAGORIS, Konstantinos ; CHATZICHRISTOFIS, Savvas A.: Fusion vs. Two-Stage for Multimodal Retrieval. In: *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, Springer-Verlag, 2011 (ECIR'11), 759–762
- [Ash 2008] ASH, Robert B.: *Basic Probability Theory*. Dover ed. Mineola, NY : Dover Publ., 2008 (Dover Books on Mathematics)
- [Assfalg et al. 2000a] ASSFALG, J. ; DEL BIMBO, A. ; PALA, P.: Image Retrieval by Positive and Negative Examples. In: *ICPR 04 (2000)*, 4267
- [Assfalg et al. 2000b] ASSFALG, J. ; DEL BIMBO, Alberto ; PALA, P.: Using Multiple Examples for Content-Based Image Retrieval. In: *IEEE International Conference on Multimedia and Expo (I)'00*, 2000, 335–338
- [Aucouturier & Pachet 2004] AUCOUTURIER, Jean-Julien ; PACHET, Francois: Improving Timbre Similarity: How High Is The Sky? In: *Journal of Negative Results in Speech and Audio Sciences* Bd. 1. 2004
- [Awang Iskandar et al. 2007] AWANG ISKANDAR, D. N. F. ; THOM, James A. ; TAHAGHOGHI, S. M. M.: Content-based Image Retrieval Using Image Regions as Query Examples.

- In: *Proceedings of the Nineteenth Conference on Australasian Databases - Volume 75*, Australian Computer Society, Inc., 2007 (ADC '08), 38–46
- [Azzopardi et al. 2011] AZZOPARDI, Leif (Ed.) ; JÄRVELIN, Kalervo (Ed.) ; KAMPS, Jaap (Ed.) ; SMUCKER, Mark D. (Ed.): *Report on the SIGIR 2010 Workshop on the Simulation of Interaction*. Bd. 44. New York, NY, USA : ACM, 2011
- [Baeza-Yates & Ribeiro-Neto 2008] BAEZA-YATES, Ricardo ; RIBEIRO-NETO, Berthier: *Modern Information Retrieval*. [Nachdr.]. Harlow : Pearson Addison-Wesley [u.a.], 2008
- [Baeza-Yates & Ribeiro-Neto 2011] BAEZA-YATES, Ricardo ; RIBEIRO-NETO, Berthier: *Modern Information Retrieval: The Concepts and Technology behind Search*. Second ed. Harlow : Pearson Addison-Wesley [u.a.], 2011
- [Balko & Schmitt 2012] BALKO, Sören ; SCHMITT, Ingo ; BRANDENBURG UNIVERSITY OF TECHNOLOGY (Ed.): *Signature Indexing and Self-Refinement in Metric Spaces*. Cottbus, 2012 (Computer Science Reports 6)
- [Baskaya et al. 2011] BASKAYA, Feza ; KESKUSTALO, Heikki ; JÄRVELIN, Kalervo: Simulating Simple and Fallible Relevance Feedback. In: *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, Springer-Verlag, 2011 (ECIR'11), 593–604
- [Baskaya et al. 2012] BASKAYA, Feza ; KESKUSTALO, Heikki ; JÄRVELIN, Kalervo: Time drives Interaction: Simulating Sessions in Diverse Searching Environments. In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2012 (SIGIR '12), 105–114
- [Bates 1989] BATES, Marcia J.: The Design of Browsing and Berrypicking Techniques for the Online Search Interface. In: *Online Review* 13 (1989), Nr. 5, 407–424
- [Bay et al. 2006] BAY, Herbert ; TUYTELAARS, Tinne ; VAN GOOL, AND L.: SURF: Speeded Up Robust Features. In: *9th European Conference on Computer Vision*, 2006
- [Becker 2008] BECKER, Gary Stanley: *The Economic Approach to Human Behavior*. Paperback ed., [Nachdr.]. Chicago : Univ. of Chicago Press, 2008
- [Beckers 2009] BECKERS, Thomas: Supporting Polyrepresentation and Information Seeking Strategies. In: *Proceedings of the 3rd Symposium on Future Directions in Information Access (FDIA)*, 2009
- [Belkin 1980] BELKIN, Nicholas: Anomalous States of Knowledge as a Basis for Information Retrieval. In: *Canadian Journal of Information Science* (1980), Nr. 5, 133–143
- [Belkin 1993] BELKIN, Nicholas: Interaction with Texts: Information Retrieval as Information-Seeking Behavior. In: *Information Retrieval*, 1993, 55–66

## Bibliography

- [Belkin 1996] BELKIN, Nicholas: Intelligent Information Retrieval: Whose Intelligence? In: *ISI '96: Proceedings of the Fifth International Symposium for Information Science*, 1996, 25–31
- [Belkin et al. 1993] BELKIN, Nicholas ; MARCHETTI, P. G. ; COOL, C.: BRAQUE: Design of an Interface to Support User Interaction in Information Retrieval. In: *Inf. Process. Manage.* 29 (1993), Nr. 3, 325–344
- [Belkin et al. 1982] BELKIN, Nicholas ; ODDY, R. N. ; BROOKS, H. M.: ASK for Information Retrieval: Part I. Background and Theory. In: *Journal of Documentation* 38 (1982), Nr. 2, 61–71
- [Ben-Bassat et al. 2006] BEN-BASSAT, Tamar ; MEYER, Joachim ; TRACTINSKY, Noam: Economic and Subjective Measures of the Perceived Value of Aesthetics and Usability. In: *ACM Trans. Comput.-Hum. Interact.* 13 (2006), Nr. 2, 210–234
- [Bentley et al. 1978] BENTLEY, J. L. ; KUNG, H. T. ; SCHKOLNICK, M. ; THOMPSON, C. D.: On the Average Number of Maxima in a Set of Vectors and Applications. In: *J. ACM* 25 (1978), Nr. 4, 536–543
- [Bilal 2000] BILAL, Dania: Children's Use of the Yahoo! Search Engine: Cognitive, Physical, and Affective Behaviors on Fact-based Search Tasks. In: *J. Am. Soc. Inf. Sci.* 51 (2000), Nr. 7, 646–665
- [Birkhoff 1993] BIRKHOFF, Garrett (Ed.): *American Mathematical Society Colloquium Publications*. Bd. 25: *Lattice Theory: Colloquium Publications*. 3. ed., 7 printing with corrections. Providence, RI : Amer. Math. Soc. and American Mathematical Society, 1993
- [Birkhoff & von Neumann 1936] BIRKHOFF, Garrett ; VON NEUMANN, John: The Logic of Quantum Mechanics. In: *Annals of Mathematics* 37 (1936), 823–843
- [Blanken et al. 2007] BLANKEN, Henk M. ; BLOK, Henk Ernst ; FENG, Ling ; DE VRIES, Arjen P.: *Multimedia Retrieval*. Berlin, Heidelberg : Springer-Verlag Berlin Heidelberg, 2007
- [Bleiholder et al. 2011] BLEIHOLDER, Jens ; HERSCHEL, Melanie ; NAUMANN, Felix: Eliminating NULLs with Subsumption and Complementation, 2011, 18–25
- [Bleiholder & Naumann 2009] BLEIHOLDER, Jens ; NAUMANN, Felix: Data Fusion. In: *ACM Comput. Surv.* 41 (2009), Nr. 1, 1:1–1:41
- [Böckelmann 2009] BÖCKELMANN, Bianca: *Entwurf einer GUI zum Lernen von Gewichten beim visuellen Multimedia-Retrieval (German); Bachelor's thesis*. Brandenburg University of Technology, Cottbus, Diplomarbeit, 2009
- [Borlund 2003] BORLUND, Pia: The IIR Evaluation Model: A Framework for Evaluation of Interactive Information Retrieval Systems. In: *Inf. Res.* 8 (2003), Nr. 3

- [Börzsönyi et al. 2001] BÖRZSÖNYI, Stephan ; KOSSMANN, Donald ; STOCKER, Konrad: The Skyline Operator. In: *Proceedings of the 17th International Conference on Data Engineering*, IEEE Computer Society, 2001, 421–430
- [Böttcher et al. 2013] BÖTTCHER, Thomas ; ZELLHÖFER, David ; SCHMITT, Ingo: BTU DBIS' Personal Photo Retrieval Runs at ImageCLEF 2013. In: *CLEF 2013 Labs and Workshop, Notebook Papers, 23-26 September 2013, Valencia, Spain*. 2013
- [Bradley et al. 1992] BRADLEY, Stephen P. ; HAX, Arnaldo C. ; MAGNANTI, Thomas L.: *Applied mathematical programming*. [19. Dr.]. Reading, Mass. : Addison-Wesley, 1992
- [Brafman & Domshlak 2009] BRAFMAN, Ronen I. ; DOMSHLAK, Carmel ; BEN-GURION UNIVERSITY (Ed.): *Preference Handling – An Introductory Tutorial*. 2009 (Technical Report TR 08-04)
- [Bruns & Meyer-Wegener 2005] BRUNS, Kai ; MEYER-WEGENER, Klaus: *Taschenbuch der Medieninformatik: Mit 39 Tabellen*. München : Fachbuchverl. Leipzig im Carl Hanser-Verl., 2005
- [Buckow 2009] BUCKOW, Christian: *Integration einer Multiparadigmen-Ähnlichkeitssuche in eine Kunstdatenbank (German); Bachelor's thesis*. Brandenburg University of Technology, Cottbus, Diplomarbeit, 2009
- [Burkard et al. 1991] BURKARD, Franz-Peter ; WIEDMANN, Franz ; KUNZMANN, Peter: *dto-Atlas Philosophie*. 15. München : Deutscher Taschenbuch Verlag, 1991
- [Burmester et al. 2008] BURMESTER, Michael ; HASSENZAHL, Marc ; KAISER, Karin ; KOLLER, Franz: Editorial: Usability & Ästhetik. In: *i-com 7* (2008), Nr. 3, 25–29
- [Callan et al. 1992] CALLAN, J. P. ; CROFT, W. Bruce ; HARDING, Stephen M.: The INQUERY Retrieval System. In: *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*. (1992), 78–83
- [Calonder et al. 2010] CALONDER, Michael ; LEPETIT, Vincent ; STRECHA, Christoph ; FUA, Pascal: BRIEF: Binary Robust Independent Elementary Features. In: *Proceedings of the 11th European Conference on Computer Vision: Part IV*, Springer-Verlag, 2010 (ECCV'10), 778–792
- [Campbell 2000] CAMPBELL, Iain: Interactive Evaluation of the Ostensive Model Using a New Test Collection of Images with Multiple Relevance Assessments. In: *Inf. Retr.* 2 (2000), 89–114
- [Caputo et al. 2013] CAPUTO, Barbara ; MÜLLER, Henning ; THOME, Bart ; VILLEGAS, Mauricio ; PAREDES, Roberto ; ZELLHÖFER, David ; GOEAU, Herve ; JOLY, Alexis ; BONNET, Pierre ; MARTINEZ GOMEZ, Jesus ; VAREA, Ismael Garcia ; CAZORLA, Miguel: ImageCLEF 2013: The Vision, the Data and the Open Challenges. In: FORNER, Pamela (Ed.) ; MÜLLER, Henning (Ed.) ; PAREDES, Roberto (Ed.) ;

## Bibliography

- ROSSO, Paolo (Ed.) ; STEIN, Benno (Ed.): *Information Access Evaluation. Multilinguality, Multimodality, and Visualization* Bd. 8138. Springer Berlin Heidelberg, 2013, 250-268
- [Carroll 2000] CARROLL, John M.: Introduction to this Special Issue on "Scenario-Based System Development. In: *Interacting with Computers* 13 (2000), Nr. 1, 41–42
- [Carterette 2011] CARTERETTE, Ben: System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information, ACM, 2011 (SIGIR '11)*, 903–912
- [Chatzichristofis & Boutalis 2008a] CHATZICHRISTOFIS, Savvas A. ; BOUTALIS, Yiannis S.: CEDD: Color and Edge Directivity Descriptor: A Compact Descriptor for Image Indexing and Retrieval. In: *Proceedings of the 6th International Conference on Computer Vision Systems, Springer-Verlag, 2008 (ICVS'08)*, 312–322
- [Chatzichristofis & Boutalis 2008b] CHATZICHRISTOFIS, Savvas A. ; BOUTALIS, Yiannis S.: FCTH: Fuzzy Color and Texture Histogram - A Low Level Feature for Accurate Image Retrieval. In: *Proceedings of the 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services, IEEE Computer Society, 2008 (WIAMIS '08)*, 191–196
- [Chaudhuri et al. 2006] CHAUDHURI, S. ; DALVI, N. ; KAUSHIK, R.: Robust Cardinality and Cost Estimation for Skyline Operator. In: *Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference on, 2006*, 64–64
- [Chaudhuri et al. 2005] CHAUDHURI, Surajit ; WEIKUM, Ramakrishnan ; RAGHU, Gerhard: Integrating DB and IR Technologies: What is the Sound of One Hand Clapping? In: *CIDR, 2005*, 1–12
- [Chomicki et al. 2003] CHOMICKI, J. ; GODFREY, P. ; GRYZ, J. ; LIANG, D.: Skyline with Presorting. In: *Data Engineering, 2003. Proceedings. 19th International Conference on, 2003*, 717–719
- [Chomicki 2002] CHOMICKI, Jan: Querying with Intrinsic Preferences. In: *EDBT '02: Proceedings of the 8th International Conference on Extending Database Technology, Springer-Verlag, 2002*, 34–51
- [Chomicki 2003] CHOMICKI, Jan: Preference Formulas in Relational Queries. In: *ACM Trans. Database Syst.* 28 (2003), Nr. 4, 427–466
- [Chomicki 2007] CHOMICKI, Jan: Database Querying under Changing Preferences. In: *Annals of Mathematics and Artificial Intelligence* 50 (2007), Nr. 1–2, 79–109
- [Cieplinski et al. 2001] CIEPLINSKI, Leszek ; JEANNIN, Sylvie ; OHM, Jens-Rainer ; KIM, Munchurl ; PICKERING, Mark ; YAMADA, Akio: *MPEG-7 Visual XM version 8.1*. Pisa, Italy, 2001. ( M6808)

