Katholische Universität Eichstätt-Ingolstadt
Wirtschaftswissenschaftliche Fakultät
Lehrstuhl für ABWL und Wirtschaftsinformatik
Prof. Dr. Klaus D. Wilde

Erfolgsmessung Informationsorientierter Websites

Doktorarbeit zur Erlangung des Grades eines
Doktors der Wirtschaftswissenschaften (Dr. rer. pol.) an der
Wirtschaftswissenschaftlichen Fakultät
der Katholischen Universität Eichstätt-Ingolstadt

Dipl. Kfm. Carsten Dirk Stolz
Holzstr. 23
80469 München
Deutschland
carsten.stolz@ku-eichstaett.de

Eingereicht von:	Dipl. Kfm. Carsten Dirk Stolz
Betreuer und Erstkorrektor:	Prof. Dr. Klaus D. Wilde
Zweitkorrektor:	Prof. Dr. Ulrich Küsters
Prüfer:	Prof. Dr. Marco Wilkens
Prüfer:	Prof. Dr. Anton Burger
Tag der mündlichen Prüfung:	26. Oktober 2007

Zusammenfassung

Das Internet erlaubt dem Menschen als aktives Medium den Zugriff, das Speichern, Modifizieren, Verlinken, Suchen und Filtern fast beliebiger Informationen in digitaler Form. Unternehmen haben das Internet als Möglichkeit erkannt, mit Kunden, Mitarbeitern, Zulieferern und Interessenten zu kommunizieren und zu interagieren. Vor dem Hintergrund der wachsenden Verlagerung von Unternehmensdarstellungen, Kommunikation, Marketing und Vertrieb auf das Internet erlangt die Analyse und Optimierung der Web Auftritte eine immer größere Bedeutung. Es wird ein Überblick der Internet-Geschäftsmodelle gegeben und daraus die Möglichkeiten zur Messung des Erfolgs von Unternehmenswebsites abgeleitet.

Die Messung des wirtschaftlichen Erfolgs von transaktionsorientierten Websites ist gut erforscht und findet breite Anwendung. Dahingegen ist eine Aussage über den Erfolg von Websites, die lediglich Informationen anbieten, nur sehr eingeschränkt möglich. Bei transaktionsorientierten Websites wird deren Erfolg durch abgeschlossene bzw. abgebrochene Transaktionen bestimmt. In einer Transaktion gibt der User und Kunde ein direktes Feedback über den durch die Website gestifteten Nutzen. Der Erfolg lässt sich als monetäre Größe direkt ermitteln. Diese Möglichkeit fehlt informationsorientierten Websites.

Das Ziel dieser Dissertation besteht darin, Wege zur Erfolgsmessung von informationsorientierten Websites zu untersuchen, um daraus einen Indikator für den Erfolg dieser Art von Websites abzuleiten.

Nach dem Überblick der Geschäftsmodelle werden die zur Verfügung stehenden Daten beschrieben. Dabei wird auf die Unterschiede zwischen einer client- und serverseitigen Erfassung der Daten und der daraus resultierenden Einschränkungen in der Interpretierbarkeit eingegangen. Die verfügbaren Daten werden nach ihrer Quelle in Usage-, Content und Structure-Daten unterteilt.

Die bereits bekannten Maße und Web Metriken werden vorgestellt und auf ihre Eignung als Erfolgsmaß für informationsorientierte Websites hin untersucht. In einer Systematik der bekannten Web Maße und Metriken wird deutlich, dass ein Mangel an Möglichkeiten besteht, den Erfolg informationsorientierter Websites zu messen.

Es werden Web Mining-Analysen vorgestellt, die ein tieferes Verständnis von Inhalt, Struktur, Benutzung und Benutzern einer Website ermöglichen. Bestehende Web Mining-Analysen alleine sind ebenfalls nicht geeignet, den Erfolg informationsorientierter Websites zu messen. Allerdings stellen sie eine hervorragenden Grundlage dar, um daraus ein Erfolgsmaß ableiten zu können.

Mit Hilfe geeigneter Web Mining-Analysen wird ein Verfahren entwickelt, das die Zielerreichung einer Website aus Sicht des Website-Besitzers beurteilt. Jede einzelne User-Aktion auf der analysierten Website wird dahingehend bewertet, ob sie zum Website-Ziel der Informationsverbreitung beiträgt. Es sollen nur solche Daten verwendet werden, die auf der Serverseite

verfügbar sind. Für die beobachtbaren User-Aktionen werden drei Teilmaße gebildet: Effektivitätstyp und -gewicht, sowie Effizienzgewicht.

Der Effektivitätstyp bewertet die Grundtendenz einer User-Aktion, die aus den Übergangen zwischen Navigations- und Zielseiten ermittelt wird. Darin liegt die Annahme zugrunde, daß Navigationsseiten den User nur unterstützen sollen, die eigentlichen Zielseiten zu erreichen. Das Effektivitätsgewicht analysiert den Inhalt der aufgesuchten Webpages und vergleicht diesen mit dem Inhalt der kompletten User Session. Um diese inhaltliche Bewertung zu ermöglichen ist eine umfangreiche Text Mining-Analyse der Website notwendig. Das Effizienzmaß nutzt die Aufenthaltsdauer auf einer Webpage und interpretiert sie als die Zeit, die der User zur Verfügung hatte, um den Inhalt der Webpage wahrnehmen zu können.

Aus den drei Teilmaßen wird für jeden Click ein Gesamtmaß erzeugt, der Guidance Performance Indicator. Da lediglich serverseitige Informationen vorliegen, spiegelt der GPI die Perspektive des Website-Autors wider. Dieses Maß beurteilt die Fähigkeit einer Website, einen User zu Informationen zu führen. Der GPI kann von der Bewertung eines einzelnen Clicks auf Ebene von Session oder Webpages aggregiert werden. Das neu erstellte Erfolgsmaß wird in einer Anwendungsstudie auf drei Websites eines großen Unternehmens angewandt und evaluiert.

Um die bei der Erstellung des GPI getroffenen Annahmen richtig einordnen zu können, wird ein formales Modell einer Website und ihrer Benutzung erstellt. Darin wird zwischen syntaktischer Ebene, semantischer Ebene und Maßebene unterschieden. Es wird gezeigt, wie das formale Modell geeignet ist, um daraus die erstellten Maße nachzubilden und neue Zusammenhänge zu erkennen.

Um den serverseitigen GPI mit einem unabhängigen Instrument empirisch zu überprüfen, wird eine clientseitige Userstudie durchgeführt. Darin wird die Website-Perspektive mit der User-Perspektive verglichen. Die meisten User bestätigen die Bewertung des GPI. Einige Fälle weisen jedoch Abweichungen auf.

