Automatische Indexierung von deutschsprachigen bibliographischen Metadaten im Bereich Bauwesen

Dimitri Busch

Fraunhofer-Informationszentrum

Raum und Bau IRB, Stuttgart

Email: dimitri.busch@irb.fraunhofer.de

Inhalt

1 Einführung

Im Fraunhofer-Informationszentrum Raum und Bau (IRB) wird Fachliteratur im Bereich Planen und Bauen bibliografisch erschlossen. Die bibliografischen Metadaten (Dokumente) werden u.a. bei der Produktion der bibliografischen Datenbank RSWB®plus verwendet.

Zu den Dokumenten werden Deskriptoren von einer Nomenklatur (Schlagwortliste IRB) und Notationen von einem Klassifikationssystem (Fachgliederung IRB) zugeordnet. In der **Abbildung 1** ist ein Dokument dargestellt, das einen Konferenzbeitrag beschreibt.

Momentan wird die Indexierung intellektuell durchgeführt. Die Intellektuelle Indexierung ist zeitaufwendig und teuer.

In dem Beitrag geht es um ein automatisiertes Indexierungssystem, das entwickelt wurde, um o.g. Probleme zu lösen.

Originaltitel	Praxisbeispiel "Umbau Hauptbahnhof Hannover" - Anwendung der BIM-Methodik
Autor	Horstmann, Wolfgang
Referat	Im Hauptbahnhof Hannover sind sechs Bahnsteige mit den Bahnsteigdächern, Beleuchtungsanlagen, Aufzügen, Fahrtreppen, Fahrgastinformationsanlagen und den Bahnsteigausstattungen zu erneuern, fünf Eisenbahnbrücken sind zu ersetzen. Ein Gewölbe sowie die Entwässerungsleitungen des Bahnhofes sollen erneuert bzw. teilerneuert werden. Im Projekt wird die BIM-Methodik angewendet…
Schlagwörter	Hauptbahnhof; Personenverkehr; Verkehrsanlage; Bahnsteig; Eisenbahnbrücke; Umbau; Planungsmethode; Erneuerung; Building Information Modeling…
Fachgebiet	31.160 Ingenieurhochbau: Verkehrsanlage; 36.010 Bahnbau: Allgemein
Publikationstyp	Buchkapitel; Konferenzbeitrag
Quelle	Vorträge zum Deutschen Bautechnik-Tag am 27. und 28. April 2017 in Stuttgart, Seiten 109-110

Abbildung 1: Beispiel-Dokument

2 Kontrollierte Vokabularien

2.1 Schlagwortliste IRB

Die Schlagwortliste IRB ist eine Nomenklatur zu den Themen Bauwesen, Raumordnung, Städtebau, Wohnungswesen. Sie wurde vom Fraunhofer-Informationszentrum Raum und Bau (IRB) erstellt. Die Nomenklatur enthält ca. 42400 Terme (Stand 06.2018). Die Nomenklatur unterstützt keine Beziehungen zwischen den Termen und ist zweisprachig: Deutsch und Englisch.

2.2 Fachgliederung IRB

Die Fachgliederung IRB ist ein Klassifikationssystem zu den Themen Bauwesen, Raumordnung, Städtebau, Wohnungswesen. Sie wurde vom Fraunhofer-Informationszentrum Raum und Bau (IRB) erstellt, und enthält ca. 1000 Notationen (Stand 06.2018). Das Klassifikationssystem enthält 2 Hierarchieebenen, Fachbereiche und Fachgebiete und ist zweisprachig: Deutsch und Englisch.

2.3 FINDEX-Thesauri

FINDEX BAU und FINDEX RAUM sind facettierte Thesauri zu den Themen Bauwesen und Raumordnung. Die Thesauri wurden vom Fraunhofer-Informationszentrum Raum und Bau (IRB) erstellt, und enthalten insgesamt ca. 7500 Terme. Alle Terme der FINDEX-Thesauri sind auch in der Schlagwortliste IRB enthalten. Die Thesauri unterstützen u.a. Äquivalenz-, und Hierarchie-Beziehungen und sind zweisprachig: Deutsch und Englisch. Obwohl die FINDEX-Thesauri zu Legacy-Ressourcen gehören und nicht mehr aktiv geführt werden, sind sie immer noch von Bedeutung, da die Äquivalenz-Beziehungen (Synonymie) bei der automatischen Indexierung berücksichtigt werden.