- [Cleverdon 1962] CLEVERDON, Cyril W. ; ASLIB CRANFIELD RESEARCH PROJECT (Ed.): *Aslib Cranfield Research Project: Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*. Cranfield, USA, 1962
- [Codd 1970] CODD, E. F.: A Relational Model of Data for Large Shared Data Banks. In: *Commun. ACM* 13 (1970), Nr. 6, 377–387
- [Codd 1972] CODD, E. F.: Relational Completeness of Data Base Sublanguages. In: R. RUSTIN (Ed.): *Data Base Systems* Bd. 6. Englewood Cliffs, NJ : Prentice Hall, 1972, 65–98
- [Cohn 2008] COHN, Mike: *User Stories Applied: For Agile Software Development*. 12. print. Boston, Mass. : Addison-Wesley, 2008 (Addison-Wesley Signature Series)
- [Cole 2011] COLE, Michael J.: Simulation of the IIR User: Beyond the Automagic. In: AZZOPARDI, Leif (Ed.) ; JÄRVELIN, Kalervo (Ed.) ; KAMPS, Jaap (Ed.) ; SMUCKER, Mark D. (Ed.): *Report on the SIGIR 2010 Workshop on the Simulation of Interaction* Bd. 44, ACM, 2011, 1–2
- [Conitzer 2010] CONITZER, Vincent: Making Decisions Based on the Preferences of Multiple Agents. In: ACM (Ed.): *Communications of the ACM* Bd. 53. New York : ACM Press, 2010
- [Cooper 1999] COOPER, Alan: *The Inmates Are Running the Asylum*. Indianapolis, IN, USA : Macmillan Publishing Co., Inc., 1999
- [Cooper et al. 2007] COOPER, Alan ; REIMANN, Robert ; CRONIN, Dave: *About Face 3: The Essentials of Interaction Design*. Completely rev. and updated. Indianapolis, Ind. : Wiley, 2007
- [Craik 1943] CRAIK, K. J. W.: *The Nature of Explanation*. Cambridge : Cambridge University Press, 1943
- [Cribbin & Chen 2001] CRIBBIN, Timothy ; CHEN, Chaomei: Exploring Cognitive Issues in Visual Information Retrieval. In: *Proceedings of the 8th IFIP TC.13 Conference on Human-Computer Interaction (INTERACT 2001)*, 2001, 166–173
- [Croft & Harper 1988] CROFT, W. Bruce ; HARPER, D. J.: *Using Probabilistic Models of Document Retrieval Without Relevance Information: Document Retrieval Systems*. London, UK, UK : Taylor Graham Publishing, 1988, 161–171
- [Croft et al. 2009] CROFT, W. Bruce ; METZLER, Donald ; STROHMAN, Trevor: *Search Engines: Information Retrieval in Practice*. International Edition. Boston, Mass. : Pearson, 2009
- [Croft & Thompson 1987] CROFT, W. Bruce ; THOMPSON, R. H.: I3R: A New Approach to the Design of Document Retrieval Systems. In: *J. Am. Soc. Inf. Sci.* 38 (1987), Nr. 6, 389–404

## Bibliography

- [Cunningham & Masoodian 2006] CUNNINGHAM, Sally Jo ; MASOODIAN, Masood: Looking for a Picture: An Analysis of Everyday Image Information Searching. In: *JCDL '06: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM, 2006, 198–199
- [Cunningham & Masoodian 2007] CUNNINGHAM, Sally Jo ; MASOODIAN, Masood: Identifying Personal Photo Digital Library Features. In: *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM, 2007 (JCDL '07), 400–401
- [Dahm 2006] DAHM, Markus: *Grundlagen der Mensch-Computer-Interaktion*. München : Pearson Studium, 2006 (Informatik)
- [Date 1982] DATE, C. J.: A Formal Definition of the Relational Model. In: *SIGMOD Rec.* 13 (1982), Nr. 1, 18–29
- [Date 2000] DATE, C. J.: *An Introduction to Database Systems*. 7. ed., repr. with corr. Reading, Mass. : Addison-Wesley, 2000
- [Datta et al. 2008] DATTA, Ritendra ; JOSHI, Dhiraj ; LI, Jia ; WANG, James Z.: Image Retrieval: Ideas, Influences, and Trends of the New Age. In: *ACM Computing Surveys* 40 (2008), Nr. 2, 1–60
- [Del Bimbo 1999] DEL BIMBO, Alberto: *Visual Information Retrieval*. San Francisco, Calif. : Kaufmann, 1999
- [deMey 1977] DEMEY, M.: The Cognitive Viewpoint: Its Development and Its Scope. In: *CC 77: International Workshop on the Cognitive Viewpoint*. Ghent, Belgium, 1977, xvi–xxxii
- [Deselaers et al. 2005] DESELAERS, Thomas ; KEYSERS, Daniel ; NEY, Hermann: FIRE – Flexible Image Retrieval Engine: ImageCLEF 2004 Evaluation. In: *Proceedings of the 5th Conference on Cross-Language Evaluation Forum: Multilingual Information Access for Text, Speech and Images*, Springer-Verlag, 2005 (CLEF'04), 688–698
- [Deselaers et al. 2008] DESELAERS, Thomas ; KEYSERS, Daniel ; NEY, Hermann: Features for Image Retrieval: An Experimental Comparison. In: *Information Retrieval* 11 (2008), 77–107
- [Dirac 1958] DIRAC, Paul: *The Principles of Quantum Mechanics*. 4th. Oxford University Press, 1958
- [Döcke 2012] DÖCKE, Michael: *Vergleichende Untersuchung von Quantenmechanik-basierten Information-Retrieval-Modellen (German)*; Master's thesis. Brandenburg University of Technology, Cottbus, Diplomarbeit, 2012
- [Dominich 2008] DOMINICH, Sándor: *The Modern Algebra of Information Retrieval*. Berlin, Heidelberg : Springer-Verlag Berlin Heidelberg, 2008 (Springer-11645 / Dig. Serial)]

- [Dushnik & Miller 1941] DUSHNIK, Ben ; MILLER, E. W.: Partially Ordered Sets. In: *American Journal of Mathematics* 63 (1941), Nr. 3, 600-610
- [Egnor & Lord 2000] EGNOR, Daniel ; LORD, Robert: Structured Information Retrieval using XML. In: ACM (Ed.): *Proceedings of the ACM SIGIR 2000 Workshop on XML and Information Retrieval*, 2000
- [Eidenberger 2003] EIDENBERGER, Horst: How Good are the Visual MPEG-7 Features? In: *SPIE & IEEE Visual Communications and Image Processing Conference*, 2003, 476-488
- [Eitz et al. 2009] EITZ, Mathias ; HILDEBRAND, Kristian ; BOUBEKEUR, Tamy ; ALEXA, Marc: PhotoSketch: A Sketch Based Image Query and Compositing System. In: *SIGGRAPH 2009: Talks*, ACM, 2009 (SIGGRAPH '09), 60:1-60:1
- [Ellis 1989] ELLIS, D.: A Behavioural Model for Information Retrieval System Design. In: *J. Inf. Sci.* 15 (1989), Nr. 4-5, 237-248
- [Ellis & Haugan 1997] ELLIS, David ; HAUGAN, Merete: Modelling the Information Seeking Patterns of Engineers and Research Scientists in an Industrial Environment. In: *Journal of Documentation* 53 (1997), Nr. 4, 384-403
- [Elmasri & Navathe 2011] ELMASRI, Ramez ; NAVATHE, Shamkant B.: *Database Systems: Models, Languages, Design, and Application Programming*. 6. ed., global ed. Boston, Mass. : Pearson, 2011
- [Fagin & Wimmers 2000] FAGIN, R. ; WIMMERS, E. L.: A Formula for Incorporating Weights into Scoring Rules. In: *Special Issue of Theoretical Computer Science* (2000), Nr. 239, 309-338
- [Fagin & Wimmers 1997] FAGIN, Ronald ; WIMMERS, E. L.: Incorporating User Preferences in Multimedia Queries. In: AFRATI, Foto (Ed.) ; KOLAITIS, Phokion (Ed.): *Database Theory@ ICDT'97* Bd. 1186. Springer Berlin Heidelberg, 1997, 247-261
- [Falk et al. 2002] FALK, Michael ; MAROHN, Frank ; TEWES, Bernward: *Foundations of Statistical Analyses and Applications with SAS*. Basel , Boston : Birkhauser Verlag, 2002
- [Faria et al. 2010] FARIA, Fabio F. ; VELOSO, Adriano ; ALMEIDA, Humberto M. ; VALLE, Eduardo ; TORRES, Ricardo da S. ; GONÇALVES, Marcos A. ; MEIRA, Jr. Wagner: Learning to Rank for Content-based Image Retrieval. In: *MIR '10: Proceedings of the International Conference on Multimedia Information Retrieval*, ACM, 2010, 285-294
- [Fei-Fei et al. 2004] FEI-FEI, L. ; FERGUS, R. ; PERONA, P.: Learning Generative Visual Models from Few Training Examples an Incremental Bayesian Approach Tested on 101 Object Categories. In: *Proceedings of the Workshop on Generative-Model Based Vision*, 2004
- [Feng et al. 2003] FENG, D. (Ed.) ; SIU, W. C. (Ed.) ; ZHANG, H. J. (Ed.): *Multimedia Information Retrieval and Management: Technological Fundamentals and Applications*. Berlin, Heidelberg : Springer-Verlag, 2003



## Bibliography

- [Fishburn 1968] FISHBURN, Peter C.: Utility Theory. In: *Management Science* 14 (1968), Nr. 5, 335-378
- [Fishburn 1970] FISHBURN, Peter C.: *Utility theory for Decision Making*. Wiley, 1970 (Publications in Operations Research)
- [Fitts 1954] FITTS, Paul Morris: The Information Capacity of the Human Motor System in Controlling Amplitude of Movement. In: *Journal of Experimental Psychology* Bd. 47. 1954, 381-391
- [Flickner et al. 1995] FLICKNER, M. ; SAWHNEY, H. S. ; ASHLEY, J. ; HUANG, Q. ; DOM, B. ; GORKANI, M. ; HAFNER, J. ; LEE, D. ; PETKOVIC, D. ; STEELE, D. ; YANKER, P.: Query by Image and Video Content: The QBIC System. In: *IEEE Computer* 28 (1995), Nr. 9, 23-32
- [Forner et al. 2012] FORNER, Pamela (Ed.) ; KARLGREN, Jussi (Ed.) ; WOMSER-HACKER, Christa (Ed.): *CLEF 2012 Evaluation Labs and Workshop*. 2012
- [Fox et al. 1992] FOX, Edward A. ; BETRABET, S. ; KOUSHIK, M. ; LEE, W.: Extended Boolean Models. In: FRAKES, W. B. (Ed.) ; BAEZA-YATES, R. (Ed.): *Information Retrieval: Data Structures and Algorithms*. Prentice Hall, 1992, 393-418
- [Frommholz & van Rijsbergen 2009] FROMMHOLZ, I. ; VAN RIJSBERGEN, Cornelis J.: Towards a Geometrical Model for Polyrepresentation of Information Objects. In: *Proc. of the "Information Retrieval 2009" Workshop at LWA 2009*, 2009
- [Frommholz et al. 2010] FROMMHOLZ, Ingo ; LARSEN, Birger ; PIWOWARSKI, Benjamin ; LALMAS, Mounia ; INGWERSEN, Peter ; VAN RIJSBERGEN, Cornelis J.: Supporting Polyrepresentation in a Quantum-inspired Geometrical Retrieval Framework. In: *Proceedings of the 2010 Information Interaction in Context Symposium*, ACM, 2010, 115-124
- [Fuhr 1992] FUHR, Norbert: Probabilistic Models in Information Retrieval. In: *The Computer Journal* 35 (1992), 243-255
- [Fuhr 1995] FUHR, Norbert: Probabilistic Datalog — A Logic for Powerful Retrieval Methods. In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 1995 (SIGIR '95), 282-290
- [Fuhr 2000] FUHR, Norbert: Probabilistic Datalog: Implementing Logical Information Retrieval for Advanced Applications. In: *J. Am. Soc. Inf. Sci.* 51 (2000), 95-110
- [Fuhr 2001] FUHR, Norbert: Information Retrieval Methods for Multimedia Objects: 9 (Information Retrieval Methods for Multimedia Objects). In: VELTKAMP, C. R. (Ed.) ; BURKHARDT, Hans (Ed.) ; KRIEGEL, P. H. (Ed.): *State-of-the-Art in Content-Based Image and Video Retrieval*. Kluwer Academic Publishers, Netherlands, 2001, 191-212
- [Fuhr 2008] FUHR, Norbert: A Probability Ranking Principle for Interactive Information Retrieval. In: *Information Retrieval* 11 (2008), Nr. 3, 251-265

- [Fuhr 2012] FUHR, Norbert: Information Retrieval as Engineering Science: Salton Award Lecture. In: *SIGIR Forum* 46 (2012), Nr. 2, 19–28
- [Fuhr & Rölleke 1994] FUHR, Norbert ; RÖLLEKE, Thomas: A Probabilistic Relational Algebra for the Integration of Information Retrieval and Database Systems. In: *ACM Transactions on Information Systems* 15 (1994), 32–66
- [Fürnkranz & Hüllermeier 2010] FÜRNKRANZ, Johannes ; HÜLLERMEIER, Eyke: Preference Learning: An Introduction. In: FÜRNKRANZ, Johannes (Ed.) ; HÜLLERMEIER, Eyke (Ed.): *Preference Learning*. Springer-Verlag, 2010, 1–17
- [Galindo et al. 2005] GALINDO, J. ; URRUTIA, A. ; PIATTINI, M.: *Fuzzy Databases: Modeling, Design and Implementation*. Idea Group Publishing, 2005
- [Gamma et al. 1995] GAMMA, E. ; HELM, R. ; JOHNSON, R. ; VLISSIDES, J.: *Design Patterns: Elements of Reusable Object-Oriented Software: Elements of reusable object-oriented software*. 36th print. Reading, MA // Boston : Addison-Wesley, 1995 // 2008
- [Garcia-Molina et al. 2000] GARCIA-MOLINA, Hector ; ULLMAN, Jeffrey D. ; WIDOM, Jennifer: *Database system implementation*. Upper Saddle River, NJ : Prentice Hall, 2000
- [Ghias et al. 1995] GHIAS, Asif ; LOGAN, Jonathan ; CHAMBERLIN, David ; SMITH, Brian C.: Query By Humming: Musical Information Retrieval in an Audio Database. In: *Proceedings of the Third ACM International Conference on Multimedia*, ACM, 1995 (MULTIMEDIA '95), 231–236
- [Gini 1997] GINI, Corrado: Concentration and Dependency Ratios: Translation from the Italian original (1909). In: *Rivista di politica economica*. (1997), Nr. 87, 769–789
- [Gleason 1957] GLEASON, A.: Measures on the Closed Subspaces of a Hilbert Space. In: *Journal of Mathematics and Mechanics* 6 (1957), 885–893
- [Godfrey et al. 2007] GODFREY, Parke ; SHIPLEY, Ryan ; GRYZ, Jarek: Algorithms and Analyses for Maximal Vector Computation. In: *The VLDB Journal* 16 (2007), Nr. 1, 5–28
- [Gould & Lewis 1987] GOULD, J. D. ; LEWIS, C.: *Designing for Usability: Key Principles and What Designers Think: Human-computer interaction*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1987, 528–539
- [Govindarajan et al. 2001] GOVINDARAJAN, Kannan ; JAYARAMAN, Bharat ; MANTHA, Surya: Preference Queries in Deductive Databases. In: *New Generation Computing* 19 (2001), Nr. 1, 57–86
- [Griffin et al. 2007] GRIFFIN, G. ; HOLUB, A. ; PERONA, P.: *Caltech-256 Object Category Dataset*. Version: 2007. ( 7694)