Um diese Fälle erklären zu können, wird die Einbeziehung der User-Perspektive in ein Erfolgsmaß für informationsorientierte Websites vorgeschlagen.

Dies wird durch die bislang in der Literatur nicht verwendeten Referrer-Informationen erreicht, die in einigen Fällen die Suchanfrage von Usern an Suchmaschinen enthalten. Mit Hilfe dieser Information wird es möglich, auf der Serverseite einen Einblick in die Absichten des Users auf der Clientseite zu erhalten. Die bereits verwendete inhaltliche Analyse der Webpages wird mit der Analyse der Suchstrings kombiniert. Anstelle einer manuellen Festlegung zwischen Navigations- und Zielseiten, wird bei diesem Vorgehen für jeden User individuell ein Ranking der Webpages in inhaltlicher Abhängigkeit zum verwendeten Suchstring berechnet.

Mit Hilfe des formalen Modells wird eine Funktion höherer Ordnung erstellt, die die user-individuelle Rankingfunktion beschreibt. Dadurch können Effektivitätstyp und -gewicht bei der GPI-Berechnung ersetzt werden. Es wird ein neues Maß, der intention based GPI (ibGPI)

erstellt, der die User-Perspektive berücksichtigt und die Abweichungen der Userstudie erklären kann. Aus diesem Maß wird zusätzlich ein Erfolgsmaß für eine User Session als Ganzes berechnet, der Session Success Indicator (SSI).

Mit den drei Maßen GPI, ibGPI und SSI steht nun Erfolgskennzahlen für informationsorientierte Websites zur Verfügung. Aus deren paralleler Anwendung und dem Vergleich der drei Kennzahlen lässt sich ein umfangreiches Bild der Benutzung einer Website und deren Erfolg in der Informationsvermittlung sowohl aus Perspektive der Website-Autoren wie der User darstellen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Vorwort

Diese Arbeit ist während meiner Tätigkeit als Doktorand bei Corporate Technology IC4 und IC1 bei der Siemens AG in München und als Wissenschaftlicher Mitarbeiter am Lehrstuhl für Wirtschaftsinformatik an der Universität Eichstätt entstanden.

Bei meiner Arbeit als Doktorand bei Siemens unterstützte mich Herr Dr. Ralph Neuneier und Herr Michal SkubaczBei beiden möchte ich mich für ihre wissenschatliche Förderung und das ausgezeichnete Arbeitsumfeld bedanken, das beide geschaffen haben. Ihre Arbeit war für mich Vorbild und Ansporn zu dieser Dissertation und Keim bei der Themenfindung.

Außerdem möchte ich mich bei meinem Ko-Doktoranden bei Corporate Technology, Maximilian Viermetz, für die angenehme und produktive Zusammenarbeit bei vielen Publikationen, Ausarbeitung gemeinsamer Ideen bedanken und die langen Nächte für die Fertigstellung unserer Veröffentlichungen hier erwähnen. Herrn Malte Buck gilt mein Dank, daß er jederzeit ein Ohr und eine Lösung für meine technischen Fragen hatte. Beiden möchte für ihre ständige Bereitschaft zu fachlichen Diskussionen danken.

Herrn Prof. Dr. Klaus D. Wilde danke ich sehr für die Betreuung der Dissertation und die freundliche Aufnahme als Mitarbeiter an seinem Lehrstuhl. Durch seine Unterstützung und Anregungen wurde diese Arbeit ermöglicht. Herrn Prof. Dr. Ulrich Küsters gilt mein Dank für seine freundliche Bereitschaft zur Übernahme des Zweitgutachtens sowie seine Bereitschaft zu fachlichen Diskussionen während der Dissertation.

Ohne die spannenden fachlichen Gespräche und zahllosen Kaffees mit Herrn Dr. Michael Barth wäre diese Arbeit in dieser Form nicht entstanden. Seinem Vorbild folgte ich bei meinen ersten wissenschaftlichen Publikationen, denen viele gemeinsame nachfolgten und hoffentlich noch weitere entstehen. Für seine Geduld bei der gemeinsamen Suche nach neuen Ansätzen und der nächtelangen Ausarbeitung von Modellen und Formulierungen kann ich ihm gar nicht genug danken. Der Wert von den dadurch gewonnen neuen Einsichten ist für mich nicht zu bemessen.

Diese Arbeit widme ich meiner Familie, insbesonere meinen Eltern, die mich immer wieder aufs neue motiviert und unterstützt haben. Ihrem Rückhalt und ihrer Geduld möchte ich mit der fertiggestellten Dissertation besonders danken. Ihre unermüdliche moralische Aufbauarbeit hat mir den notwendigen familiären Rückhalt zur Durchführung dieser Arbeit gegeben. Für ihre ausdauernde Unterstützung und zeitlichen Einsatz bei der Drucklegung dieser Arbeit möchte ich mich bei meiner Tante Dorothee Burkhardt und meiner Mutter herzlich bedanken.

München, 15.Juni 2008 Carsten Dirk Stolz

1 Einleitung

Die traditionellen Medien wie Zeitungen, Photographie oder Radio sind rein passive Medien. Das Internet dagegen erlaubt dem Menschen als aktives Medium den Zugriff, das Speichern, Modifizieren, Verlinken, Suchen und Filtern fast beliebiger Informationen in digitaler Form [12]. Auch Unternehmen haben das Internet bzw. das World Wide Web [25, 26] als Möglichkeit erkannt, mit Kunden, Mitarbeitern, Zulieferern und Interessenten zu kommunizieren und zu interagieren.

Sowohl unternehmensinterne als auch -externe Prozesse finden vermehrt über dieses Medium statt. Vor dem Hintergrund der wachsenden Verlagerung von Unternehmensdarstellungen, Kommunikation, Marketing und Vertrieb auf das Internet erlangt lt. Wilde et al. [124] die Analyse der Internet-Aktivitäten eine immer größere Bedeutung. Nach der verklungenen Euphorie der New Economy überprüfen Unternehmen ihre Internet Aktivitäten auf deren Wirtschaftlichkeit. Die Frage nach dem ökonomischen Nutzen ist noch immer Gegenstand wissenschaftlicher Untersuchungen. So stellt auch im Jahr 2006 Grob et al.[108] die Frage nach dem wirtschaftlichen Erfolg von Internetauftritten von Unternehmen.

Dies verlangt nach Instrumenten zur Steuerung, Qualitätskontrolle und stetigen Verbesserung. Eine Website soll die Ziele des Unternehmens erreichen helfen, ohne die Interessen der User außer acht zu lassen.