3 Vorarbeiten im Fraunhofer IRB

Das Fraunhofer IRB hat bereits ein System für die halbautomatische Indexierung englischsprachiger Dokumente mit Deskriptoren von der Schlagwortliste IRB entwickelt. Bei der Entwicklung des Systems wurde eine freie Indexierungssoftware JEX verwendet, die eine Java-

Programmierschnittstelle unterstützt. Das Indexierungssystem wird momentan hauptsächlich für die bibliografische Erschließung der Literatur verwendet, die von CIB (International Council for Building) publiziert wird. Die daraus resultierenden Metadaten werden in den Datenbanken ICONDA®Bibliographic und ICONDA®CIB Library verwendet.

4 Ein System für automatische Indexierung

Das Fraunhofer IRB hat eine Weiterentwicklung des bestehenden englischsprachigen Indexierungssystems durchgeführt, um Besonderheiten Deutscher Sprache zu berücksichtigen, und die Indexierungsqualität zu verbessern. Das System verwendet ein lineares profilbasiertes Verfahren und informationslinguistische Verfahren für die Indexierung mit Deskriptoren und ein instanzbasiertes Verfahren für die Indexierung mit Notationen.

Im Folgenden wird das System genauer betrachtet.

4.1 Indexierung mit Deskriptoren

4.1.1 Lineares profilbasiertes Verfahren

Für die automatische Indexierung mit Deskriptoren wird ein lineares profilbasiertes Verfahren verwendet. In dem profilbasierten Modell wird jeder Deskriptor C_i durch einen Vektor $\overrightarrow{C_i} = (w_{i1}, w_{i2}, \cdots, w_{it})$ repräsentiert, der zu einem $|t|$-dimensionalen Vektorraum gehört, wobei w_{ik} das Gewicht des Terms k in Bezug auf den Deskriptor C_i bezeichnet. Dieser Vektor wird als Profil benannt. Die **Tabelle 1** zeigt eine vereinfachte Darstellung eines Beispielprofils für den Deskriptor „Baustil“. Jeder Eintrag des Profils enthält einen Term und dessen Gewicht.

Tabelle 1: Profil für den Deskriptor „Baustil“

Term	Gewicht
Gotik	0.263
Barock	0.255
Klassizismus	0.234
Historismus	0.187
Jugendstil	0.185
…	…

Jedes Dokument D_j, das zu indexieren ist, wird auch durch einen Vektor $\overrightarrow{D_j} = (w_{j1}, w_{j2}, \cdots, w_{ij})$ repräsentiert, wobei w_{jk} das Gewicht des Terms k in dem Dokument D_j bezeichnet. Bei der Indexierung eines Dokuments werden zuerst Ähnlichkeiten zwischen Deskriptoren und dem Dokument berechnet, und dann zu dem Dokument Deskriptoren zugeordnet, die dem Dokument am ähnlichsten sind. Die Ähnlichkeit zwischen dem Deskriptor C_i und dem Dokument D_j wird als Skalarprodukt $\sum_{k=1}^{t}(w_{ik}w_{jk})$ von Vektoren $\overrightarrow{C_i}$ und $\overrightarrow{D_j}$ berechnet.

Die Deskriptoren-Profile werden aus Trainingsdokumenten erzeugt. Die Trainingsdokumente sind Dokumente, die bereits indexiert sind. Um Profile zu erzeugen, wird die parametrische Rocchio-Methode verwendet [vgl. Basili/Moschitti 2005]. Nach dieser Methode wird jedes Element w_{ik} eines Profils nach folgender Formel berechnet:

$$W_{ik} = max\left\{0, \frac{1}{|T_i|}\sum_{t\in T_i} w_{tk} - p_i \frac{1}{|N_i|}\sum_{t\in N_i} w_{tk}\right\} \quad (1)$$