## Bibliography

- [Hajek et al. 1995] HAJEK, Petr ; GODO, Lluís ; ESTEVA, Francesc: Fuzzy Logic and Probability. In: *In Uncertainty in Artificial Intelligence. Proc. of 11th Conference, 1995*, 237–244
- [Hansson & Grüne-Yanoff 2012] HANSSON, Sven Ove ; GRÜNE-YANOFF, Till: Preferences. In: ZALTA, N. E. (Ed.): *The Stanford Encyclopedia of Philosophy*. 2012
- [Harman 1992] HARMAN, Donna: Relevance Feedback Revisited. In: *SIGIR '92: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 1992, 1–10
- [Hearst 2009] HEARST, Marti A.: *Search User Interfaces*. Cambridge : Cambridge Univ. Press, 2009
- [Hearst & Pedersen 1996] HEARST, Marti A. ; PEDERSEN, Jan O.: Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 1996 (SIGIR '96), 76–84
- [Heesch 2008] HEESCH, Daniel: A Survey of Browsing Models for Content Based Image Retrieval. In: *Multimedia Tools Appl.* 40 (2008), Nr. 2, 261–284
- [Holland et al. 1996] HOLLAND, John H. ; HOLYOAK, Keith J. ; NISBETT, Richard E. ; THAGARD, Paul R.: *Induction: Processes of Inference, Learning, and Discovery*. 5th print. Cambridge, Mass. : MIT Press, 1996 (Computational Models of Cognition and Perception)
- [Hollas 2007] HOLLAS, Boris: *Grundkurs Theoretische Informatik: Mit Aufgaben und Prüfungsfragen*. 1. Aufl. München : Elsevier Spektrum Akad. Verl., 2007
- [Hu et al. 2008] HU, Y. ; LI, M. ; YU, N.: Multiple-instance Ranking: Learning to Rank Images for Image Retrieval. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 2008
- [Huang et al. 1997] HUANG, Jing ; KUMAR, S. R. ; MITRA, Mandar ; ZHU, Wei-Jing ; ZABIH, Ramin: Image Indexing Using Color Correlograms. In: *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, IEEE Computer Society, 1997 (CVPR '97), 762–
- [Huiskes & Lew 2008a] HUISKES, Mark J. ; LEW, Michael S.: Performance Evaluation of Relevance Feedback Methods. In: *CIVR '08: Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, ACM, 2008, 239–248
- [Huiskes & Lew 2008b] HUISKES, Mark J. ; LEW, Michael S.: The MIR Flickr Retrieval Evaluation. In: *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, ACM, 2008 (MIR '08), 39–43

- [Huiskes et al. 2010] HUISKES, Mark J. ; THOMEE, Bart ; LEW, Michael S.: New Trends and Ideas in Visual Concept Detection: The MIR Flickr Retrieval Evaluation Initiative. In: *Proceedings of the International Conference on Multimedia Information Retrieval*, ACM, 2010 (MIR '10), 527–536
- [Hull 1997] HULL, David A.: Using Structured Queries for Disambiguation in Cross-Language Information Retrieval. In: *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval Electronic Working Notes*, 1997, 24–26
- [Hume 2009] HUME, David: *A Treatise of Human Nature*. Oxford : Oxford Univ. Press, 2009 (Oxford Philosophical Texts)
- [Hwang & Salvendy 2010] HWANG, Wonil ; SALVENDY, Gavriel: Number of People Required for Usability Evaluation: The  $10 \pm 2$  Rule. In: *Commun. ACM* 53 (2010), Nr. 5, 130–133
- [Ide 1971] IDE, E.: New Experiments in Relevance Feedback. In: *The SMART retrieval system: experiments in automatic document processing* (1971), 337–354
- [Ingwersen 1992] INGWERSEN, Peter: *Information Retrieval Interaction*. London : Taylor Graham, 1992
- [Ingwersen 1996] INGWERSEN, Peter: Cognitive Perspectives of Information Retrieval Interaction: Elements of a Cognitive IR Theory. In: *Journal of Documentation* 52 (1996), 3–50
- [Ingwersen & Järvelin 2005] INGWERSEN, Peter ; JÄRVELIN, Kalervo: *The Turn: Integration of Information Seeking and Retrieval in Context*. Dordrecht : Springer, 2005
- [Ip & Rui 2012] IP, Horace H. S. (Ed.) ; RUI, Yong (Ed.): *ICMR '12: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*. New York, NY, USA : ACM, 2012 (ICMR '12)
- [Jacobson et al. 1992] JACOBSON, Ivar ; CHRISTERSON, Magnus ; JONSSON, Patrik ; ÖVERGAARD, Gunnar: *Object-Oriented Software Engineering: A Use Case Driven Approach*. Reading : Addison-Wesley, 1992
- [Jain & Prabhakaran 2013] JAIN, Ramesh (Ed.) ; PRABHAKARAN, Balakrishnan (Ed.): *ICMR '13: Proceedings of the 3rd ACM International Conference on Multimedia Retrieval*. New York, NY, USA : ACM, 2013
- [Järvelin 2011] JÄRVELIN, Kalervo: Evaluation. In: RUTHVEN, Ian (Ed.) ; KELLY, Diane (Ed.): *Interactive Information Seeking, Behaviour and Retrieval*. London : Facet Publ., 2011, 113–138
- [Järvelin & Kekäläinen 2002] JÄRVELIN, Kalervo ; KEKÄLÄINEN, Jaana: Cumulated Gain-based Evaluation of IR Techniques. In: *ACM Trans. Inf. Syst.* 20 (2002), Nr. 4, 422–446

## Bibliography

- [Jøsang 2001] JØSANG, Audun: A Logic for Uncertain Probabilities. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 9 (2001), Nr. 3, 279–212
- [Kamps et al. 2012] KAMPS, Jaap (Ed.) ; KRAAIJ, Wessel (Ed.) ; FUHR, Norbert (Ed.): *Proceedings of the 4th Information Interaction in Context Symposium*. New York, NY, USA : ACM, 2012 (IIIX '12)
- [Käppler 2009] KÄPPLER, Heiko: *Design and Evaluation of Supportive User Interfaces for the Learning of Weights in Multimedia Retrieval (German); Diploma's thesis*. Brandenburg University of Technology, Cottbus, Diplomarbeit, 2009
- [Karlgrén et al. 2011] KARLGRÉN, Jussi ; JÄRVELIN, Anni ; ERIKSSON, Gunnar ; HANSEN, Preben: Use Cases as a Component of Information Access Evaluation. In: *Proceedings of the 2011 Workshop on Data Infrastructures for Supporting Information Retrieval Evaluation*, ACM, 2011 (DESIRE '11), 19–24
- [Kaufman & Rousseeuw 2009] KAUFMAN, Leonard ; ROUSSEEUW, Peter J.: *Wiley Series in Probability and Statistics*. Bd. v.344: *Finding Groups in Data: An Introduction to Cluster Analysis*. 99th ed. Hoboken : John Wiley & Sons Inc, 2009
- [Kelly 2009] KELLY, Diane: Methods for Evaluating Interactive Information Retrieval Systems with Users. In: *Found. Trends Inf. Retr.* 3 (2009), 1–224
- [Kelly & Teevan 2003] KELLY, Diane ; TEEVAN, Jaime: Implicit Feedback for Inferring User Preference: A Bibliography. In: *SIGIR Forum* 37 (2003), Nr. 2, 18–28
- [Kerre et al. 1986] KERRE, Etienne E. ; ZENNER, Rembrand B. R. C. ; DE CALUWE, Rita M. M.: The Use of Fuzzy Set Theory in Information Retrieval and Databases: A Survey. In: *Journal of the American Society for Information Science (JASIS)* 37 (1986), Nr. 5, 341–345
- [Kherfi et al. 2003] KHERFI, M. L. ; ZIOU, D. ; BERNARDI, A.: Combining Positive and Negative Examples in Relevance Feedback for Content-based Image Retrieval. In: *Journal of Visual Communication and Image Representation* 14 (2003), Nr. 4, 428–457
- [Kießling & Köstler 2002] KIESSLING, W. ; KÖSTLER, G.: Preference SQL - Design, Implementation, Experiences. In: *Proc. of the 28th Int. Conf. on Very Large Data Bases VLDB'02, Hong Kong, China, August, 2002*, Morgan Kaufmann Publishers, 2002, 990–1001
- [Kießling 2002] KIESSLING, Werner: Foundations of Preferences in Database Systems. In: *VLDB '02: Proceedings of the 28th International Conference on Very Large Data Bases*, VLDB Endowment, 2002, 311–322
- [Knuth 2013] KNUTH, Donald E.: *The Art of Computer Programming*. Bd. 1: *Fundamental Algorithms*. 3. ed., 31. printing, newly updated and rev. Upper Saddle River, NJ : Addison-Wesley, 2013

- [Kohonen et al. 2000] KOHONEN, T. ; KASKI, S. ; LAGUS, K. ; SALOJARVI, J. ; HONKELA, J. ; PAATERO, V. ; SAARELA, A.: Self Organization of a Massive Document Collection. In: *IEEE Transactions on Neural Networks* 11 (2000), Nr. 3, 574–585
- [Kohonen 1995] KOHONEN, Teuvo: *Springer Series in Information Sciences*. Bd. 30: *Self-Organizing Maps*. Berlin : Springer, 1995
- [Kokar et al. 2004] KOKAR, Mieczyslaw M. ; TOMASIK, Jerzy A. ; WEYMAN, Jerzy: Formalizing Classes of Information Fusion Systems. In: *Information Fusion* 5 (2004), Nr. 3, 189–202
- [Kosch & Maier 2010] KOSCH, Harald ; MAIER, Paul: Content-Based Image Retrieval Systems - Reviewing and Benchmarking. In: *JDIM* 8 (2010), Nr. 1, 54–64
- [Kossmann et al. 2002] KOSSMANN, D. ; RAMSAK, F. ; ROST, S.: Shooting Stars in the Sky: An Online Algorithm for Skyline Queries. In: *Proc. of the 28th Int. Conf. on Very Large Data Bases VLDB'02, Hong Kong, China, August, 2002*, Morgan Kaufmann Publishers, 2002, 275–286
- [Kraft & Buell 1983] KRAFT, Donald H. ; BUELL, Duncan A.: Fuzzy Sets and Generalized Boolean Retrieval Systems. In: *International Journal of Man-Machine Studies* 19 (1983), Nr. 1, 45–56
- [Krippendorff 2004] KRIPPENDORFF, Klaus: Reliability in Content Analysis: Some Common Misconceptions and Recommendations. In: *Human Communication Research* 30 (2004), Nr. 3, 411–433
- [Kruse et al. 1993] KRUSE, Rudolf ; GEBHARDT, Joerg ; KLAWONN, Frank: *Fuzzy-Systeme*. Stuttgart, Germany : Teubner, 1993
- [Kuhlthau 1991] KUHALTHAU, C. C.: Inside the Search Process: Information Seeking from the User's Perspective. In: *Journal of the American Society for Information Science* 42 (1991), Nr. 5, 361–371
- [Kühn et al. 2011] KÜHN, Romina ; KELLER, Christine ; SCHLEGEL, Thomas: Von modellbasierten Storyboards zu kontextsensitiven Interaction-Cases: From Model-Based Storyboards to Context-Aware Interaction-Cases. In: *i-com* 10 (2011), Nr. 3, 12–18
- [Kullback & Leibler 1951] KULLBACK, S. ; LEIBLER, R. A.: On Information and Sufficiency. In: *The Annals of Mathematical Statistics* 22 (1951), Nr. 1, 79–86
- [Laaksonen et al. 1999] LAAKSONEN, Jorma ; KOSKELA, Markus ; LAAKSO, Sami ; OJA, Erkki: PicSOM: Self-organizing Maps for Content-based Image Retrieval. In: *Proceedings of IJCNN* Bd. 4. 1999, 2470–2473
- [Lacroix & Lavency 1987] LACROIX, M. ; LAVENCY, Pierre: Preferences; Putting More Knowledge into Queries. In: *VLDB '87: Proceedings of the 13th International Conference on Very Large Data Bases*, Morgan Kaufmann Publishers Inc., 1987, 217–225

## Bibliography

- [Lalmas 1996] LALMAS, Mounia: *Theories of Information and Uncertainty for the Modelling of Information Retrieval: An Application of Situation Theory and Dempster-Shafer's Theory of Evidence*. Glasgow, University of Glasgow, Diss., 1996
- [Lalmas 1997] LALMAS, Mounia: Dempster-Shafer's Theory of Evidence Applied to Structured Documents: Modelling Uncertainty. In: *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 1997 (SIGIR '97), 110–118
- [Lancaster 1966] LANCASTER, Kelvin J.: A New Approach to Consumer Theory. In: *Journal of Political Economy* 74 (1966), 132
- [Lancaster 1991] LANCASTER, Kelvin J.: *Moderne Mikroökonomie*. 4. Aufl. Frankfurt/-Main : Campus-Verl., 1991
- [Landay & Myers 1996] LANDAY, James A. ; MYERS, Brad A.: Sketching Storyboards to Illustrate Interface Behaviors. In: *Conference Companion on Human Factors in Computing Systems: Common Ground*, ACM, 1996 (CHI '96), 193–194
- [Larsen 2004] LARSEN, Birger: *References and Citations in Automatic Indexing and Retrieval Systems: Experiments with the Boomerang Effect: Dissertation*. Copenhagen, Denmark, Royal School of Library and Information Sciences, Diss., 2004
- [Larsen & Ingwersen 2005] LARSEN, Birger ; INGWERSEN, Peter: Cognitive Overlaps along the Polyrepresentation Continuum. In: SPINK, Amanda (Ed.) ; COLE, Charles (Ed.): *New Directions in Cognitive Information Retrieval* Bd. 19. Springer Netherlands, 2005, 43-60
- [Larsen et al. 2006] LARSEN, Birger ; INGWERSEN, Peter ; KEKÄLÄINEN, Jaana: The Polyrepresentation Continuum in IR. In: *IiX: Proceedings of the 1st International Conference on Information Interaction in Context*, ACM, 2006, 88–96
- [Larsen et al. 2009] LARSEN, Birger ; INGWERSEN, Peter ; LUND, Berit: Data Fusion According to the Principle of Polyrepresentation. In: *J. Am. Soc. Inf. Sci. Technol.* 60 (2009), Nr. 4, 646–654
- [Larson et al. 2008] LARSON, Martha ; NEWMAN, Eamonn ; JONES, Gareth J. F.: Overview of VideoCLEF 2008: Automatic Generation of Topic-Based Feeds for Dual Language Audio-Visual Content. In: PETERS, Carol (Ed.) ; DESELAERS, Thomas (Ed.) ; FERRO, Nicola (Ed.) ; GONZALO, Julio (Ed.) ; JONES, Gareth J. F. (Ed.) ; KURIMO, Mikko (Ed.) ; MANDL, Thomas (Ed.) ; PEÑAS, Anselmo (Ed.) ; PETRAS, Vivien (Ed.): *CLEF* Bd. 5706, Springer, 2008 (Lecture Notes in Computer Science), 906-917
- [Lee et al. 2007] LEE, Jongwuk ; YOU, Gae-won ; HWANG, Seung-won: Telescope: Zooming to Interesting Skylines. In: KOTAGIRI, Ramamohanarao (Ed.) ; KRISHNA, P. R. (Ed.) ; MOHANIA, Mukesh (Ed.) ; NANTAJEEWARAWAT, Ekawit (Ed.): *Advances in Databases: Concepts, Systems and Applications, 12th International Conference*