Der User steht auf der einen Seite einer unübersehbaren Vielfalt an Informationsmöglichkeiten im Internet gegenüber. Auf die gleichen Probleme stößt auf der anderen Seite der Betreiber einer Website, wenn er herauszufinden versucht, was die User auf seinen Webpages machen und welche Ziele sie dabei verfolgen. Erst durch dieses Verständnis kann die Website an die User-Bedürfnisse angepasst werden.

Dabei ist man mit einer Reihe von Herausforderungen konfrontiert. Eine dieser Herausforderungen stellen unstrukturierte Daten in großer Menge, unterschiedliche Daten- und Medienformate, gepaart mit einer enormen Zahl an User-Aktivitäten, auf den Webseiten eines Unternehmens dar. Diese müssen täglich analysiert werden, um eine fortlaufende Erfolgskontrolle zu ermöglichen. Die Meßbarkeit von Erfolg und Qualität einer Website ist die Voraussetzung für einen kontinuierlichen Verbesserungsprozeß, der die Gestaltung und Inhalte einer Website an die User Bedürfnisse anpaßt.

"Miß alles, was sich messen läßt, und mach alles meßbar, was sich nicht messen läßt." (Galileo Galilei, 1564-1643).

Wie Galilei vorschlägt, untersucht dieser Arbeit zuerst, was sich in Bezug auf den Erfolg einer Website bereits messen läßt.

Die Fälle, in denen der Website-Erfolg nicht meßbar ist, bilden den Ausgangspunkt für diese Dissertation.

Daher beschäftigt sich das Kapitel in den nachfolgenden Abschnitten zunächst damit, die Geschäfts- und Erlösmodelle von Unternehmenswebsites zu beschreiben. Sind die Ziele einer Website bekannt, kann untersucht werden, wie die Zielerreichung gemessen werden kann.

1.1 Einordnung des Themengebiets

1.1.1 E-Business

Der Wandel zur Informationsgesellschaft wird in seiner Tragweite lt. Meier et al. [169, S.2] oft mit der industriellen Revolution verglichen. Der fortlaufende Prozeß der schöpferischen Zerstörung im Sinne von Schumpeter [211] wird lt. Gasos et al. [103] erneut durch das Internet eingeleitet. Durch die Zerstörung von alten Strukturen werden die Produktionsfaktoren neu geordnet. Die Zerstörung sei notwendig, damit Neuordnung stattfinden könne. Als eine derartige schöpferische Zerstörung kann die Integration von Informationssystemen, in diesem Falle des Internets bzw. Intranets, entlang der gesamten Wertschöpfungskette eines Unternehmens, angesehen werden [219, 169]. Der Faktor Information gewinnt heute gegenüber den Produktionsfaktoren Arbeit und Kapital an Bedeutung [12].

Die informationstechnologische Integration der verschiedenen Wertschöpfungsprozesse untereinander wird nach [108, 219, 169] unter dem Begriff **E-Business** zusammengefaßt. Meier et al. [169, S.2] definieren E-Business wie folgt:
"Electronic Business bedeutet Anbahnung, Vereinbarung und Abwicklung elektronischer Geschäftsprozesse, d.h. Leistungsaustausch mit Hilfe öffentlicher oder privater Kommunikationsnetze, respektive Internet, zur Erzielung einer Wertschöpfung".
Skiera et al. grenzen in [219, S.1] den Teil des E-Business, der auf internetbasierenden Informationstechnologien beruht, als *Internetökonomie* ab.

Abbildung 1.1 zeigt, daß sich E-Business sowohl innerhalb als auch über die Unternehmensgrenzen hinaus erstreckt [212]. Abbildung 1.2 von Meier et al. [169] konzentriert sich auf die Darstellung der am E-Business beteiligten Prozesse und Aufgaben innerhalb eines Unternehmens.

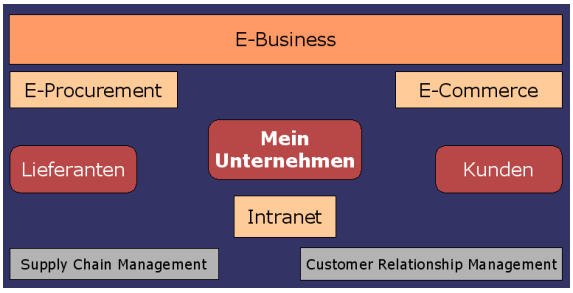


Abbildung 1.1: E-Business [230, S.54]

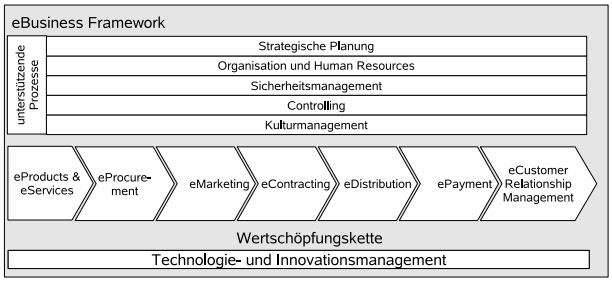


Abbildung 1.2: E-Business nach [169, S.1]

Die Wertschöpfungsmöglichkeiten von E-Business-Technologien, wie Kostenreduktion, Verbesserung der Kundengewinnung und -bindung, Komplexitätsbeherrschung und Nutzung innovativer Organisationsstrukturen, beschreiben Müller et al. [177]. Sie weisen darauf hin, daß Unternehmenswertsteigerungen nicht eindeutig E-Business-Investitionen zugewiesen werden können. Damit ist eine erste Herausforderung genannt, mit der bei der Erfolgsmessung von E-Business-Aktivitäten umgegangen werden muß. Einen Vorschlag, wie mit dieser Problematik als Ganzes umgegangen werden kann, liefern Jonen et al. [140]. Sie schlagen den Einsatz einer Balanced Scorecard vor, die zwar ein Steuerungsinstrument darstellt, aber keine Antwort auf die Meßbarkeit des Erfolges von Internet-Aktivitäten eines Unternehmens bereithält.

Einen unternehmensweiten Ansatz, wie den von Jonen et al. durch eine Balanced Scorecard, versucht diese Arbeit nicht. Es sollen nicht alle Prozeßschritte untersucht werden. Vielmehr soll die Schnittstelle zwischen Unternehmen und Kunden im World Wide Web, sprich die Internet Seiten eines Unternehmens, untersucht werden. Dazu gehören die E-Commerce-Aktivitäten und das Customer Relationship Management (CRM), wie in Abb. 1.1 zu sehen ist. Im E-Business Schema von Meier et al. in Abb. 1.2 beschränkt sich dies auf folgende Teile der Wertschöpfungskette: *eProducts & eServices*, *eMarketing* und *eCustomer Relationship Management*.