In der o.g. Formel bezeichnet w_{tk} das Gewicht des Terms k in dem Trainingsdokument D_t . Ein solches Gewicht wird als Produkt $TF_{jk} \times IDF_k$ der Frequenz TF_{jk} des Terms k in dem Dokument D_j und der inversen Frequenz IDF_k des Terms k in der Gesamtmenge von Trainingsdokumenten berechnet. Die Teilformel $\frac{1}{|T_i|}\sum_{t\in P_i} w_{tk}$ bezeichnet das Durchschnittsgewicht des Terms k in der Menge T_i von Trainingsdokumenten (positiven Beispielen), zu denen der Deskriptor C_i zugeordnet wurde. Die Teilformel $\frac{1}{|N_i|}\sum_{t\in P_i} w_{tk}$ bezeichnet das Durchschnittsgewicht des Terms k in der Menge N_i von Trainingsdokumenten (negativen Beispielen), zu denen der Deskriptor C_i nicht zugeordnet wurde. Der Parameter p_i ist eine Zahl, die die Wirkung von negativen Beispielen auf Term-Gewichten im Profil für den Deskriptor $C_i$regelt. Der Parameter p_i wird separat für jeden Deskriptor automatisch ermittelt und wird dabei so optimiert, dass die Qualität der Indexierung maximiert wird. Die Methode zur Ermittlung des Parameters wird genauer von Basili/Moschitti [2005, S. 89 ff.] beschrieben.

4.1.2 Informationslinguistische Indexierungsverfahren

Informationslinguistische Verfahren ermitteln Indexterme auf Basis sprachlicher Gesetzmäßigkeiten. Zu Informationslinguistischen Verfahren gehören u.a.:

- Lemmatisierung (Grundformreduktion);
- Erkennung und Zerlegung von Komposita;
- Mehrwortgruppenerkennung;
- Ermittlung von Synonymen

Beispiele:

- Lemmatisierung: **Schlösser -> Schloss**
- Zerlegung eines Kompositums: **Residenzschloss -> Residenz + Schloss**

- Ermittlung eines Synonyms: **Wärmeisolierung -> Wärmedämmung**

Informationslinguistische Verfahren sind genauer von Gödert/Lepsky/Nagelschmidt [2012, S. 260 ff.] beschrieben.

4.1.3 Ablauf der Indexierung

Die **Abbildung 2** zeigt den Ablauf der Indexierung mit Deskriptoren. Ein zu indexierendes Dokument wird zuerst informationslinguistischer Indexierung unterzogen, die Lemmatisierung, Erkennung von Komposita, Mehrwortgruppen und Synonymen und ein abschließender Abgleich mit der Schlagwortliste IRB beinhaltet. Die informationslinguistische Indexierung ergibt eine Liste von Deskriptoren. Dann wird eine profilbasierte Indexierung durchgeführt, wobei Grundformen, Kompositateile und Mehrwortgruppen, die bei der informationslinguistischen Indexierung ermittelt wurden, berücksichtigt werden. Die Ergebnisse der informationslinguistischen und profilbasierten Indexierung werden schließlich vereinigt, sodass eine endgültige Liste von Deskriptoren erzeugt wird. Nach der Vereinigung werden zuerst Deskriptoren ausgegeben, die sowohl von profilbasierten als auch von informationslinguistischen Verfahren erzeugt wurden. Dann wird eine bestimmte Anzahl von Deskriptoren ausgegeben, die nur profilbasiert ermittelt wurden. Diese Zahl wird empirisch ermittelt, und beträgt momentan bis 3 Deskriptoren. Dann werden Deskriptoren ausgegeben, die nur informationslinguistisch ermittelt wurden.

Die **Abbildung 3** zeigt ein Beispiel der Vereinigung von Deskriptoren, die nach unterschiedlichen Methoden ermittelt wurden. Zuerst werden Deskriptoren „Schloss“ und „Architektur“ angezeigt, die sowohl Informationslinguistisch als auch profilbasiert ermittelt wurden. Dann folgen die Deskriptoren „Baustil“ und „Baugeschichte“, die nur profilbasiert ermittelt wurden. Schließlich folgen die Deskriptoren „Residenz“ und „Beispiel“, die nur informationslinguistisch ermittelt wurden.