- on Database Systems for Advanced Applications, DASEAA 2007, Bangkok, Thailand, April 9-12, 2007, Proceedings Bd. 4443. Springer, 2007, 539–550*
- [Lee 1994] LEE, Joon Ho: Properties of Extended Boolean Models in Information Retrieval. In: *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, Springer-Verlag New York, Inc., 1994, 182–190
- [Lee et al. 1994] LEE, Joon Ho ; KIM, Myoung Ho ; LEE, Yoon Joon: Ranking Documents in Thesaurus-based Boolean Retrieval Systems. In: *Inf. Process. Manage.* 30 (1994), Nr. 1, 79–91
- [Lehrack & Schmitt 2010] LEHRACK, Sebastian ; SCHMITT, Ingo: QSQL: Incorporating Logic-Based Retrieval Conditions into SQL. In: KITAGAWA, Hiroyuki (Ed.) ; ISHIKAWA, Yoshiharu (Ed.) ; LI, Qing (Ed.) ; WATANABE, Chiemi (Ed.): *Database Systems for Advanced Applications, 15th International Conference, DASEAA 2010, Tsukuba, Japan, April 1-4, 2010, Proceedings, Part I Bd. 5981. Springer, 2010, 429–443*
- [Lesk & Salton 1968] LESK, M. E. ; SALTON, Gerard: Relevance Assessments and Retrieval System Evaluation. In: *Information Storage and Retrieval* 4 (1968), Nr. 4, 343–359
- [Lew et al. 2006] LEW, Michael S. ; SEBE, Nicu ; DJERABA, Chabane ; JAIN, Ramesh: Content-based Multimedia Information Retrieval: State of the Art and Challenges. In: *ACM Trans. Multimedia Comput. Commun. Appl.* 2 (2006), Nr. 1, 1–19
- [Lioma et al. 2012] LIOMA, Christina ; LARSEN, Birger ; INGWERSEN, Peter: Preliminary Experiments using Subjective Logic for the Polyrepresentation of Information Needs. In: *Proceedings of the 4th Information Interaction in Context Symposium, ACM, 2012 (IIIX '12)*, 174–183
- [Lioma et al. 2010] LIOMA, Christina ; LARSEN, Birger ; SCHUETZE, Hinrich ; INGWERSEN, Peter: A Subjective Logic Formalisation of the Principle of Polyrepresentation for Information Needs. In: *Proceeding of the Third Symposium on Information Interaction in Context, ACM, 2010 (IIIX '10)*, 125–134
- [Liu et al. 2010] LIU, Haiming ; MULHOLLAND, Paul ; SONG, Dawei ; UREN, Victoria ; RÜGER, Stefan: Applying Information Foraging Theory to Understand User Interaction with Content-based Image Retrieval. In: *Proceedings of the Third Symposium on Information Interaction in Context, ACM, 2010 (IIIX '10)*, 135–144
- [Liu et al. 2009] LIU, Haiming ; ZAGORAC, Srdan ; UREN, Victoria ; SONG, Dawei ; RÜGER, Stefan: Enabling Effective User Interactions in Content-Based Image Retrieval. In: *Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology, Springer-Verlag, 2009 (AIRS '09)*, 265–276
- [Liu 2011] LIU, Tie-Yan: *Learning to Rank for Information Retrieval*. Berlin, Heidelberg : Springer-Verlag Berlin Heidelberg, 2011



## Bibliography

- [Lofi et al. 2008] LOFI, Christoph ; BALKE, Wolf-Tilo ; GÜNTZER, Ulrich: Consistency Check Algorithms for Multi-Dimensional Preference Trade-Offs. In: *IJCSA* 5 (2008), Nr. 3b, 165–185
- [Lowe 2004] LOWE, David G.: Distinctive Image Features from Scale-Invariant Key-points. In: *Int. J. Comput. Vision* 60 (2004), Nr. 2, 91–110
- [Lux & Chatzichristofis 2008] LUX, Mathias ; CHATZICHRISTOFIS, Savvas A.: Lire: Lucene Image Retrieval: An Extensible Java CBIR Library. In: *MM '08: Proceeding of the 16th ACM International Conference on Multimedia*, ACM, 2008, 1085–1088
- [Malone 1983] MALONE, Thomas W.: How Do People Organize Their Desks?: Implications for the Design of Office Information Systems. In: *ACM Trans. Inf. Syst.* 1 (1983), Nr. 1, 99–112
- [Manjunath et al. 2002] MANJUNATH, B. S. ; SALEMBIER, P. ; SIKORA, T.: *Introduction to MPEG-7: Multimedia Content Description Interface*. New York, NY, USA : John Wiley & Sons, Inc., 2002
- [Manning et al. 2009] MANNING, Christopher D. ; RAGHAVAN, Prabhakar ; SCHÜTZE, Hinrich: *Introduction to Information Retrieval*. Reprinted. Cambridge : Cambridge Univ. Press, 2009
- [Marchionini et al. 2000] MARCHIONINI, Gary ; GEISLER, Gary ; BRUNK, Ben: Agileviews: A Human-Centered Framework for Interfaces to Information Spaces. In: *Proceedings of the Annual Conference of the American Society for Information Science*, 2000, 271–280
- [McDonald et al. 2001] MCDONALD, Sharon ; LAI, Ting-Sheng ; TAIT, John: Evaluating a Content Based Image Retrieval System. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2001 (SIGIR '01), 232–240
- [McDonald & Tait 2003] MCDONALD, Sharon ; TAIT, John: Search Strategies in Content-based Image Retrieval. In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, ACM, 2003 (SIGIR '03), 80–87
- [McLuhan 2010] MCLUHAN, Marshall: *Understanding Media: The Extensions of Man*. Repr. London : Routledge, 2010 (Routledge Classics)
- [Melucci 2008] MELUCCI, Massimo: A Basis for Information Retrieval in Context. In: *ACM Trans. Inf. Syst.* 26 (2008), 14:1–14:41
- [Melucci 2011] MELUCCI, Massimo: Can Information Retrieval Systems be Improved using Quantum Probability? In: *Proceedings of the Third International Conference on Advances in Information Retrieval Theory*, Springer-Verlag, 2011 (ICTIR'11), 139–150

- [Mendenhall et al. 1999] MENDENHALL, W. ; BEAVER, R. J. ; BEAVER, B. M.: *Introduction to Probability and Statistics*. ITP Duxbury Press, 1999
- [Miller & Miller 1999] MILLER, I. ; MILLER, M.: *John E. Freund's Mathematical Statistics*. Upper Saddle River, New Jersey 07458 : Prentice Hall, 1999
- [Mizuochi et al. 2013] MIZUOCHI, Masaru ; HIGUCHI, Takayuki ; KAMADA, Chie ; HARADA, Tatsuya: MIL at ImageCLEF 2013: Personal Photo Retrieval. In: *CLEF 2013 Labs and Workshop, Notebook Papers, 23-26 September 2013, Valencia, Spain*. 2013
- [Morville & Callender 2010] MORVILLE, Peter ; CALLENDER, Jeffery: *Search Patterns: Design for Discovery*. 1st ed. Sebastopol, Calif. : O'Reilly, 2010 (Safari Tech Books Online)
- [Motro 1986] MOTRO, Amihai: Extending the Relational Database Model to Support Goal Queries. In: KERSCHBERG, Larry (Ed.): *Proceedings of the First International Conference on Expert Database Systems*, 1986, 129–150
- [Müller et al. 2010] MÜLLER, Henning ; CLOUGH, Paul ; DESELAERS, Thomas ; CAPUTO, Barbara: *The Information Retrieval Series*. Bd. 32: *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*. Berlin, Heidelberg : Springer-Verlag Berlin Heidelberg, 2010
- [Mulwa et al. 2011] MULWA, Catherine ; LIU, Wei ; LAWLESS, Séamus ; JONES, Gareth J. F.: A Proposal for the Evaluation of Adaptive Information Retrieval Systems using Simulated Interaction. In: AZZOPARDI, Leif (Ed.) ; JÄRVELIN, Kalervo (Ed.) ; KAMPS, Jaap (Ed.) ; SMUCKER, Mark D. (Ed.): *Report on the SIGIR 2010 Workshop on the Simulation of Interaction* Bd. 44, ACM, 2011, 5–6
- [Muramatsu & Pratt 2001] MURAMATSU, Jack ; PRATT, Wanda: Transparent Queries: Investigation Users' Mental Models of Search Engines. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2001 (SIGIR '01), 217–224
- [Myoupo et al. 2010] MYOUPPO, Débora ; POPESCU, Adrian ; LE BORGNE, Hervé ; MOËLLIC, Pierre-Alain: Multimodal Image Retrieval over a Large Database. In: *Proceedings of the 10th international conference on Cross-language evaluation forum: multimedia experiments*, Springer-Verlag, 2010 (CLEF'09), 177–184
- [Nakazato & Huang 2001] NAKAZATO, Munehiro ; HUANG, Thomas S.: 3d Mars: Immersive Virtual Reality for Content-Based Image Retrieval. In: *Proceedings of the 2001 IEEE International Conference on Multimedia and Expo, ICME 2001, August 22-25, 2001, Tokyo, Japan*. IEEE Computer Society, 2001
- [Nardi 1993] NARDI, Bonnie A.: *A Small Matter of Programming: Perspectives on End User Computing*. MIT Press, 1993

## Bibliography

- [Nelder & Mead 1965] NELDER, J. A. ; MEAD, R.: A Simplex Method for Function Minimization. In: *Computer Journal* 7 (1965), 308–313
- [Nielsen 2009] NIELSEN, Jakob: *Usability engineering*. [Nachdr.]. Amsterdam : Kaufmann, 2009
- [Nielsen & Landauer 1993] NIELSEN, Jakob ; LANDAUER, Thomas K.: A Mathematical Model of the Finding of Usability Problems. In: *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, ACM, 1993 (CHI '93), 206–213
- [Nilsson 1986] NILSSON, Nils J.: Probabilistic logic. In: *Artif. Intell.* 28 (1986), 71–88
- [Nilsson 1994] NILSSON, Nils J.: Probabilistic logic revisited. Cambridge, MA, USA : MIT Press, 1994, 39–42
- [Nørgaard & Hornbæk 2006] NØRGAARD, Mie ; HORNBÆK, Kasper: What Do Usability Evaluators Do in Practice?: An Explorative Study of Think-aloud Testing. In: *Proceedings of the 6th Conference on Designing Interactive Systems*, ACM, 2006 (DIS '06), 209–218
- [Nottelmann & Fuhr 2003] NOTTELMANN, Henrik ; FUHR, Norbert: From Uncertain Inference to Probability of Relevance for Advanced IR Applications. In: SEBASTIANI, Fabrizio (Ed.): *Advances in Information Retrieval* Bd. 2633. Springer Berlin / Heidelberg, 2003, 79-79
- [Oddy 1977] ODDY, R. N.: Information Retrieval through Man-Machine Dialogue. In: *Journal of Documentation* 33 (1977), Nr. 1, 1–14
- [Oviatt 1999] OVIATT, Sharon: Ten Myths of Multimodal Interaction. In: *Communications of the ACM* 42 (1999), Nr. 11, 74–81
- [Pearl 2008] PEARL, Judea: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Rev. 2. print., 12. [Dr.]. San Francisco, Calif. : Kaufmann, 2008 (The Morgan Kaufmann Series in Representation and Reasoning)
- [Pirolli 2007] PIROLI, Peter: *Information Foraging Theory: Adaptive Interaction with Information*. New York : Oxford University Press, 2007 (Oxford Series in Human-Technology Interaction)
- [Pirolli & Card 1995] PIROLI, Peter ; CARD, Stuart: Information Foraging in Information Access Environments. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM Press/Addison-Wesley Publishing Co., 1995 (CHI '95), 51–58
- [Piwowarski et al. 2010] PIWOWARSKI, Benjamin ; FROMMHOLZ, Ingo ; LALMAS, Mounia ; VAN RIJSBERGEN, Cornelis J.: What Can Quantum Theory Bring to Information Retrieval? In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ACM, 2010 (CIKM '10), 59–68

- [Piwowarski & Lalmas 2009] PIWOWARSKI, Benjamin ; LALMAS, Mounia: A Quantum-Based Model for Interactive Information Retrieval. In: *Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory*, Springer-Verlag, 2009 (ICTIR '09), 224–231
- [Popescu et al. 2010] POPESCU, Adrian ; TSIKRIKA, Theodora ; KLUDAS, Jana: Overview of the Wikipedia Retrieval Task at ImageCLEF 2010. In: BRASCHLER, Martin (Ed.) ; HARMAN, Donna (Ed.) ; PIANTA, Emanuele (Ed.): *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy*. 2010
- [Preece et al. 2002] PREECE, Jennifer ; ROGERS, Yvonne ; SHARP, Helen: *Interaction design: Beyond human-computer interaction*. New York, NY : Wiley, 2002
- [Press et al. 2005] PRESS, William H. ; TEUKOLSKY, Saul A. ; VETTERLING, William T. ; FLANNERY, Brian P.: *Numerical Recipes in C++: The Art of Scientific Computing*. 2. ed., reprint. with corr. to software version 2.11. Cambridge : Cambridge Univ. Press, 2005
- [R Core Team 2012] R CORE TEAM ; R FOUNDATION FOR STATISTICAL COMPUTING (Ed.): *R: A Language and Environment for Statistical Computing*. Version: 2012. ( 2.15)
- [Reiterer et al. 2000] REITERER, Harald ; MUSSLER, Gabriela ; MANN, Thomas M. ; HANDSCHUH, Siegfried: INSYDER - An Information Assistant for Business Intelligence. In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2000 (SIGIR '00), 112–119
- [Robertson 1977] ROBERTSON, Stephen E.: The Probability Ranking Principle in IR. In: *Journal of Documentation* Bd. 4. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1977, 281–286
- [Robertson & Hancock-Beaulieu 1992] ROBERTSON, Stephen E. ; HANCOCK-BEAULIEU, Micheline: On the Evaluation of IR Systems. In: *Inf. Process. Manage.* 28 (1992), Nr. 4, 457–466
- [Robertson & Spärck Jones 1976] ROBERTSON, Stephen E. ; SPÄRCK JONES, Karen: Relevance Weighting of Search Terms. In: *Journal of the American Society for Information Science* 27 (1976), Nr. 3, 129–146
- [Rocchio 1971] ROCCHIO, J.: Relevance Feedback in Information Retrieval. In: *The SMART Retrieval System*. 1971, 313–323
- [Rodden et al. 2001] RODDEN, Kerry ; BASALAJ, Wojciech ; SINCLAIR, David ; WOOD, Kenneth: Does Organisation by Similarity Assist Image Browsing? In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2001 (CHI '01), 190–197
- [Rodden & Wood 2003] RODDEN, Kerry ; WOOD, Kenneth R.: How Do People Manage Their Digital Photographs? In: *CHI '03: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2003, 409–416