1.1.2 E-Commerce

Während E-Business alle geschäftlichen Transaktionen einbezieht, verstehen Meier et al. [169, S.2] unter E-Commerce diejenigen Transaktionen bzw. Geschäftsbeziehungen, bei denen ein Unternehmen Leistungen an Dritte im Rahmen eines elektronischen Handels anbietet. In diesen Fällen tritt das Unternehmen als Leistungsanbieter auf. Gemäß Abb. 1.3 zählen hierzu B2C, B2B und B2A. B2C beschreibt das klassische Endkundengeschäft, bei dem ein Unternehmen dem Endverbraucher auf einer Website seine Produkte oder Dienstleistungen anbietet. Die Phasen des E-Commerce können lt. Stahlknecht et al. [231, S.406f] je nach Geschäftsmo-

		Leistungsnachfrager		
		Consumer	**Business**	**Administration**
Leistungsanbieter	**Consumer**	**Consumer-to-Consumer (C2C)** z.B.Kleinanzeigen auf einer persönlichen Homepage	**Consumer-to-Business (C2B)** z.B. Website mit persönlichem Fähigkeitsprofil zur Bewerbung	**Consumer-to-Administration (C2A)** z.B. Bürger bewertet öffentliches Bauprojekt
	Business	**Business-to-Consumer (B2C)** z.B. Produkte in einem Online-Shop	**Business-to-Business (B2B)** z.B. Bestellung auf Website des Lieferanten	**Business-to-Administration (B2A)** z.B. elektronische Dienstleistung für öffentliche Verwaltung
	Administration	**Administration-to-Consumer (A2C)** z.B. Beantragung eines Personalausweises	**Administration-to-Business (A2B)** z.B. öffentliche Ausschreibung von Projekten	**Administration-to-Administration (A2A)** z.B. Zusammenarbeit von Gemeinden

Abbildung 1.3: Übersicht elektronischer Geschäftsbeziehungen, nach [169, Abb: 1.1]

dell folgende Phasen umfassen: Leistungsanbahnung, Leistungsvereinbarung und Leistungserbringung. E-Commerce umfaßt die Wertschöpfungskette des E-Business aus Abschnitt 1.1.1 Abb. 1.2 bis auf den eProcurement Schritt, da hier das betrachtete Unternehmen nicht als Leistungsanbieter auftritt. Der betriebswirtschaftliche Erfolg eines E-Commerce Angebots leitet sich aus den entstandenen Kosten und den erzielten Erlösen ab.

Die Kosten einer E-Commerce Website sind aus der internen Kostenrechnung eines Unternehmens nach dem von Stahlknecht et al. [231, S.470ff] erläuterten Prinzip des IT-Controlling, sowie der IT-Kosten- und Leistungsverrechnung zu ermitteln. Auf die Kostenseite geht diese Dissertation daher nicht weiter ein.

Die Erlöse und der gestiftete Nutzen eines Internetangebots sind dagegen erst noch zu erfassen und hängen vom jeweiligen Geschäfts- und Erlösmodell des Internetangebots ab.

Um den Erfolg messen zu können, ist es zunächst wichtig zu wissen, wie ein Unternehmen ein erfolgreiches Internetangebot definiert: Welche Besucher sollen angelockt werden? Welche

Botschaften sollen vermittelt werden? Was soll ein User auf der Website erreichen können und was soll er aus Sicht des Unternehmens tun?

Schonberg et al. geben in [210] mehrere Beispiele, welche Ziele mit einer Website verfolgt werden können: Unterhaltung, Erziehung, Vertrieb von Produkten und Dienstleistungen, Produktwerbung, Serviceangebote oder Bereitstellung von Informationen. Schonberg et al. leiten die Erfolgsmaße direkt aus der Zielfestlegung ab. Sie geben dafür als Beispiel den Verkauf eines Produkts. Auf die übrigen genannten anderen Ziele gehen Schonberg et al. nicht ein.

Im Abschnitt 1.1.4 werden die Geschäftsmodelle erläutert und anschließend in Abschnitt 1.1.5 die entsprechenden Erlösmodelle beschrieben, von denen die Meßbarkeit des Erfolgs eines Internetangebots abhängt. Das Internet und E-Business stellen neue Herausforderungen an das Beziehungsmanagement zwischen einem Unternehmen und seinen Kunden bzw. den Stakeholdern. Zur weiteren Einordnung des Dissertationsthemas wird gemäß Abb. 1.1 auf den Aspekt des Customer Relationship Managements eingegangen.

1.1.3 E-Customer Relationship Management

Wilde et al. [128] geben folgende Definition des CRM: *"Customer Relationship Management (CRM) versteht sich als kundenorientierte Unternehmensstrategie, die mit Hilfe moderner Informationstechnologie versucht, auf lange Sicht profitable Kundenbeziehungen durch ganzheitliche und individuelle Marketing-, Vertriebs- und Servicekonzepte aufzubauen und zu festigen".*

Entgegen der Darstellung von Meier et al. in Abb. 1.2 kann das eCRM nicht lediglich als letzter Schritt in der Wertschöpfungskette angesehen werden, sondern beeinflußt den kompletten Wertschöpfungsprozeß. Khalifa et al. nennen in [143] drei Kundenbeziehungsphasen, in denen eCRM ansetzen kann: die Vorkaufsphase, Kaufzeitpunkt und Nachkaufphase. Wie man bei der Behandlung der Internet-Geschäftsmodelle in Abschnitt 1.1.4 sehen wird, beschränkt sich der Einsatz von CRM nicht ausschließlich auf den Verkauf von Produkten und Dienstleistungen. Denn Hippner et al. weisen in [121, S.26] nach, daß der Nutzen einer Kundenbeziehung nicht nur aus der eigentlichen Transaktion erwächst, sondern ebenso durch “weiche“ Faktoren beeinflußt wird. Die dieser Arbeit zugrundeliegenden Aktionen der Kunden auf einer Website umfassen nach Stauss [233] und Bromberger [42] alle Phasen eines Kundenbeziehungslebenszykluses. Dies wird in Abb. 1.4 deutlich. Die dunkel eingefärbten Felder kennzeichnen die von dieser Dissertation angesprochenen Systeme, Daten, Medien, Kanäle und Aufgaben eines CRM. Der Schwerpunkt der angewandten Methoden liegt im analytischen CRM.