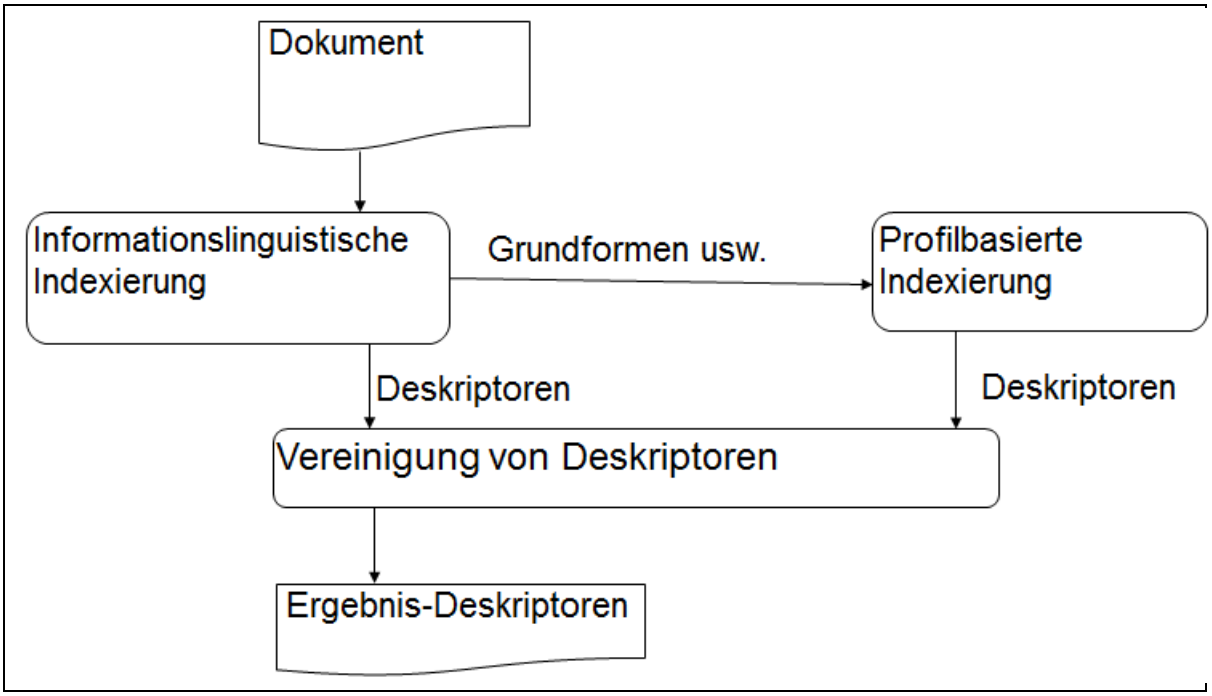

Abbildung 2: Ablauf der Indexierung mit Deskriptoren

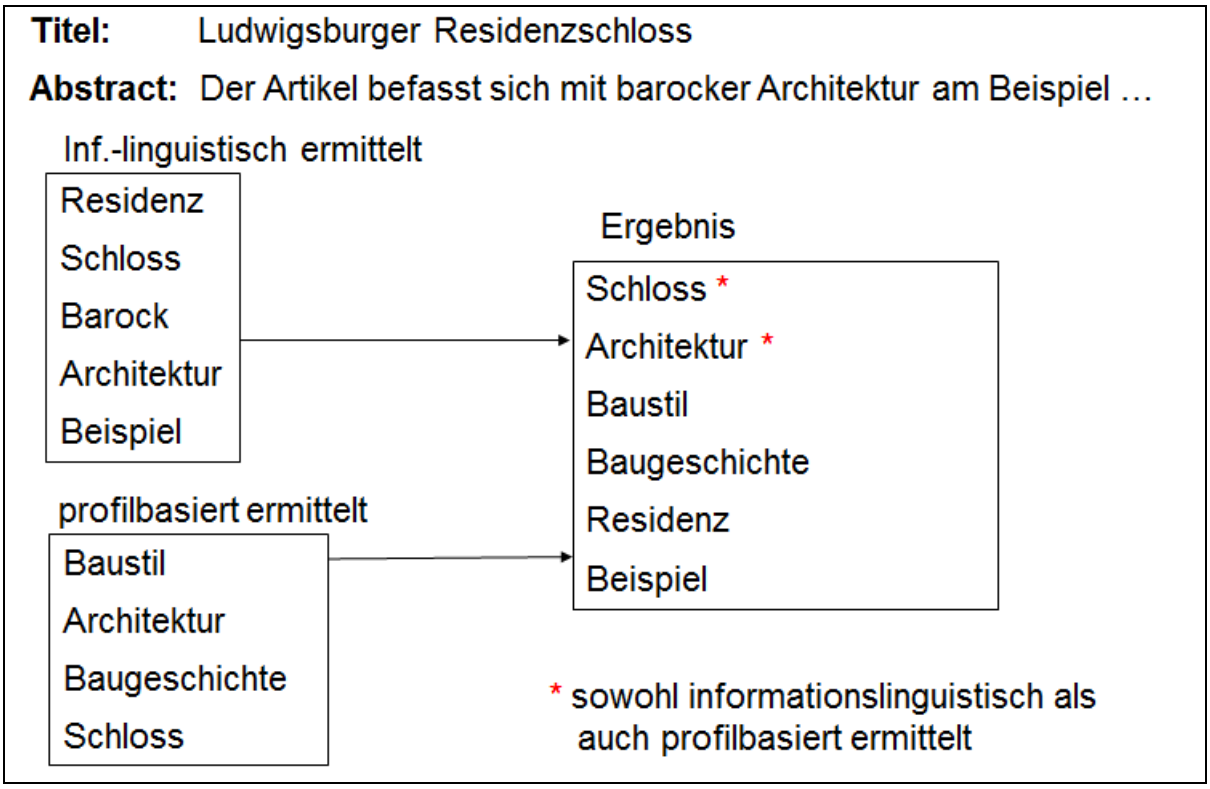

Abbildung 3: Beispiel der Vereinigung von Deskriptoren

Die **Abbildung 4** zeigt ein Beispiel der Indexierung eines Dokuments mit Deskriptoren. In dem rechten Teil der Abbildung kann man Deskriptoren sehen, die von dem Indexierungsprogramm vorgeschlagen wurden.