## Bibliography

- [Rubner et al. 1998] RUBNER, Yossi ; TOMASI, Carlo ; GUIBAS, Leonidas J.: A Metric for Distributions with Applications to Image Databases. In: *Proceedings of the Sixth International Conference on Computer Vision*, IEEE Computer Society, 1998 (ICCV '98), 59–
- [Russell et al. 2007] RUSSELL, Stuart ; NORVIG, Peter ; CANNY, John F.: *Künstliche Intelligenz: Ein moderner Ansatz*. 2. Aufl., [Nachdr.]. München : Pearson Studium, 2007 (it-InformatikKünstliche Intelligenz)
- [Russell-Rose & Tate 2013] RUSSELL-ROSE, Tony ; TATE, Tyler: *Designing the Search Experience: The Information Architecture of Discovery*. Amsterdam, Boston, Heidelberg : Morgan Kaufmann, 2013
- [Ruthven & Kelly 2011] RUTHVEN, Ian (Ed.) ; KELLY, Diane (Ed.): *Interactive Information Seeking, Behaviour and Retrieval*. London : Facet Publ., 2011
- [Ruthven & Lalmas 2003] RUTHVEN, Ian ; LALMAS, Mounia: A Survey on the Use of Relevance Feedback for Information Access Systems. In: *Knowl. Eng. Rev.* 18 (2003), Nr. 2, 95–145
- [Ruthven et al. 2002] RUTHVEN, Ian ; LALMAS, Mounia ; VAN RIJSBERGEN, Cornelis J.: Combining and Selecting Characteristics of Information Use. In: *J. Am. Soc. Inf. Sci. Technol.* 53 (2002), Nr. 5, 378–396
- [Saake et al. 2010] SAAKE, Gunter ; SATTLER, Kai-Uwe ; HEUER, Andreas: *Datenbanken: Konzepte und Sprachen*. 4. Aufl. Heidelberg, Hamburg : mitp Verl.-Gruppe Hüthig Jehle Rehm, 2010 (Biber-Buch)
- [Salton & Buckley 1990] SALTON, Gerard ; BUCKLEY, Chris: Improving Retrieval Performance by Relevance Feedback. In: *Journal of the American Society for Information Science (JASIS)* 41 (1990), Nr. 4, 288–297
- [Salton & Buckley 1988] SALTON, Gerard ; BUCKLEY, Christopher: Term-weighting Approaches in Automatic Text Retrieval. In: *Inf. Process. Manage.* 24 (1988), Nr. 5, 513–523
- [Salton et al. 1983] SALTON, Gerard ; FOX, Edward A. ; WU, Harry: Extended Boolean Information Retrieval. In: *Commun. ACM* 26 (1983), Nr. 11, 1022–1036
- [Santini 2012] SANTINI, Simone: Because Not All Displays are Lists. In: *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ACM, 2012 (ICMR '12), 10:1–10:8
- [Santini & Jain 2000] SANTINI, Simone ; JAIN, Ramesh: Integrated Browsing and Querying for Image Databases. In: *IEEE MultiMedia* 7 (2000), Nr. 3, 26–39
- [SAS Publishing 2011] SAS PUBLISHING: *Base SAS 9.3 Procedures Guide: Statistical Procedures*. Cary NC : SAS Institute Inc., 2011

- [Schaefer & Stich 2004] SCHAEFER, G. ; STICH, M.: UCID - An Uncompressed Colour Image Database. In: *Proc. SPIE, Storage and Retrieval Methods and Applications for Multimedia*. San Jose, USA, 2004, 472–480
- [Schiela 2010] SCHIELA, Karsten: *Ein CQQL-basiertes Musikretrievalsystem für GlobalMusic2one (German); Master's thesis*. Brandenburg University of Technology, Cottbus, Diplomarbeit, 2010
- [Schlegel & Raschke 2010] SCHLEGEL, Thomas ; RASCHKE, Michael: Interaction-Cases: Model-Based Description of Complex Interactions in Use Cases. In: *Proceedings of the IADIS International Conferences Interfaces and Human Computer Interaction 2010 and Game and Entertainment Technologies 2010*. 2010, 195–202
- [Schmettow 2012] SCHMETTOW, Martin: Sample Size in Usability Studies. In: *Commun. ACM* 55 (2012), Nr. 4, 64–70
- [Schmitt 2006] SCHMITT, Ingo: *Ähnlichkeitssuche in Multimedia-Datenbanken: Retrieval, Suchalgorithmen und Anfragebehandlung*. München : Oldenbourg, 2006
- [Schmitt 2007] SCHMITT, Ingo ; BRANDENBURG UNIVERSITY OF TECHNOLOGY AT COTTBUS (Ed.): *Weighting in CQQL*. Cottbus, 2007 (Computer Science Reports 4)
- [Schmitt 2008] SCHMITT, Ingo: QQL: A DB&IR Query Language. In: *The VLDB Journal* 17 (2008), Nr. 1, 39–56
- [Schmitt et al. 2009] SCHMITT, Ingo ; NÜRNBERGER, Andreas ; LEHRACK, Sebastian: On the Relation between Fuzzy and Quantum Logic. In: SEISING, Rudolf (Ed.): *Views on Fuzzy Sets and Systems from Different Perspectives* Bd. 243. Springer Berlin Heidelberg, 2009, 417-438
- [Schmitt et al. 2005] SCHMITT, Ingo ; SCHULZ, Nadine ; HERSTEL, Thomas: WS-QBE: A QBE-like Query Language for Complex Multimedia Queries. In: *Proc. of the 11th Int. Multimedia Modelling Conf. (MMM'05)*, IEEE CS Press, 2005, 222–229
- [Schmitt & Zellhöfer 2009] SCHMITT, Ingo ; ZELHÖFER, David: Lernen nutzerspezifischer Gewichte innerhalb einer logikbasierten Anfragesprache. In: FREYTAG, Christoph J. (Ed.) ; RUF, Thomas (Ed.) ; LEHNER, Wolfgang (Ed.) ; VOSSEN, Gottfried (Ed.): *Datenbanksysteme in Business, Technologie und Web (BTW 2009)*, 13. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme (DBIS), Proceedings, 2.-6. März 2009, Münster, Germany Bd. 144. GI, 2009, 137–156
- [Schmitt & Zellhöfer 2012] SCHMITT, Ingo ; ZELHÖFER, David: Condition Learning from User Preferences. In: *6th IEEE International Conference on Research Challenges in Information Science*. Valencia, Spain : IEEE, 2012
- [Schmitt et al. 2008] SCHMITT, Ingo ; ZELHÖFER, David ; NÜRNBERGER, Andreas: Towards Quantum Logic Based Multimedia Retrieval. In: IEEE (Ed.): *Proceedings of the Fuzzy Information Processing Society (NAFIPS)*, IEEE, 2008, 1-6

## Bibliography

- [Schöning 1989] SCHÖNING, U.: *Logik für Informatiker*. Bibliographisches Institut, Mannheim, 1989
- [Schulz & Schmitt 2003] SCHULZ, N. ; SCHMITT, Ingo: Relevanzwichtung in komplexen Ähnlichkeitsanfragen. In: WEIKUM, G. (Ed.) ; SCHÖNING, H. (Ed.) ; RAHM, E. (Ed.): *Datenbanksysteme in Business, Technologie und Web, BTW'03, 10. GI-Fachtagung, Leipzig Februar 2003*, Gesellschaft für Informatik, 2003 (Lecture Notes in Informatics (LNI) Volume P-26), 187–196
- [Sedgewick & Wayne 2012] SEDGEWICK, Robert ; WAYNE, Kevin: *Algorithms*. 4. ed., 3. printing. Upper Saddle River, NJ : Addison-Wesley, 2012
- [Shang & Kitsuregawa 2013] SHANG, Haichuan ; KITSUREGAWA, Masaru: Skyline Operator on Anti-correlated Distributions. In: *Proc. VLDB Endow.* 6 (2013), Nr. 9, 649–660
- [Shannon 1948] SHANNON, Claude Elwood: A Mathematical Theory of Communication. In: *Bell System Technical Journal* 27 (1948), 379–423
- [Shannon & Weaver 1949] SHANNON, Claude Elwood ; WEAVER, Warren: *The Mathematical Theory of Communication*. Urbana : Univ. of Illinois Press, 1949
- [Shneiderman 1987] SHNEIDERMAN, Ben: Direct Manipulation: A Step beyond Programming Languages. In: BAECKER, R. M. (Ed.) ; BUXTON, W. A. S. (Ed.): *Human-Computer Interaction*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1987, 461–467
- [Shneiderman 1996] SHNEIDERMAN, Ben: The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: *Proceedings of the 1996 IEEE Symposium on Visual Languages*, IEEE Computer Society, 1996 (VL '96), 336-343
- [Shneiderman 2003] SHNEIDERMAN, Ben: Promoting Universal Usability with Multi-Layer Interface Design. In: *CUU '03: Proceedings of the 2003 Conference on Universal Usability*, ACM, 2003, 1–8
- [Shneiderman & Plaisant 2005] SHNEIDERMAN, Ben ; PLAISANT, Catherine: *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. 4. ed. Boston : Pearson, 2005
- [Siefkes 1990] SIEFKES, D.: *Formalisieren und Beweisen. Logik für Informatiker*. Braunschweig : Vieweg-Verlag, 1990
- [Silberschatz et al. 1999] SILBERSCHATZ, Abraham ; KORTH, Henry F. ; SUDARSHAN, S.: *Database system concepts*. 3. ed. Boston, Mass. : WCB McGraw-Hill, 1999 (McGraw-Hill Series in Computer Science)
- [Sinha et al. 2009] SINHA, Pinaki ; PIRSIYAVASH, Hamed ; JAIN, Ramesh: Personal Photo Album Summarization. In: *Proceedings of the 17th ACM International Conference on Multimedia*, ACM, 2009 (MM '09), 1131–1132

- [Skov et al. 2004] SKOV, Metter ; PEDERSEN, Henriette ; LARSEN, Birger ; INGWERSEN, Peter: Testing the Principle of Polyrepresentation. In: INGWERSEN, Peter (Ed.) ; VAN RIJSBERGEN, Cornelis J. (Ed.) ; BELKIN, Nicholas (Ed.): *Proceedings of ACM SIGIR 2004 Workshop on "Information Retrieval in Context"*, 2004, 47–49
- [Slaney 2010] SLANEY, Malcolm: Multimodal Retrieval and Ranking: More than Waveforms. In: *Proceedings of the International Conference on Multimedia Information Retrieval*, ACM, 2010 (MIR '10), 241–242
- [Smeaton et al. 2006] SMEATON, Alan F. ; OVER, Paul ; KRAAIJ, Wessel: Evaluation Campaigns and TRECVID. In: *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, ACM Press, 2006, 321–330
- [Smeulders et al. 2000] SMEULDERS, A. W. M. ; WORRING, M. ; SANTINI, S. ; GUPTA, A. ; JAIN, R.: Content-Based Image Retrieval at the End of the Early Years. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence, (PAMI) 22* (2000), Nr. 12
- [Smucker et al. 2007] SMUCKER, Mark D. ; ALLAN, James ; CARTERETTE, Ben: A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In: *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, ACM, 2007 (CIKM '07), 623–632
- [Soergel 1985] SOERGEL, Dagobert: *Organizing Information: Principles of Data Base and Retrieval Systems*. San Diego, CA, USA : Academic Press Professional, Inc., 1985
- [Spendley et al. 1962] SPENDLEY, W. ; HEXT, G. R. ; HIMSWORTH, F. R.: Sequential Application of Simplex Designs in Optimisation and Evolutionary Operation. In: *Technometrics 4* (1962), Nr. 4, 441–461
- [Spink et al. 1998] SPINK, Amanda ; GREISDORF, Howard ; BATEMAN, Judy: From Highly Relevant to Not Relevant: Examining Different Regions of Relevance. In: *Inf. Process. Manage.* 34 (1998), Nr. 5, 599–621
- [Stehling et al. 2002] STEHLING, Renato O. ; NASCIMENTO, Mario A. ; FALCÃO, Alexandre X.: A Compact and Efficient Image Retrieval Approach Based on Border/Interior Pixel Classification. In: *Proceedings of the Eleventh International Conference on Information and Knowledge management*, ACM, 2002 (CIKM '02), 102–109
- [Strohman et al. 2004] STROHMAN, Trevor ; METZLER, Donald ; TURTLE, Howard ; CROFT, Bruce W.: Indri: A Language-model Based Search Engine for Complex Queries. In: *Proceedings of the International Conference on Intelligent Analysis*. 2004
- [Sung & Hu 2009] SUNG, Sam Yuan ; HU, Tianming: Combining Weights into Scores: A Linear Transform Approach. In: *IEEE International Conference on Systems, Man and Cybernetics, 2009. SMC 2009, 2009*, 2636–2641
- [Tahaghoghi et al. 2002] TAHAGHOGHI, Seyed ; THOM, James ; WILLIAMS, Hugh ; LAENDER, Alberto (Ed.) ; OLIVEIRA, Arlindo (Ed.): *Multiple Example Queries in Content-Based Image Retrieval*. Version: 2002 (Lecture Notes in Computer Science)



## Bibliography

- [Tamura et al. 1978] TAMURA, Hideyuki ; MORI, Shunji ; YAMAWAKI, Takashi: Texture Features Corresponding to Visual Perception. In: *IEEE Transactions on System, Man and Cybernetic* 8 (1978), Nr. 6, 460–472
- [Tankard 1979] TANKARD, James W.: The H.G. Wells Quote on Statistics: A Question of Accuracy. In: *Historia Mathematica* 6 (1979), Nr. 1, 30-33
- [Taylor 2006] TAYLOR, Arlene G.: *Introduction to Cataloging and Classification*. 10th ed. Westport, Conn. : Libraries Unlimited, 2006 (Library and Information Science Text Series)
- [Thomee & Popescu 2012] THOMEE, Bart ; POPESCU, Adrian: Overview of the Image-CLEF 2012 Flickr Photo Annotation and Retrieval Task. In: FORNER, Pamela (Ed.) ; KARLGREN, Jussi (Ed.) ; WOMSER-HACKER, Christa (Ed.): *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*. 2012
- [Torlone & Ciaccia 2002] TORLONE, Riccardo ; CIACCIA, Paolo: Which Are My Preferred Items? In: *Workshop on Recommendation and Personalization in E-Commerce* (2002)
- [Trotter 1992] TROTTER, T. W.: *Combinatorics and partially ordered sets: Dimension theory*. Baltimore : Johns Hopkins Univ. Pr., 1992 (Johns Hopkins Series in the Mathematical Sciences)
- [Tsikrika et al. 2011] TSIKRIKA, Theodora ; POPESCU, Adrian ; KLUDAS, Jana: Overview of the Wikipedia Image Retrieval Task at ImageCLEF 2011. In: *CLEF (Notebook Papers/Labs/Workshops) : CLEF 2011 Working Notes*, 2011
- [Tukey 1977] TUKEY, John W.: *Exploratory Data Analysis*. Addison-Wesley, 1977
- [Tunkelang 2011] TUNKELANG, Daniel: Using QPP to Simulate Query Refinement. In: AZZOPARDI, Leif (Ed.) ; JÄRVELIN, Kalervo (Ed.) ; KAMPS, Jaap (Ed.) ; SMUCKER, Mark D. (Ed.): *Report on the SIGIR 2010 Workshop on the Simulation of Interaction* Bd. 44, ACM, 2011, 7–8
- [Turtle & Croft 1991] TURTLE, Howard ; CROFT, W. Bruce: Evaluation of an Inference Network-based Retrieval Model. In: *ACM Trans. Inf. Syst.* 9 (1991), Nr. 3, 187–222
- [Uhlig 2010] UHLIG, Markus: *Visualization of Queries and Query Results for a Retrieval System (German); Bachelor's thesis*. Brandenburg University of Technology, Cottbus, Diplomarbeit, 2010
- [Urban & Jose 2006] URBAN, Jana ; JOSE, Joemon M.: Can a Workspace Help to Overcome the Query Formulation Problem in Image Retrieval? In: LALMAS, Mounia (Ed.) ; MACFARLANE, Andy (Ed.) ; RÜGER, Stefan (Ed.) ; TOMBROS, Anastasios (Ed.) ; TSIKRIKA, Theodora (Ed.) ; YAVLINSKY, Alexei (Ed.): *Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006, Proceedings* Bd. 3936. Springer, 2006, 385–396