Es werden nicht nur Endkunden im engsten Sinne als *Kunde* angesehen, sondern es werden alle Besucher einer Unternehmenswebsite mit eingeschlossen, wie beispielsweise Investoren,

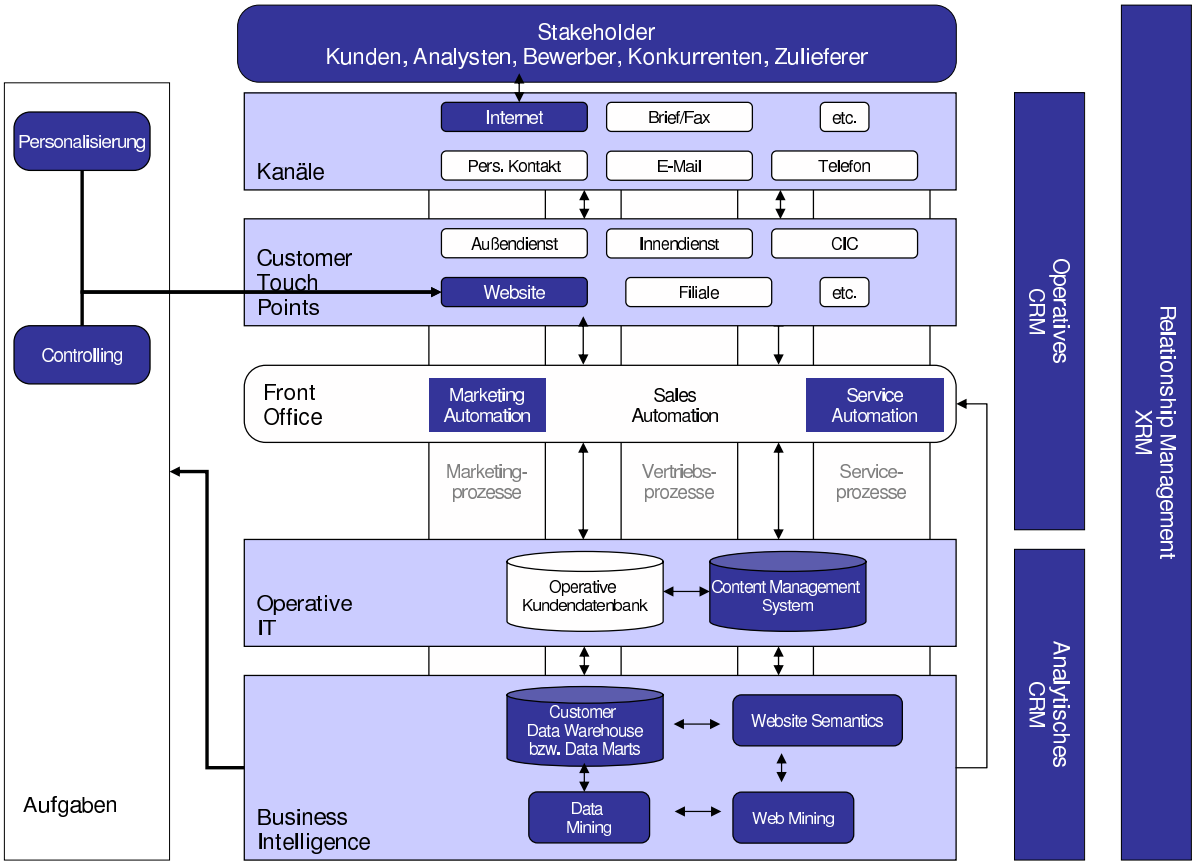


Abbildung 1.4: Komponenten eines CRM-Systems, nach Hippner et al. [126, Abb:1]

Zulieferer, Journalisten, Mitarbeiter und Arbeitssuchende, die Meier et al. [169] als Stakeholder zusammenfassen. Hierfür wird auch der Begriff Relationship Marketing verwendet [121, S.19][27].

Das Ziel eines profitablen Kundenbeziehungsmanagements kann als Maximierung des Kundenwertes ausgedrückt werden. Entgegen der traditionellen Unterteilung des Kundenwertes in monetäre und nicht monetäre Größen, unterscheidet Hippner [121, S.27ff] zwischen dem *Transaktionspotential* und dem *Relationspotential* eines Kunden. Das Transaktionspotential beschreibt den gegenwärtigen und zukünftigen Verkaufserfolg. Das Relationspotential umfaßt ein **Referenzpotential**, das die Einflußnahme auf die Kaufentscheidung Dritter widerspiegelt [72, S.161], sowie das **Informationspotential** und **Kooperationspotential**, die den Wert des Kunden durch Informationsbereitstellung oder Zusammenarbeit beschreiben.

1.1.4 Internet Geschäftsmodelle

An welcher Stelle ein E-Commerce Angebot bzw. eine Unternehmens-Website den Kundenwert beeinflussen kann, hängt wiederum vom zugrundeliegenden Geschäftsmodell der Website ab. Anand et al. sehen in [10, S.27] eine Website als erfolgreich an, wenn die Ziele des Webautors erfüllt sind. In diesem Kapitel werden daher die Ziele beschrieben, die mit einer Website verfolgt werden können und in denen sich die Ziele des Website-Besitzers widerspiegeln.

Bis sich das Internet als zusätzliches Kontaktmedium etabliert hatte, wurden seine Möglichkeiten mit sehr viel Euphorie bewertet. Von der Vielzahl an Geschäftsideen haben sich inzwischen unterschiedliche Geschäftsmodelle herausgebildet.

Ein Geschäftsmodell beschreibt nach Wirtz et al. [264], welche Kombinationen von Produktionsfaktoren durch die Geschäftsstrategie eines Unternehmens eingesetzt werden. Schwickert [212] geht davon aus, daß ein E-Business Geschäftsmodell die Unternehmensstruktur, die Marktleistung, den Prozeß der Leistungserstellung und die Erlösgenerierung enthält. Krüger et al. [152] beschreiben ein Geschäftsmodell als Darstellung des sozialen Systems eines Unternehmens. Es charakterisiert, welche externen Ressourcen in ein Unternehmen fließen und dort durch die internen Ressourcen weiterverarbeitet werden.

Wirtz et al. [264] unterscheiden vier Basisgeschäftsmodelltypen : Content, Commerce, Context und Connection. Bei der Charakterisierung einer spezifischen Website sind Überlappungen zwischen den Geschäftsmodelltypen möglich. Wie bei Wirtz werden in dieser Arbeit die Internet Geschäftsmodelle auf die Teilaspekte fokussiert, die relevant für das Medium Internet sind.

Content -Geschäftsmodelle werden von Wirtz et al.[264] in E-Information, E-Education und E-Entertainment unterteilt. Das Ziel von Content-orientierten Websites ist es, dem Nutzer Inhalte anzubieten. Ergänzend zu Wirtz et al. soll betont werden, daß dem Nutzer der von ihm gewünschte Inhalt außerdem bequem und leicht zugänglich bereitgestellt werden soll.