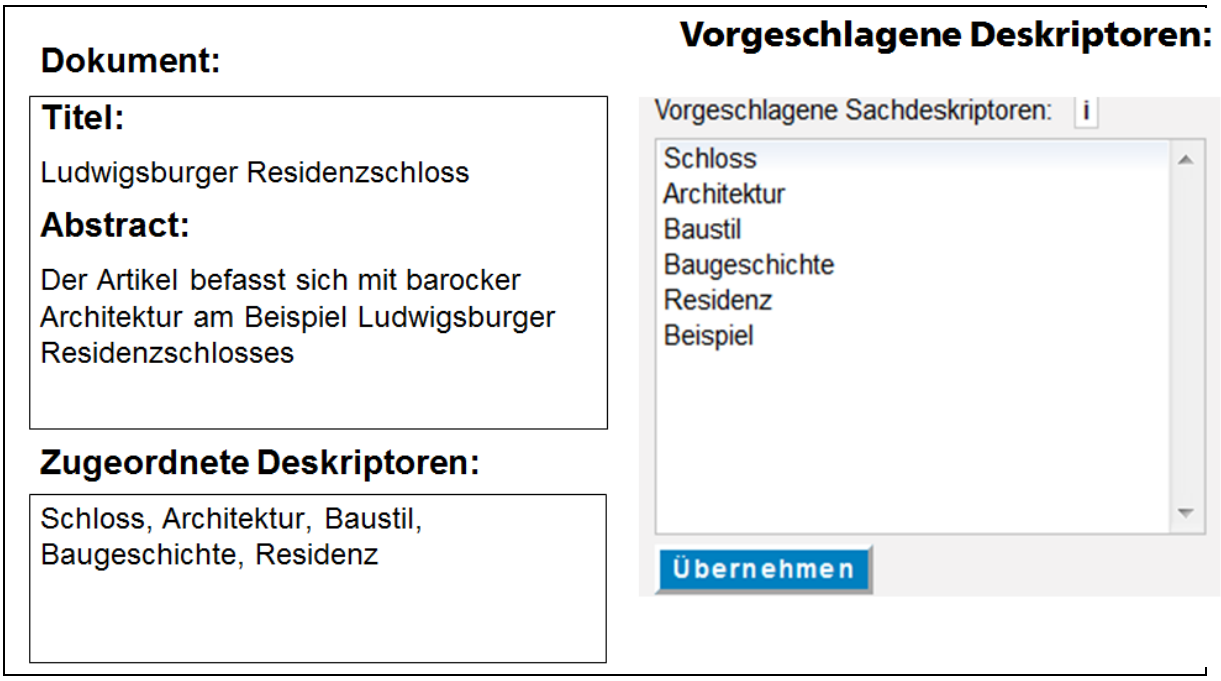

Abbildung 4: Beispiel der Indexierung eines Dokuments mit Deskriptoren

4.2 Indexierung mit Notationen

Die automatische Indexierung mit Notationen der Fachgliederung IRB erfolgt nach dem instanzbasierten Verfahren k-NN (k-Nearest Neighbors). Das Verfahren orientiert sich an den Notationen, die jene Trainingsdokumente aufweisen, die dem zu indexierenden Dokument am ähnlichsten sind. Um ein Dokument zu indexieren, wird zuerst eine bestimmte Anzahl (k) der Trainingsdokumente gefunden, die dem zu indexierenden Dokument am ähnlichsten sind. Diese Zahl wird empirisch ermittelt, und beträgt momentan 21 Trainingsdokumente. Dann werden für das Dokument eine oder mehrere Notationen vorgeschlagen, die in den gefundenen Trainingsdokumenten am häufigsten vorkommen. Das k-NN-Verfahren ist genauer von Manning/Raghavan/Schütze [vgl. 2008, S. 273 ff.] beschrieben.

Die Abbildung 5 zeigt rechts Fachbereiche/Fachgebiete, die für das Dokument von der Abbildung 4 vorgeschlagen wurden.