- [Urban et al. 2006] URBAN, Jana ; JOSE, Joemon M. ; VAN RIJSBERGEN, Cornelis J.: An Adaptive Technique for Content-based Image Retrieval. In: *Multimedia Tools Appl.* 31 (2006), Nr. 1, 1–28
- [van der Heijden 2003] VAN DER HEIJDEN, Hans: Factors Influencing the Usage of Websites: The Case of a Generic Portal in The Netherlands. In: *Inf. Manage.* 40 (2003), Nr. 6, 541–549
- [van Rijsbergen 1979] VAN RIJSBERGEN, Cornelis J.: *Information Retrieval*. 2. London : Butterworths, 1979
- [van Rijsbergen 1986a] VAN RIJSBERGEN, Cornelis J.: A New Theoretical Framework for Information Retrieval. In: *SIGIR Forum* 21 (1986), Nr. 1-2, 23–29
- [van Rijsbergen 1986b] VAN RIJSBERGEN, Cornelis J.: A Non-classical Logic for Information Retrieval. In: *The Computer Journal* Bd. 29(6). San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1986, 481–485
- [van Rijsbergen 2004] VAN RIJSBERGEN, Cornelis J.: *The Geometry of Information Retrieval*. Cambridge, England : Cambridge University Press, 2004
- [van Rijsbergen 2009] VAN RIJSBERGEN, Cornelis J.: *A Brief Introduction to Information Retrieval: Keynote at the 7th ESSIR*. Padua, 2009 (European Summer School in Information Retrieval)
- [van Rijsbergen & Lalmas 1996] VAN RIJSBERGEN, Cornelis J. ; LALMAS, Mounia: Information Calculus for Information Retrieval. In: *J. Am. Soc. Inf. Sci.* 47 (1996), Nr. 5, 385–398
- [Veltkamp & Tanase 2002] VELTKAMP, Remco C. ; TANASE, Mirela: *Content-based Image Retrieval Systems: A Survey: Revised and Extended Version 2002*. Version: 2002
- [von Neumann 1932] VON NEUMANN, John: *Grundlagen der Quantenmechanik*. Berlin, Heidelberg, New York : Springer Verlag, 1932
- [Voorhees 2008] VOORHEES, Ellen M.: On Test Collections for Adaptive Information Retrieval. In: *Information Processing & Management* 44 (2008), Nr. 6, 1879–1885
- [Voorhees & Harman 2005] VOORHEES, Ellen M. ; HARMAN, Donna K.: *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, Mass. : MIT Press, 2005 (Digital Libraries and Electronic Publishing)
- [Waller & Kraft 1979] WALLER, W. G. ; KRAFT, Donald H.: A Mathematical Model of a Weighted Boolean Retrieval System. In: *Information Processing and Management* 15 (1979), Nr. 5, 235–245
- [Wang et al. 2001] WANG, James Z. ; LI, Jia ; WIEDERHOLD, Gio: SIMPLiCity: Semantics-Sensitive Integrated Matching for Picture Libraries. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (2001), 947–963

## Bibliography

- [Wang et al. 2009] WANG, Meng ; YANG, Linjun ; HUA, Xian-Sheng ; MICROSOFT RESEARCH (Ed.): *MSRA-MM: Bridging Research and Industrial Societies for Multimedia Information Retrieval*. 2009. ( MSR-TR-2009-30)
- [Weikum 2007] WEIKUM, Gerhard: DB&IR: Both Sides Now. In: ACM (Ed.): *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, ACM, 2007, 25–30
- [White 2004a] WHITE, Ryen W.: Contextual Simulations for Information Retrieval Evaluation. In: *SIGIR 2004: Information Retrieval in Context Workshop*, 2004
- [White 2004b] WHITE, Ryen W.: *Implicit Feedback for Interactive Information Retrieval*, University of Glasgow, Diss., 2004
- [White 2006] WHITE, Ryen W.: Using Searcher Simulations to Redesign a Polyrepresentative Implicit Feedback Interface. In: *Inf. Process. Manage.* 42 (2006), Nr. 5, 1185–1202
- [White & Roth 2009] WHITE, Ryen W. ; ROTH, Resa: *Exploratory Search: Beyond the Query-Response Paradigm*. San Rafael : Morgan & Claypool Publishers, 2009 (Synthesis Lectures on Information Concepts, Retrieval, and Services)
- [White & Ruthven 2006] WHITE, Ryen W. ; RUTHVEN, Ian: A Study of Interface Support Mechanisms for Interactive Information Retrieval. In: *J. Am. Soc. Inf. Sci. Technol.* 57 (2006), Nr. 7, 933–948
- [White et al. 2005] WHITE, Ryen W. ; RUTHVEN, Ian ; JOSE, Joemon M.: A Study of Factors Affecting the Utility of Implicit Relevance Feedback. In: *SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2005, 35–42
- [Wiedenbeck & Zila 1997] WIEDENBECK, Susan ; ZILA, Patti L.: Hands-on Practice in Learning to Use Software: A Comparison of Exercise, Exploration, and Combined Formats. In: *ACM Trans. Comput.-Hum. Interact.* 4 (1997), 169–196
- [Wilcoxon 1945] WILCOXON, Frank: Individual Comparisons by Ranking Methods. In: *Biometrics Bulletin* 1 (1945), Nr. 6, 80–83
- [Wittgenstein 2004] WITTGENSTEIN, Ludwig: *Edition Suhrkamp*. Bd. 12: *Tractatus logico-philosophicus: Logisch-philosophische Abhandlung*. 1. Aufl., [Nachdr.]. Frankfurt am Main : Suhrkamp, 2004
- [Yager 1988] YAGER, R. R.: On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decision Making. In: *IEEE Trans. on Systems, Man, and Cybernetics* 18 (1988), Nr. 1, 183–190
- [Yang et al. 2001] YANG, Jun ; ZHUANG, Yueting ; LI, Qing: Multi-modality Retrieval for Multimedia Digital Libraries: Issues, Architecture, and Mechanisms. In: *Proc. 7th International Workshop on Multimedia Information Systems*, 2001, 81–88

- [Yee et al. 2003] YEE, Ka-Ping ; SWEARINGEN, Kirsten ; LI, Kevin ; HEARST, Marti: Faceted Metadata for Image Search and Browsing. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2003 (CHI '03), 401–408
- [Zadeh 1965] ZADEH, Lotfi A.: Fuzzy Sets. In: *Information and Control* (1965), Nr. 8, 338–353
- [Zadeh 1988] ZADEH, Lotfi A.: Fuzzy Logic. In: *IEEE Computer* 21 (1988), Nr. 4, 83–93
- [Zadeh 2005] ZADEH, Lotfi A.: Toward a Generalized Theory of Uncertainty (GTU): An Outline. In: *Inf. Sci. Inf. Comput. Sci.* 172 (2005), Nr. 1-2, 1–40
- [Zech 2008] ZECH, Sebastian: *Evaluierung des LIRE-Frameworks zur MPEG-7-basierten Feature-Extraktion (German); Bachelor's thesis*. Brandenburg University of Technology, Cottbus, Diplomarbeit, 2008
- [Zellhöfer 2010a] ZELLHÖFER, David: Eliciting Inductive User Preferences for Multimedia Information Retrieval. In: BALKE, Wolf-Tilo (Ed.) ; LOFI, Christoph (Ed.): *Proceedings of the 22nd Workshop "Grundlagen von Datenbanken 2010"* Bd. 581, 2010
- [Zellhöfer 2010b] ZELLHÖFER, David: Inductive User Preference Manipulation for Multimedia Retrieval. In: BÖSZÖRMENYI, Laszlo (Ed.) ; BURDESCU, Dumitru (Ed.) ; DAVIES, Philip (Ed.) ; NEWELL, David (Ed.): *Proc. of the Second International Conference on Advances in Multimedia (MMEDIA)*, IEEE, 2010, 90–95
- [Zellhöfer 2012a] ZELLHÖFER, David: A Permeable Expert Search Strategy Approach to Multimodal Retrieval. In: KAMPS, Jaap (Ed.) ; KRAAIJ, Wessel (Ed.) ; FUHR, Norbert (Ed.): *Proceedings of the 4th Information Interaction in Context Symposium*, ACM, 2012 (IIIX '12), 62–71
- [Zellhöfer 2012b] ZELLHÖFER, David ; BRANDENBURG UNIVERSITY OF TECHNOLOGY (Ed.): *An Evaluation of the Principle of Polyrepresentation in Multimodal and Content-based Image Retrieval*. Cottbus, 2012 (Computer Science Reports 8)
- [Zellhöfer 2012c] ZELLHÖFER, David: An Extensible Personal Photograph Collection for Graded Relevance Assessments and User Simulation. In: IP, Horace H. S. (Ed.) ; RUI, Yong (Ed.): *ICMR '12: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ACM, 2012 (ICMR '12), 29:1–29:8
- [Zellhöfer 2012d] ZELLHÖFER, David ; BRANDENBURG UNIVERSITY OF TECHNOLOGY (Ed.): *On the Usability of PythiaSearch*. Cottbus, 2012 (Computer Science Reports 9)
- [Zellhöfer 2012e] ZELLHÖFER, David: Overview of the Personal Photo Retrieval Pilot Task at ImageCLEF 2012. In: FORNER, Pamela (Ed.) ; KARLGREN, Jussi (Ed.) ; WOMSER-HACKER, Christa (Ed.): *CLEF 2012 Evaluation Labs and Workshop*, 2012
- [Zellhöfer 2012f] ZELLHÖFER, David: Personas - The Missing Link between User Simulations and User-Centered Design?: Linking the Persona-based Design of Adaptive

## Bibliography

- Multimedia Retrieval Systems with User Simulations. In: *10th International Workshop on Adaptive Multimedia Retrieval*. 2012, to appear
- [Zellhöfer 2013] ZELLHÖFER, David: Overview of the ImageCLEF 2013 Personal Photo Retrieval Subtask. In: *CLEF 2013 Labs and Workshop, Notebook Papers, 23-26 September 2013, Valencia, Spain*. 2013
- [Zellhöfer et al. 2012] ZELLHÖFER, David ; BERTRAM, Maria ; BÖTTCHER, Thomas ; SCHMIDT, Christoph ; TILLMANN, Claudius ; SCHMITT, Ingo: PythiaSearch – A Multiple Search Strategy-supportive Multimedia Retrieval System. In: *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, ACM, 2012 (ICMR '12)*, 59:1–59:2
- [Zellhöfer & Böttcher 2011] ZELLHÖFER, David ; BÖTTCHER, Thomas: BTU DBIS' Multimodal Wikipedia Retrieval Runs at ImageCLEF 2011. In: PETRAS, Vivien (Ed.) ; FORNER, Pamela (Ed.) ; CLOUGH, Paul D. (Ed.): *CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands*. 2011
- [Zellhöfer et al. 2013] ZELLHÖFER, David ; BÖTTCHER, Thomas ; BERTRAM, Maria ; SCHMIDT, Christoph ; TILLMANN, Claudius ; UHLIG, Markus ; ZIERENBERG, Marcel ; SCHMITT, Ingo: PythiaSearch - Interaktives, Multimodales Multimedia-Retrieval. In: MARKL, Volker (Ed.) ; SAAKE, Gunter (Ed.) ; SATTLER, Kai-Uwe (Ed.) ; HACKENBROICH, Gregor (Ed.) ; MITSCHANG, Bernhard (Ed.) ; HÄRDER, Theo (Ed.) ; KÖPPEN, Veit (Ed.): *Datenbanksysteme für Business, Technologie und Web (BTW), 15. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), 11.-15.3.2013 in Magdeburg, Germany. Proceedings Bd. 214. GI, 2013, 495–498*
- [Zellhöfer et al. 2011] ZELLHÖFER, David ; FROMMHOLZ, Ingo ; SCHMITT, Ingo ; LALMAS, Mounia ; VAN RIJSBERGEN, Cornelis J.: Towards Quantum-Based DB+IR Processing Based on the Principle of Polyrepresentation. In: CLOUGH, P. (Ed.) ; FOLEY, C. (Ed.) ; GURRIN, C. (Ed.) ; JONES, G. (Ed.) ; KRAAIJ, W. (Ed.) ; LEE, H. (Ed.) ; MURDOCH, V. (Ed.): *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings Bd. 6611. Springer, 2011, 729–732*
- [Zellhöfer & Lehrack 2008] ZELLHÖFER, David ; LEHRACK, Sebastian: Nutzerzentriertes maschinenbasiertes Lernen von Gewichten beim Multimedia-Retrieval. In: HÖPFNER, Hagen (Ed.) ; KLAN, Friederike (Ed.): *Proceedings of the 20. GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), Apolda, Thüringen, Germany, May 13-16, 2008 Bd. 01/2008. School of Information Technology, International University in Germany, 2008, 51–55*
- [Zellhöfer & Schmitt 2010a] ZELLHÖFER, David ; SCHMITT, Ingo: A Poset Based Approach for Condition Weighting. In: *Proceedings of the 6th International Conference on Adaptive Multimedia Retrieval: Identifying, Summarizing, and Recommending Image and Music. Berlin, Heidelberg : Springer-Verlag, 2010 (AMR'08), 28–39*