Commerce -Geschäftsmodelle umfassen lt. Wirtz et al.[264] die Anbahnung, Aushandlung und Abwicklung von Geschäftstransaktionen. “Das Ziel ist eine Unterstützung, Ergänzung oder Substitution der traditionellen Phasen einer Transaktion durch das Internet“.

Context -Anbieter klassifizieren und systematisieren die im Internet verfügbaren elektronischen Informationen, fassen diese zusammen und dienen als Navigationshilfe. Hierzu zählen zum Beispiel Suchmaschinen oder Webkataloge.

Connection -Websites bieten die Möglichkeit des Informationsaustauschs mit anderen Nutzern innerhalb eines Netzwerkes.

Schwickert [212] und Kruerger [152] geben jeweils eine Übersicht verschiedener Geschäftsmodelltypologien. Neben der von Wirtz unterscheidet Schwickert [212], inwieweit die Glieder der Wertschöpfungskette durch das Internet-Geschäftsmodell abgedeckt werden. Je mehr Bestandteile des Geschäftsprozesses innerhalb des Internets abgewickelt werden, desto ausgeprägter sei der E-Business-Charakter des Geschäftsmodells. Schwickert erläutert weiter, daß E-Business-Geschäftsmodelle meist aus einer Ansammlung von Teilgeschäftsmodellen bestehen.

Von den Geschäftsmodellen hängen die Erlösmodelle ab, da vom Angebot einer Website die Zahlungsbereitschaft der User abhängt. Besteht bei den Usern keine Zahlungsbereitschaft für das Internetangebot eines Unternehmens, muß auf andere Finanzierungs- bzw. Erlösquellen zurückgegriffen werden.

1.1.5 Erlösmodelle

Skiera et al. [218] weisen darauf hin, daß Erlöse nur dann erzielt werden können, wenn ein Wert geschaffen wird, für den ein anderer Marktteilnehmer zu zahlen bereit ist. Zerdick et al. [272] unterscheiden dabei zwischen direkten und indirekten Erlösen. Direkte Erlöse erhält man von den Nutzern einer geschaffenen Leistung. Indirekte Erlöse stammen von Dritten, die selber einen Nutzen daraus ziehen, daß ein User die Leistung einer Website in Anspruch nimmt.

Abgesehen von den direkten Erlösen einer transaktionsorientierten Website, können durch E-Business auch anderer Werte geschaffen werden. Browne et al. [43] nennen beispielsweise: effizientere Kundenbeziehungen, Kostenreduktion, sowie ein schnellerer und einfacherer Marktzugang. Die effizientere Kundenbeziehung kann in dem von Hippner [121] beschriebenen Relations- und Referenzpotential bestehen, das zum Kundenwert beiträgt. Hierbei wird deutlich, wie eine Unternehmens-Website zu den Zielen eines eCRM-Projekts sowohl durch monetäre wie auch nicht monetäre Erlöse beitragen kann.

Schwickert et al. systematisieren in [212] mögliche Erlösformen. Die Transaktionsabhängigkeit des Geschäftsmodells ist dabei der entscheidende Faktor. Außerdem werden wie bei Krüger et al.[152]zwischen direkter und indirekter Erlösgenerierung unterschieden.

Meier et al. [169, S.49f] beschreiben ein Erlösmodell als den materiellen und immateriellen Nutzen aus der Geschäftstätigkeit, wobei direkte Erträge aus der Geschäftstätigkeit des Unternehmens resultieren und indirekte Erträge aus dem unternehmensinternen oder -externen Kapitalmarkt stammen. Abb. 1.5 zeigt diese Sichtweise der Ertragsmodelle von Meier et al.

und Birkhofer [31], der die indirekten Ertragsmodelle (weiß) und die direkten Ertragsmodelle (grau) dargestellt hat.

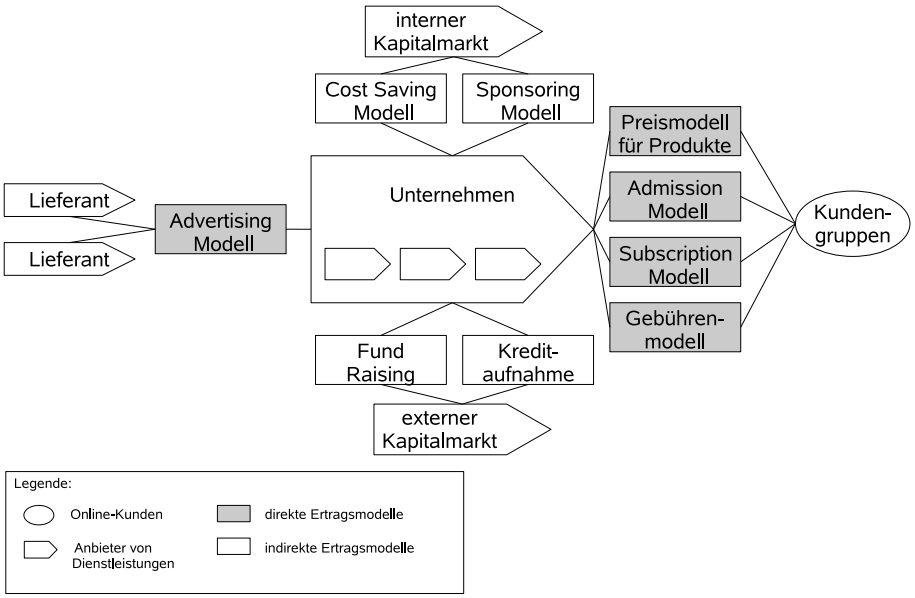


Abbildung 1.5: Ertragsmodelle in elektronischen Märkten nach Birkhofer [31]

Meier et al. [169] erläutern die direkten Erlösmodelle aus Abb. 1.5 wie folgt:

Advertising Modell : verkauft Werbeflächen auf stark frequentierten Websites. Dieses Modell ist besonders für Unternehmen bzw. Websites mit einer starken Marktposition interessant. Beispiel: Google

Preismodell für Produkte : sind der Standardfall von transaktionsorientierten E-Commerce Websites, bei denen ein Produkt verkauft wird. Beispiel: Amazon

Admission Modell : verlangt vom Kunden eine Eintrittsgebühr zur zeitlich befristeten Nutzung der Inhalte. Beispiel: Online-Literatur-Datenbanken: ACM, IEEE.

Subscription Modell : die verkauften Inhalte werden in fixen oder periodischen Abonnementsbeiträgen abgerechnet.