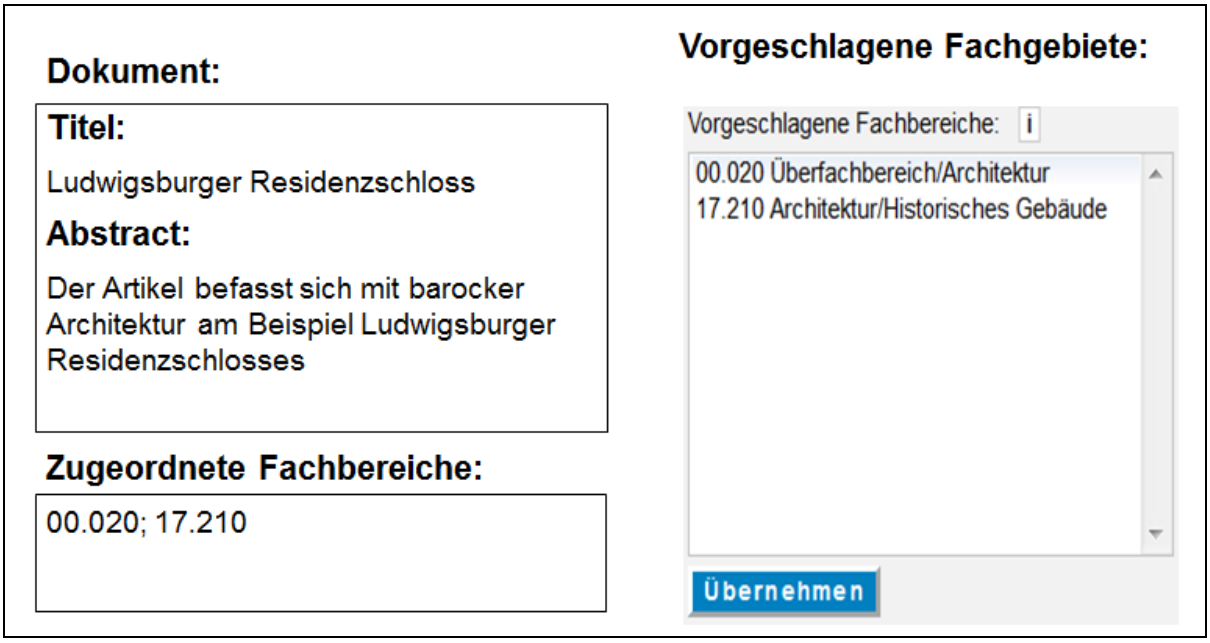

Abbildung 5: Beispiel der Indexierung eines Dokuments mit Notationen

4.3 Anpassung an thematische Domänen

Um die Indexierungsqualität zu verbessern können Indexierungsprozeduren an bestimmte thematische Domänen angepasst werden, z.B. Betonbau, Gebäudetechnik. Eine Anpassungsmethode besteht in der Verwendung von Metadaten über einschlägige Zeitschriften für das Trainieren von Indexierungsprozeduren. Eine andere Anpassungsmethode besteht in dem Aufbau von domäne-spezifischen Wörtebüchern. Die domäne-spezifischen Begriffe können z.B. aus entsprechenden Facetten von FINDEX-Thesauri (s. Abs. 2.3) ausgewählt werden.

4.4 Details zur Implementierung

Die Programme für profilbasierte und instanzbasierte Indexierung wurden mit der Programmiersprache Java implementiert. Für die instanzbasierte Indexierung wurde auch eine freie Retrievalsoftware LUCENE verwendet. Für die informationslinguistische Indexierung wird eine freie Indexierungsoftware LINGO verwendet, die eine Ruby-Umgebung benötigt [Gödert/Lepsky/Nagelschmidt 2012, S. 269 ff.].

Die automatische Ermittlung von Indextermen (Deskriptoren und Notationen) erfolgt im Batch-Modus. Die Zuordnung der Indexterme wird von menschlichen Indexierern über eine webbasierte Oberfläche durchgeführt, die mit JSF implementiert wurde.

5 Zusammenfassung

In dem Beitrag wird ein System für automatische Indexierung von deutschsprachigen Dokumenten mit Deskriptoren und Notationen beschrieben. Für die Indexierung mit Deskriptoren werden profilbasierte und informationslinguistische Verfahren verwendet. Die Indexierung mit Notationen erfolgt nach einem instanzbasierten Verfahren.

Der vorgestellte Einsatz eignet sich insbesondere für halbautomatische Indexierung, bei der ein menschlicher Indexierer eine endgültige Entscheidung über die Zuordnung von Deskriptoren und Notationen trifft.

Nach entsprechenden Anpassungen wird es voraussichtlich möglich, auch eine vollautomatische Indexierung für bestimmte thematische Domänen durchzuführen

6 Danksagung

Ich danke meinen Kollegen, Herrn Dr. Friedmar Voormann und Herrn Hans-Martin Barth, die mir stets Ansprechpartner waren, für ihre Unterstützung bei der Durchführung dieses Projektes

7 Literatur

- Basili, Roberto ; Moschitti, Alessandro [2005]. Automatic text categorization. From information retrieval to support vector learning. Rom: Arachne

- Fraunhofer-Informationszentrum Raum und Bau, Hrsg. [1985]: FINDEX Bau: Facettenartiges Indexierungssystem für das Bauwesen, 2. Auflage. Stuttgart: IRB Verlag
- Fraunhofer-Informationszentrum Raum und Bau, Hrsg. FINDEX Raum [1985]: Facettenartiges Indexierungssystem für Raumordnung, Städtebau, Wohnungswesen, 1. Auflage. Stuttgart: IRB Verlag
- Gödert, Winfried; Lepsky, Klaus; Nagelschmidt, Matthias [2012]: Informationserschließung und automatisches Indexieren. Berlin, Heidelberg: Springer
- Manning, Christopher; Raghavan, Prabhakar; Schütze, Hinrich [2008]: Introduction to information retrieval. Cambridge, New York: Cambridge University Press: 2008
- Steinberger, Ralf; Ebrahim, Mohamed; Turchi, Marco [2012]: JRC EuroVoc Indexer JEX– A freely available multi-label categorisation tool. In Proceedings of the 8th international conference on language resources and evaluation (LREC'2012), Istambul, 21-27. Mai 2012, S.798-805