- [Zellhöfer & Schmitt 2010b] ZELLHÖFER, David ; SCHMITT, Ingo: A Preference-based Approach for Interactive Weight Learning: Learning Weights within a Logic-Based Query Language. In: *Distributed and Parallel Databases 27* (2010), Nr. 1, 31-51
- [Zellhöfer & Schmitt 2010c] ZELLHÖFER, David ; SCHMITT, Ingo: Ein Polyrepräsentatives Anfrageverfahren für das Multimedia Retrieval. In: ATZMÜLLER, M. (Ed.) ; BENZ, D. (Ed.) ; HOTH, A. (Ed.) ; STUMME, G. (Ed.): *Proceedings of LWA2010 - Workshop-Woche: Lernen, Wissen & Adaptivität*, 2010
- [Zellhöfer & Schmitt 2011a] ZELLHÖFER, David ; SCHMITT, Ingo: A User Interaction Model based on the Principle of Polyrepresentation. In: *Proceedings of the 4th Workshop on Workshop for Ph.D. Students in Information & Knowledge Management*. New York, NY, USA : ACM, 2011 (PIKM '11), 3-10
- [Zellhöfer & Schmitt 2011b] ZELLHÖFER, David ; SCHMITT, Ingo: Approaching Multimedia Retrieval from a Polyrepresentative Perspective. In: DETYNIĘCKI, Marcin (Ed.) ; KNEES, Peter (Ed.) ; NÜRNBERGER, Andreas (Ed.) ; SCHEDL, Markus (Ed.) ; STÖBER, Sebastian (Ed.): *Adaptive Multimedia Retrieval. Context, Exploration, and Fusion - 8th International Workshop, AMR 2010, Linz, Austria, August 17-18, 2010, Revised Selected Papers* Bd. 6817. Springer, 2011, 46-60
- [Zhang et al. 2000] ZHANG, Dengsheng ; WONG, Aylwin ; INDRAWAN, Maria ; LU, Guojun: Content-based Image Retrieval Using Gabor Texture Features. In: *IEEE Transactions PAMI*, 2000, 13-15
- [Zhang 2008] ZHANG, Jin: *Springer-11645 / Dig. Serial*. Bd. 23: *Visualization for Information Retrieval*. Berlin, Heidelberg : Springer-Verlag Berlin Heidelberg, 2008
- [Zhao et al. 2003] ZHAO, W. ; CHELLAPPA, R. ; PHILLIPS, P. J. ; ROSENFELD, A.: Face Recognition: A Literature Survey. In: *ACM Comput. Surv.* 35 (2003), Nr. 4, 399-458
- [Zimmermann 1996] ZIMMERMANN, H.-J.: *Fuzzy Set Theory - And Its Applications*. 3. Norwell, MA, USA : Kluwer Academic Publishers, 1996
- [Zloof 1975] ZLOOF, M. M.: Query By Example. In: *Proc. of AFIPS National Computer Conference* Bd. 44, AFIPS Press, 1975, 431-438

# List of Figures

2.1	Schematic illustration of system-centric IR . . . . .	15
2.2	The semantic gap in CBIR . . . . .	27
2.3	A Balinese demon statue . . . . .	29
3.1	Venn diagram of a sample cognitive overlap in MIR . . . . .	36
3.2	The polyrepresentation continuum . . . . .	38
3.3	Browsing in the ostensive model . . . . .	41
3.4	Browsing with the starting point “Eiffel tower” in Google’s image swirl . . . . .	42
3.5	Faceted navigation in the Flamenco fine arts IR system . . . . .	43
4.1	The cluster hypothesis versus multi-clustered information needs . . . . .	60
4.2	Predefined queries in commercial image search engines for the query term “1960s” . . . . .	62
4.3	A pure IN in a IN/term space . . . . .	69
4.4	Venn-like diagram of different document representations forming a penetrable cognitive overlap . . . . .	75
4.5	The extended polyrepresentation continuum . . . . .	77
5.1	Skyline of $r_1$ with $DIFF = Type$ and $MIN = Avg.Price$ . . . . .	93
5.2	Extended Hasse diagram of a Chomicki’s preference example I . . . . .	97
5.3	Different understanding of edges in an image . . . . .	101
5.4	Interactive preference information flow within PrefCQQL . . . . .	106
5.5	Creation of a preference . . . . .	108
5.6	Confirmation of a preference . . . . .	109
5.7	Inversion of a preference . . . . .	109
5.8	Removal of a preference . . . . .	109
5.9	Indifferent preference . . . . .	109
5.10	Hasse diagram (sample) . . . . .	110
5.11	Preferences as a generalization of traditional relevance feedback . . . . .	111
5.12	Preference utilities . . . . .	113
5.13	Cases of utility of the preference $d_1 \succeq d_2$ with respect to the zero level . . . . .	113
5.14	Implication and overlap of preferences . . . . .	114
5.15	Simplex redefinition strategies . . . . .	117
5.16	Simplified flow diagram of the downhill simplex algorithm . . . . .	117
5.17	Runtime of the learning algorithm . . . . .	120
5.18	Preference conflict . . . . .	121
5.19	Conflicting preference poset; arrows point to the preferred element . . . . .	122

5.20	Resolved preference poset; arrows point to the preferred element, bi-directed edges indicate indifference . . . . .	122
6.1	Persona-based user-centered system design in relation to user simulations . . . . .	135
6.2	Early GUI for the Pythia MIR system . . . . .	138
6.3	The global model of polyrepresentation . . . . .	150
6.4	Matrix of the four intrinsic information need extreme cases . . . . .	151
6.5	Polyrepresentative user interaction model . . . . .	153
6.6	Thematical groups of user stories . . . . .	159
6.7	Apple Aperture’s central GUI elements . . . . .	162
6.8	GUI mockup of the Pythia MIR system . . . . .	163
6.9	The Polaroid and Post-it note interface metaphor . . . . .	168
6.10	Self-organizing map visualization . . . . .	171
6.11	Search bar . . . . .	173
6.12	Mockup of CQQL-based polythetic clustering . . . . .	176
6.13	Concentric circles for preference elicitation . . . . .	180
6.14	Preference elicitation per drag and drop and visual relevance feedback	182
6.15	Linear search history widget . . . . .	184
6.16	Branching search history . . . . .	184
6.17	Polyrepresentation visualization . . . . .	186
6.18	Weight inspector . . . . .	187
6.19	Central system components of the Pythia MIR system . . . . .	189
6.20	Main GUI elements of the Pythia MIR system . . . . .	193
6.21	Screenshot of the browsing functionality in the Pythia MIR system . .	194
6.22	The uInteract interface . . . . .	196
7.1	Schematic illustration of the Cranfield laboratory setting . . . . .	203
8.1	Samples from Caltech 101’s “stegosaurus” topic . . . . .	214
8.2	Contribution to the Pythia collection per photographer . . . . .	218
8.3	Job types of the assessors . . . . .	220
8.4	Position of a persona-based user simulation in the relevance feedback loop . . . . .	222
8.5	Performance comparison of single representations, part 1 . . . . .	230
8.6	Performance comparison of single representations, part 2 . . . . .	231
8.7	Performance comparison of single representations, part 3 . . . . .	232
8.8	Performance comparison of representation fusions using different NULL value policies, part 1 . . . . .	238
8.9	Performance comparison of representation fusions using different NULL value policies, part 2 . . . . .	239
8.10	Sample distribution analysis of nDCG@20 values . . . . .	240
8.11	Box plot of the ranks obtained by basic representation and combined matching functions over all collections . . . . .	242



## List of Figures

8.12	Performance comparison of basic combinations of representations, pt. 1	244
8.13	Performance comparison of basic combinations of representations, pt. 2	245
8.14	Performance comparison of representation combinations and standard aggregations, part 1	248
8.15	Performance comparison of representation combinations and standard aggregations, part 2	249
8.16	Box plot of the ranks obtained by different matching functions over all collections	253
8.17	RF performance comparison of characteristic matching functions, part 1	258
8.18	RF performance comparison of characteristic matching functions, part 2	259
8.19	Performance comparison of RF-enabled representation combinations and standard aggregations, part 1	260
8.20	Performance comparison of RF-enabled representation combinations and standard aggregations, part 2	261
8.21	Relevance feedback effectiveness development trends	263
8.22	Effects of long-enduring relevance feedback	271
8.23	Boxplot of the mean weights values over all matching functions	273
8.24	Boxplot of the standard deviation of the weights values for different matching functions	274
8.25	Hypothetic weight development analysis	276
8.26	Correlation of different measures with the weighting scheme distance between RF iterations	279
8.27	Correlation of query incompatibilities and average preference poset size	280
8.28	Ratio between query incompatibility percentage and number of specified preferences	281
8.29	Bubble chart of query incompatibilities	283
9.1	Initial QBE document of the usability test	286
9.2	Sample target images of the usability test	287
9.3	Available widgets in GUI variant T2	289
9.4	Similarity search feature in Google's image search	289
9.5	Available widgets in GUI variant T3	290
9.6	Available widgets in GUI variant T4	291
9.7	Daily computer usage and job type of the subjects	293
9.8	Year of birth and expertise of CBIR of the subjects	294
9.9	Usage of other CBIR systems	294
9.10	Ranking of the tested GUI variants	307
9.11	General utility of the different UI components	308
9.12	Online help information channels in preference rings widget	310
9.13	Frequency of the transitions between search strategies and visualizations per GUI variant	312
9.14	Result personalization window of the Pythia MIR system	316

10.1	Performance comparison of RF-enabled representation combinations and standard aggregations, part 1 . . . . .	331
A.1	Hasse diagram of a lattice with the ordering criterion “subset of” . . . . .	338
B.1	Transformation of CQQL queries . . . . .	347
B.2	RF performance comparison of characteristic matching functions and standard aggregations, part 1 . . . . .	353
B.3	RF performance comparison of characteristic matching functions and standard aggregations, part 2 . . . . .	354
B.4	RF performance comparison of conjunction, disjunction and weighted arithmetic mean, part 1 . . . . .	355
B.5	RF performance comparison of conjunction, disjunction and weighted arithmetic mean, part 2 . . . . .	356
B.6	RF performance comparison of characteristic matching functions (conjunctions, disjunctions, and Q10), part 1 . . . . .	357
B.7	RF performance comparison of characteristic matching functions (conjunctions, disjunctions, and Q10), part 2 . . . . .	358
B.8	RF performance comparison of conjunction and Eidenberger variants, part 1 . . . . .	359
B.9	RF performance comparison of conjunction and Eidenberger variants, part 2 . . . . .	360
B.10	RF performance comparison of conjunction and semantic group conjunctions, part 1 . . . . .	361
B.11	RF performance comparison of conjunction and semantic group conjunctions, part 2 . . . . .	362
B.12	Conjunction variants, averaged weight development . . . . .	363
B.13	Disjunction variants, averaged weight development . . . . .	364
B.14	Eidenberger conjunctions, averaged weight development . . . . .	365
B.15	Eidenberger disjunctions, averaged weight development . . . . .	366
B.16	Semantic group conjunctions, averaged weight development . . . . .	367
B.17	Q10 & weighted average, averaged weight development . . . . .	368
B.18	Conjunction, weight development and distribution (Caltech 101) . . . . .	369
B.19	Conjunction, weight development and distribution (Caltech 256) . . . . .	370
B.20	Conjunction, weight development and distribution (MSRA-MM) . . . . .	370
B.21	Conjunction, weight development and distribution (Pythia) . . . . .	371
B.22	Conjunction, weight development and distribution (UCID) . . . . .	371
B.23	Conjunction, weight development and distribution (Wang) . . . . .	372
B.24	Conjunction (best), weight development and distribution (Caltech 101) . . . . .	372
B.25	Conjunction (best), weight development and distribution (Caltech 256) . . . . .	373
B.26	Conjunction (best), weight development and distribution (MSRA-MM) . . . . .	373
B.27	Conjunction (best), weight development and distribution (Pythia) . . . . .	374
B.28	Conjunction (best), weight development and distribution (UCID) . . . . .	374
B.29	Conjunction (best), weight development and distribution (Wang) . . . . .	375

## List of Figures

B.30	Disjunction, weight development and distribution (Caltech 101) . . . . .	375
B.31	Disjunction, weight development and distribution (Caltech 256) . . . . .	376
B.32	Disjunction, weight development and distribution (MSRA-MM) . . . . .	376
B.33	Disjunction, weight development and distribution (Pythia) . . . . .	377
B.34	Disjunction, weight development and distribution (UCID) . . . . .	377
B.35	Disjunction, weight development and distribution (Wang) . . . . .	378
B.36	Disjunction (best), weight development and distribution (Caltech 101)	378
B.37	Disjunction (best), weight development and distribution (Caltech 256)	379
B.38	Disjunction (best), weight development and distribution (MSRA-MM)	379
B.39	Disjunction (best), weight development and distribution (Pythia) . . . .	380
B.40	Disjunction (best), weight development and distribution (UCID) . . . . .	380
B.41	Disjunction (best), weight development and distribution (Wang) . . . . .	381
B.42	Q10, weight development and distribution (Caltech 101) . . . . .	382
B.43	Q10, weight development and distribution (Caltech 256) . . . . .	382
B.44	Q10, weight development and distribution (MSRA-MM) . . . . .	383
B.45	Q10, weight development and distribution (Pythia) . . . . .	383
B.46	Q10, weight development and distribution (UCID) . . . . .	384
B.47	Q10, weight development and distribution (Wang) . . . . .	384
B.48	Eidenberger conjunction 1, weight development & distribution (Caltech 101) . . . . .	385
B.49	Eidenberger conjunction 1, weight development & distribution (Caltech 256) . . . . .	385
B.50	Eidenberger conjunction 1, weight development & distribution (MSRA-MM) . . . . .	386
B.51	Eidenberger conjunction 1, weight development & distribution (Pythia)	386
B.52	Eidenberger conjunction 1, weight development & distribution (UCID)	387
B.53	Eidenberger conjunction 1, weight development & distribution (Wang)	387
B.54	Eidenberger disjunction 1, weight development & distribution (Caltech 101) . . . . .	388
B.55	Eidenberger disjunction 1, weight development & distribution (Caltech 256) . . . . .	388
B.56	Eidenberger disjunction 1, weight development & distribution (MSRA-MM) . . . . .	389
B.57	Eidenberger disjunction 1, weight development & distribution (Pythia)	389
B.58	Eidenberger disjunction 1, weight development & distribution (UCID)	390
B.59	Eidenberger disjunction 1, weight development & distribution (Wang)	390
B.60	Eidenberger conjunction 2, weight development & distribution (Caltech 101) . . . . .	391
B.61	Eidenberger conjunction 2, weight development & distribution (Caltech 256) . . . . .	391
B.62	Eidenberger conjunction 2, weight development & distribution (MSRA-MM) . . . . .	392
B.63	Eidenberger conjunction 2, weight development & distribution (Pythia)	392
B.64	Eidenberger conjunction 2, weight development & distribution (UCID)	393