Gebührenmodell : Diese Transaktionsgebühr richtet sich entweder nach der Nutzungsdauer, den heruntergeladenen Inhalten oder einem angebotenen Support.

Die indirekten Erlösmodelle werden von Meier et al. [169] nur kurz erläutert. Im **Cost Saving Modell** soll die Investition in eine Website durch Kosteneinsparungen refinanziert werden. Finanziert ein interner Geschäftsbereich den Webauftritt für eine bestimmte Zeit, liegt das **Sponsoring Modell** vor. Beide Modelle können sich auch kombinieren, indem die Kosteneinsparungen nicht unbedingt beim finanzierenden Geschäftsbereich liegen müssen. Die Erlösmodelle, die sich aus dem externen Kapitalmarkt finanzieren, werden nicht weiter erläutert und finden selten Anwendung. Die Erlösmodell sollten einem der anderen Modelle entsprechen, da insbesondere bei einer Finanzierung über den externen Kapitalmarkt entsprechende Erträge in Aussicht stehen müssen.

Meier et al. gehen nicht auf die Meßbarkeit der direkten oder indirekten Erlöse ein. Die Höhe der indirekten Erlöse und deren ursächlicher Entstehungspunkt bleiben bei Meier et al. [169] unbestimmt. Somit stellt sich die Frage der Meßbarkeit von direkten und indirekten Erlösen, denn erst dadurch wäre eine Aussage zum Erfolg einer Website möglich.

Der Nutzen von Internet-Auftritten für Unternehmen, die keinen E-Commerce anbieten, d.h. keine Produkte oder Dienstleistungen per Internet absetzen, besteht entweder in indirekten Erlösen oder eingesparten Kosten. Dies entspricht dem Cost-Saving-Modell in Abb. 1.5. Indirekte Erlöse können durch Marketingmaßnahmen im Internet entstehen. Da Vervielfältigungs- und Distributionskosten im Internet vernachlässigbar sind [264], werden im Internet gegenüber anderen Medien Kosten eingespart. Ebenso entsteht für ein Unternehmen ein Nutzen, wenn sich Kunden und potentielle Kunden im Internet über Produkte informieren, Serviceangebote wahrnehmen und sich Analysten, Arbeitsplatzbewerber, Pressevertreter, Mitarbeiter oder Geschäftspartner über das Unternehmen informieren. Im Rahmen einer Mehrkanalstrategie kann der Erlös dann auf einem anderen Kanal generiert werden. Der in einem anderen Kanal entstandene Nutzen ist ursächlich schwer dem Internet zuzurechnen.

Einen weiteren Überblick über Erlösquellen von Internet Geschäftsmodellen geben Skiera et al. in [218]. Skiera beschreibt, wie die für ein Unternehmen denkbaren Erlösquellen im Internet von der Position des Unternehmens entlang der Wertschöpfungskette von Porter [195] abhängen. Skiera kommt zu dem Schluß, daß die Stufen in der Wertschöpfungskette, die näher am Kunden sind, ein tendenziell höheres Potential, Erlöse zu erzielen, aufweisen. Skiera et al. nennen drei Erlösarten und unterscheiden diese wie [169, 264] zwischen direkten und indirekten Erlösen:

- Produkte, Dienstleistungen: Bücher, DVD, Software, Musik direkte Erlöse
- Kontakte: Werbung, Bannerwerbung, E-Mail-Werbung indirekte Erlöse
- Informationen: Nutzerprofile, Paneldaten indirekte Erlöse

Man hat in den unterschiedlichen Erlösmodellen gesehen, daß durch direkte Erlöse der Nutzen einer Unternehmens-Website direkt ersichtlich wird. Die indirekten Erlösmodelle generieren

für ein Unternehmen einen Nutzen, der erst an anderer Stelle zu Tage tritt. Mit dem Problem der nicht direkten Zurechenbarkeit der Wertschöpfung einer Website beschäftigen sich Teltzrow et al. in [248]. Sie stellen die Website eines Unternehmens in den Zusammenhang einer Multichannel-Strategie, in der zum Beispiel die Informationsphase auf einer Website stattfindet, die Transaktion aber im traditionellen Einzelhandel abgeschlossen wird. Wie man in Abb. 1.4 sehen kann, stellt das Internet nur einen unter mehreren, oft parallelen Kundenkontaktkanälen dar.

Bezogen auf den von Hippner et al. [121] beschriebenen Kundenwert, zielen indirekte Erlösmodelle mehr auf des Relationspotential und direkte Erlösmodelle auf das Transaktionspotential. Analog dazu wird im folgenden Abschnitt untersucht, wie die Erlöse transaktionsorientierter Websites mit direkten Erlösmodellen gemessen werden können und wie die indirekten Erlöse nicht transaktionsorientierter Websites mit ihren indirekten Erlösmodellen erfaßt werden können.

1.1.6 Erfolgsmessung

Anhand der vier Basisgeschäftsmodelle von Wirtz et al. [264] und der oben beschriebenen Erlösmodelle, wird in diesem Abschnitt untersucht, wie und woran der Erfolg einer Website gemessen werden kann. Die oben vorgestellten Geschäfts- und Erlösmodelle werden hierzu in Abb. 1.6 zusammengefaßt.

Dabei nennen sie folgende mögliche Ziele: User zu Konsumenten machen, diese wiederum zum häufigen Besuch der Website bringen, bestimmte Informationen vermitteln oder einfach die Zahl der Besucher erhöhen. Der zentrale Punkt für die Meßbarkeit des Erfolges bei allen Geschäftsmodellen ist die Verfügbarkeit eines User Feedbacks bezüglich des Ziels der Internetseite. Die Art und Verfügbarkeit eines Feedbacks hängt wiederum vom Geschäftsmodelltyp und Erlösmodell ab.

Commerce Die zu diesem Geschäftsmodell gehörenden Websites eines Unternehmens sind auf Transaktionen ausgerichtet. Es werden Produkte oder Dienstleistungen angeboten, für die der User je nach Erlösmodell bezahlen oder persönliche Informationen bereitstellen muß. In beiden Fällen identifiziert sich der User und bleibt nicht länger anonym, sodaß er in Zukunft persönlich angesprochen werden kann und CRM-Systeme direkt anwendbar werden. Der Erfolg der Website wird nicht nur durch die Tatsache meßbar, daß eine Transaktion stattgefunden hat, sondern der Erfolg wird durch die Höhe der Transaktion auch quantifizierbar. Der User gibt durch seine Zahlungsbereitschaft ein direktes Feedback über die Höhe des Erfolgs bei einer erfolgreich abgeschlossenen Transaktion auf transaktionsbasierten Websites.