B.65	Eidenberger conjunction 2, weight development & distribution (Wang)	393
B.66	Eidenberger disjunction 2, weight development & distribution (Caltech 101)	394
B.67	Eidenberger disjunction 2, weight development & distribution (Caltech 256)	394
B.68	Eidenberger disjunction 2, weight development & distribution (MSRA-MM)	395
B.69	Eidenberger disjunction 2, weight development & distribution (Pythia)	395
B.70	Eidenberger disjunction 2, weight development & distribution (UCID)	396
B.71	Eidenberger disjunction 2, weight development & distribution (Wang)	396
B.72	Semantic group conjunction (best), weight development & distribution (Caltech 101)	397
B.73	Semantic group conjunction (best), weight development & distribution (Caltech 256)	397
B.74	Semantic group conjunction (best), weight development & distribution (MSRA-MM)	398
B.75	Semantic group conjunction (best), weight development & distribution (Pythia)	398
B.76	Semantic group conjunction (best), weight development & distribution (UCID)	399
B.77	Semantic group conjunction (best), weight development & distribution (Wang)	399
B.78	Semantic group conjunction, weight development & distribution (Caltech 101)	400
B.79	Semantic group conjunction, weight development & distribution (Caltech 256)	400
B.80	Semantic group conjunction, weight development & distribution (MSRA-MM)	401
B.81	Semantic group conjunction, weight develop. & distribution (Pythia)	401
B.82	Semantic group conjunction, weight develop. & distribution (UCID)	402
B.83	Semantic group conjunction, weight develop. & distribution (Wang)	402
B.84	Q10, weight development and distribution (Caltech 101)	403
B.85	Q10, weight development and distribution (Caltech 256)	403
B.86	Q10, weight development and distribution (MSRA-MM)	404
B.87	Q10, weight development and distribution (Pythia)	404
B.88	Q10, weight development and distribution (UCID)	405
B.89	Q10, weight development and distribution (Wang)	405
B.90	Correlation of Gini coefficient and nDCG at 10	406
B.91	Correlation of Gini coefficient and nDCG at 10	407
B.92	Correlation of Gini coefficient and weight distance betw. RF iterations	408
B.93	Correlation of Gini coefficient and weight distance betw. RF iterations	409
B.94	Correlation of Gini coefficient and nDCG at 10	410
B.95	Correlation of Gini coefficient and nDCG at 10	411
B.96	Histogram of grand random values	446

## List of Figures

C.1	Usability test instructions . . . . .	453
D.1	Initial result set . . . . .	456
D.2	Result set after location facet "Same as current" has been chosen . . . .	456
D.3	Preferences at iteration #1 . . . . .	457
D.4	Weighting characteristics of the query after iteration #1 . . . . .	457
D.5	Result set after iteration #1 . . . . .	457
D.6	Preferences at iteration #2 . . . . .	458
D.7	Weighting characteristics of the query after iteration #2 . . . . .	458
D.8	Result set after iteration #2 . . . . .	458
E.1	Table of contents of the DVD . . . . .	460



# List of Tables

2.1	Available Features and Origin . . . . .	25
3.1	Query by example as a DB query approach . . . . .	39
4.1	Related concepts from database querying and quantum mechanics . . . . .	50
4.2	Impact of Weights in CQQL . . . . .	55
4.3	Related concepts from QIR, CQQL, and quantum mechanics . . . . .	69
4.4	Related concepts from MQRM, CQQL, and quantum mechanics . . . . .	70
4.5	Related concepts from CQQL and the principle of polyrepresentation . . . . .	74
5.1	Comparison of runtime and retrieval effectiveness of the two winning groups at the ImageCLEF 2013 Personal Photo Retrieval subtask . . . . .	84
5.2	Sample instance $r_1$ of relation <i>Restaurant</i> ( <i>Type, Name, Avg.Price</i> ) . . . . .	90
5.3	Dominant tuples of the sample instance $r_1$ of relation <i>Restaurant</i> . . . . .	92
5.4	Sample instance $r_1$ of relation <i>Restaurant</i> ( <i>Type, Name, Avg.Price</i> ) ordered by their utility . . . . .	95
5.5	Sample instance $r_1$ of relation <i>Book</i> . . . . .	96
5.6	Comparison of the mathematical properties of preferences and posets . . . . .	107
6.1	Multi-layer interface elements of the Pythia MIR system . . . . .	165
6.2	Lines of code of the Pythia MIR system . . . . .	188
6.3	Central namespaces of the Pythia MIR system . . . . .	190
6.4	Main GUI elements of the Pythia MIR system (Legend) . . . . .	192
8.1	Characteristics of the Cranfield-based test collections . . . . .	216
8.2	Motif duplicates and their type . . . . .	219
8.3	Topics (visual concepts) of the Pythia collection . . . . .	221
8.4	Event distribution in the Pythia collection . . . . .	221
8.5	Characteristics of the Pythia test sets . . . . .	224
8.6	Overview over the examined test collections . . . . .	226
8.7	Excerpt from the examined matching functions . . . . .	227
8.8	Processing duration of the analyzed collections . . . . .	228
8.9	Best performing representations and rank of color histogram . . . . .	233
8.10	Worst performing representations . . . . .	234
8.11	Sample rank, sorted by arithmetic average . . . . .	236
8.12	Sample rank, sorted by arithmetic average with NULL ignorance . . . . .	237
8.13	Missing representations per collection . . . . .	237

8.14	Results of test for normality . . . . .	240
8.15	Results of the Wilcoxon signed-rank test for baseline matching functions	243
8.16	Matching functions and their characteristics, ordered by mean rank . .	250
8.17	Results of the Wilcoxon signed-rank test between arithmetic mean and conjunction . . . . .	251
8.18	Matching functions characteristics during relevance feedback . . . . .	264
8.19	Wilcoxon signed-rank test $p$ -values of differences between RF iterations' nDCG@20 of various matching functions . . . . .	267
8.20	Wilcoxon signed-rank test $p$ -values of the differences between the RF iterations' nDCG@20 values for the weighted arithmetic mean and the conjunction (best features variants) . . . . .	268
8.21	Wilcoxon signed-rank test $p$ -value results between long-enduring RF it- erations based on their nDCG@20 values . . . . .	272
8.22	Summary statistics of the weights values over all RF iterations and col- lections for different matching functions . . . . .	273
9.1	Functional comparison of the GUI variants . . . . .	292
9.2	Study and work field of the subjects . . . . .	293
9.3	Interaction duration with the GUI variants . . . . .	296
9.4	General usability of the GUI variants . . . . .	298
9.5	General user criticism . . . . .	309
9.6	Legend to Figure 9.13 . . . . .	313
A.1	Semantics of Boolean logical connectors and propositions <b>A</b> and <b>B</b> . . .	339
A.2	Truth table for the material implication and propositions <b>A</b> and <b>B</b> . . . .	342
B.1	Available Features and Origin . . . . .	346
B.2	Suitability for the task (T2) . . . . .	414
B.3	Self-descriptiveness (T2) . . . . .	416
B.4	Controllability (T2) . . . . .	417
B.5	Conformity with user expectations (T2) . . . . .	419
B.6	Error tolerance (T2) . . . . .	420
B.7	Suitability for individualization (T2) . . . . .	422
B.8	Suitability for learning (T2) . . . . .	423
B.9	Suitability for the task (T3) . . . . .	425
B.10	Self-descriptiveness (T3) . . . . .	426
B.11	Controllability (T3) . . . . .	428
B.12	Conformity with user expectations (T3) . . . . .	429
B.13	Error tolerance (T3) . . . . .	431
B.14	Suitability for individualization (T3) . . . . .	432
B.15	Suitability for learning (T3) . . . . .	434
B.16	Suitability for the task (T4) . . . . .	435
B.17	Self-descriptiveness (T4) . . . . .	437
B.18	Controllability (T4) . . . . .	438



## List of Tables

B.19	Conformity with user expectations (T4)	440
B.20	Error tolerance (T4)	441
B.21	Suitability for individualization (T4)	443
B.22	Suitability for learning (T4)	444
C.1	Demographics of the Contributors	449
C.2	Demographics of the Assessors	450



# List of Definitions

2.1	Medium . . . . .	11
2.2	Multimedia . . . . .	11
2.3	Modality . . . . .	12
2.4	Information need . . . . .	13
2.5	Query . . . . .	13
2.6	Query formulation problem . . . . .	13
2.7	Query representation . . . . .	13
2.8	Semantic gap . . . . .	13
2.9	Document . . . . .	14
2.10	Document representation . . . . .	14
2.11	Document storage . . . . .	14
2.12	Index vocabulary . . . . .	14
2.13	Relevance . . . . .	14
2.14	Matching . . . . .	14
2.15	Matching function . . . . .	14
2.16	Relevance feedback . . . . .	14
2.17	Fuzzy set . . . . .	17
2.18	Probability of relevance (POR) . . . . .	19
2.19	Uncertain inference . . . . .	20
2.20	Content-based image retrieval . . . . .	22
2.21	Multimodal retrieval . . . . .	23
2.22	Global low-level feature . . . . .	24
2.23	Local low-level feature . . . . .	24
2.24	Minkowski distance . . . . .	26
3.1	Directed search . . . . .	38
3.2	Exploratory search . . . . .	40
4.1	Hilbert space . . . . .	49
4.2	State vector . . . . .	49
4.3	Projector . . . . .	49
4.4	Quantum measurement . . . . .	49
4.5	Tensor product . . . . .	49
4.6	Meet (lattice operator) . . . . .	50
4.7	Join (lattice operator) . . . . .	50
4.8	Orthocomplement (lattice operator) . . . . .	51

4.9	Commuting projectors . . . . .	51
4.10	Basic CQQL elements . . . . .	52
4.11	CQQL conditions . . . . .	52
4.12	CQQL evaluation basics . . . . .	53
4.13	CQQL conjunction (evaluation) . . . . .	53
4.14	CQQL disjunction (evaluation) . . . . .	53
4.15	CQQL negation (evaluation) . . . . .	54
4.16	CQQL atomic condition (evaluation) . . . . .	54
4.17	CQQL weighted conjunction . . . . .	55
4.18	CQQL weighted disjunction . . . . .	55
4.19	Cluster hypothesis . . . . .	60
4.20	Gleason's theorem . . . . .	65
5.1	Strict preference . . . . .	85
5.2	Indifferent preference . . . . .	85
5.3	Weak preference . . . . .	85
5.4	Anti-symmetry (Preference) . . . . .	85
5.5	Symmetry of indifference (Preference) . . . . .	85
5.6	Reflexivity of indifference (Preference) . . . . .	85
5.7	Incompatibility of preference & indifference . . . . .	85
5.8	Transitivity (Preference) . . . . .	85
5.9	Pareto optimality . . . . .	89
5.10	Ceteris paribus . . . . .	89
5.11	Preference formula . . . . .	89
5.12	Winnow operator . . . . .	90
5.13	Utility function . . . . .	94
5.14	Incompletene preferences . . . . .	103
5.15	Core concepts in the CQQL-based retrieval model PrefCQQL . . . . .	105
5.16	Creation . . . . .	108
5.17	Confirmation . . . . .	108
5.18	Inversion . . . . .	108
5.19	Removal . . . . .	109
5.20	Indifference . . . . .	109
5.21	Weighting scheme $\omega$ . . . . .	111
5.22	Evaluation function <i>eval()</i> . . . . .	111
5.23	Preference fulfillment . . . . .	112
5.24	Preference utility . . . . .	112
5.25	Preference utility categories . . . . .	112
6.1	Usability principles . . . . .	143
6.2	8 golden rules of interface design . . . . .	144
7.1	Document collection . . . . .	203
7.2	Topics . . . . .	203

## List of Definitions

7.3	Relevance assessments . . . . .	203
8.1	Precision . . . . .	208
8.2	Recall . . . . .	208
8.3	Average precision . . . . .	208
8.4	Mean average precision . . . . .	209
8.5	Precision at n . . . . .	210
8.6	Discounted Cumulated Gain . . . . .	210
8.7	Ideal Gain Vector . . . . .	211
8.8	Normalized Discounted Cumulated Gain . . . . .	211
8.9	Effectiveness stability . . . . .	250
A.1	Partially ordered set (Poset) . . . . .	337
A.2	Anti-symmetry . . . . .	337
A.3	Reflexivity . . . . .	337
A.4	Transitivity . . . . .	337
A.5	Strict partially ordered set . . . . .	337
A.6	Total order . . . . .	337
A.7	Hasse diagram . . . . .	338
A.8	Axiom of Associativity . . . . .	339
A.9	Axiom of Commutativity . . . . .	339
A.10	Axiom of Distributivity . . . . .	339
A.11	Axiom of Identity . . . . .	339
A.12	Axiom of Complements . . . . .	340
A.13	Axiom of Idempotence . . . . .	340
A.14	Ket vector . . . . .	340
A.15	Bra vector . . . . .	340
A.16	Inner product in bra-ket form . . . . .	340
A.17	Norm of a ket vector . . . . .	340
A.18	Outer product in bra-ket form . . . . .	340



# List of Theorems, Lemmata, and Proofs

4.1	Lemma . . . . .	52
4.1	Theorem (CQQL is a Boolean algebra) . . . . .	53
4.1	Proof (CQQL is a Boolean algebra) . . . . .	53





# List of Matching Functions

1	Conjunction . . . . .	346
2	Conjunction, best features . . . . .	347
3	Disjunction . . . . .	347
4	Disjunction, best features . . . . .	347
5	Eidenberger conjunction, variant 1 . . . . .	348
6	Eidenberger conjunction, variant 2 . . . . .	348
7	Eidenberger disjunction, variant 1 . . . . .	348
8	Eidenberger disjunction, variant 2 . . . . .	348
9	Q10 . . . . .	348
10	Semantic group conjunction . . . . .	348
11	Semantic group conjunction, best features . . . . .	349
12	Bielefeld conjunction . . . . .	349
13	Arithmetic mean . . . . .	349
14	Arithmetic mean, best features . . . . .	350
15	Weighted arithmetic mean, best features . . . . .	350
16	Geometric mean . . . . .	350
17	Geometric mean, best features . . . . .	350
18	Max . . . . .	350
19	Max, best features . . . . .	351
20	Min . . . . .	351
21	Min, best features . . . . .	351



# List of User Stories

1	Fast and intuitive adaptivity . . . . .	140
2	Controllable relevance feedback . . . . .	141
3	Fuzzy filters . . . . .	141
4	Get an overview . . . . .	141
5	Diversity of the results . . . . .	141
6	Avoid empty result sets . . . . .	141
7	Exploit all representations . . . . .	141
8	Similarity search . . . . .	141
9	Dynamic information needs . . . . .	141
10	Dynamic search strategies . . . . .	141
11	Exploratory visualization . . . . .	141
12	Similarity grouping . . . . .	141
13	Sorting support . . . . .	141
14	Ease of use . . . . .	141
15	Refindability of documents . . . . .	142
16	Visualize the unusual . . . . .	142
17	Multimodal query specification . . . . .	142
18	Weight representations . . . . .	142
19	Weight visualization . . . . .	142
20	Direct weight manipulation . . . . .	142



# Listings

3.1 Sample SQL query . . . . .	39
5.1 Sample preference query in the Lacroix-Lavency SQL dialect . . . . .	88
5.2 Preference utility calculation using the SumMin strategy . . . . .	118

Ceterum censeo Carthaginem esse delendam.