Wirtz et al. [264, 263] unterteilen das Commerce-Geschäftsmodell weiter in Anbahnung, Aushandeln und Abwickeln. Nur wenn alle drei Schritte erfolgreich abgeschlossen sind, ist auch

		Geschäftsmodell			
		Commerce	**Content**	**Connection**	**Context**
Erlösmodell	**direkt**	Preismodell, Advertising, Subscription, Admission, Gebührenmodell	Advertising, Subscription, Admission, Gebührenmodell z.B. Online-Zeitungsartikel gegen Gebühr	Advertising, Subscription, Admission, Gebührenmodell z.B. Zugang zu elektronischer Bibliothek	Advertising, z.B. Werbung bei Google
	indirekt	Cost Saving, Sponsoring, Multichannel Strategy	Cost Saving, Sponsoring, Multichannel Strategy	Cost Saving, Sponsoring, Multichannel Strategy	Cost Saving, Sponsoring, Multichannel Strategy

Abbildung 1.6: Zusammenfassung von Geschäfts- und Erlösmodellen

die Transaktion abgeschlossen. Aussagen über den Erfolg der Website lassen sich dennoch in allen Fällen treffen, da alle Erfolgsmaße sich an dem Punkt orientieren, ob die Transaktion erfolgreich war oder nicht. Ist dieser Meßpunkt nicht verfügbar, wie beim Content-Geschäftsmodell, ist eine Erfolgsmessung problematisch.

Content Das Ziel des Content-Geschäftsmodells ist es, dem User Inhalte bzw. Informationen zur Verfügung zu stellen. Bei den direkten Erlösmodellen wird eine Erfolgsmessung analog zum Commerce-Geschäftsmodell durchgeführt.

Bietet ein Unternehmen auf seiner Website lediglich Informationen über sich und seine Produkte frei zugänglich an, liegt ein indirektes Erlösmodell vor. Hier erhält man kein direktes Feedback des Users. Man kann lediglich verfolgen, ob der User Informationen abgerufen hat, nicht aber ob er gefunden hat, wonach er gesucht hat.

Freshwater drückt es in [220] so aus, daß die einzig harte Währung im Internet die Aufmerksamkeit ist. Verfehlt eine Website es, ihren Inhalt in kurzer und attraktiver Zeit zu vermitteln, verliert ein User schnell das Interesse, und der zeitliche und monetäre Aufwand zur Erstellung der Website war umsonst. Die Aufmerksamkeit ist allerdings auf der Server-Seite nicht meßbar, sondern muß in aufwendigen Studien direkt in persönlichem Kontakt mit den Usern erfragt werden.

Connection Websites des Geschäftsmodelltyps Connection bieten dem User die Möglichkeit, mit anderen Usern in Kontakt zu treten. Da die Kommunikation zwischen den Usern über die jeweilige Internetseite abgewickelt wird, wie zum Beispiel in Foren, kann der Erfolg der Seite, sprich die Kommunikation, direkt verfolgt und ermittelt werden. Auch der Inhalt der Kommunikation gibt Aufschluß über eine für die User hilfreiche und zufriedenstellende Kommunikation, sofern dies der Datenschutz erlaubt. Je nach Angebot können über Zusatzdienste bei entsprechender Zahlungsbereitschaft Erlöse generiert werden, die wiederum eine direkte monetäre Rückmeldung über den Erfolg der Website darstellen.

Context Suchmaschinen und Webkataloge, die in diese Geschäftsmodellkategorie fallen, bieten dem User einen Nutzen, indem sie ihn bei seiner Informationssuche unterstützen und ihm Informationen in geordneter Form aufbereiten. Meist sind die eigentlichen Inhalte lediglich verlinkt, sodaß sie außerhalb der Website des Anbieters liegen. Die Ergebnislisten von Suchmaschinen verlinken nahezu ausschließlich auf andere Websites.

Wie bei informationsbasierten Seiten des Typs Content, kann der Anbieter auch nicht feststellen, ob der User mit der Leistung einer Suchmaschine zufrieden ist. Der Erfolg einer Suchmaschine besteht in einer möglichst großen Anzahl an Nutzern. Das Geschäftsmodell beruht nicht auf der Vermittlung bestimmter Inhalte, sondern auf der bequemen, zielgerichteten, schnellen und hochqualitativen Suchfunktion. Erlöse können indirekt durch Werbung oder Suchmaschinenmarketing erzielt werden.

Der Erfolg dieses Geschäftsmodelltyps läßt sich durch das Weitervermitteln von Usern an die Werbeauftraggeber direkt messen. Zufriedene Nutzer werden das Angebot auch zukünftig nutzen. Wiederkehrende User können als zufrieden angesehen werden, wenn sich durch wiederholte Nutzung ihre Zufriedenheit ausdrückt.

Zusammenfassend kann man feststellen, daß sich die Erfolgsmessung von Internet Geschäftsmodellen bei indirekten Erlösmodellen, insbesondere beim Content-Geschäftsmodell, am schwierigsten gestaltet.

Von den oben vorgestellten Geschäftsmodellen ist Commerce das unproblematischste. Anhand einer erfolgreichen oder nicht erfolgten Transaktion läßt sich ein Bündel an Kennzahlen und Analysen ausrichten und somit der Erfolg der Website sogar in monetären Werten ausdrücken. Diese Kennzahlen werden in Kapitel 3 erläutert.

Websites des Geschäftsmodelltyps Connection und Context nutzen oft indirekte Erlösmodelle, die sich mit einfachen Statistiken, wie Anzahl der Besucher und Anzahl der angeklickten Banner, zufriedenstellend auswerten lassen. Da die indirekten Erlöse von werbeschaltenden Unternehmen stammen, können sie genau beziffert werden.

Anders sieht es mit Internetseiten des Geschäftsmodelltyps Content aus. Im folgenden werden Websites vom Geschäftsmodelltyp Content als *informationsorientierte Websites* bezeichnet.

Der Anbieter des Contents ist daran interessiert, daß der Besucher der Website den von ihm gewünschten Inhalt findet und ihn zur Kenntnis nimmt. Der Anbieter kann aber auch den User auf andere Inhalte aufmerksam machen wollen, den der User ursprünglich beim Betreten der Website nicht gesucht hat. Dieser Website-Perspektive stehen die Ziele der User gegenüber. Spiliopoulou et al. drücken es in [226, S.1] so aus, daß die Qualität einer informationsorientierten Website nicht alleine durch deren ansprechendes Äußeres bestimmt wird, vielmehr sei eine Website ein komplexes Netzwerk von Seiten, das den User darin unterstützen soll, Informationen auf eine intuitive Weise zu finden.

Wie kann nun der Erfolg von informationsorientierten Websites ermittelt werden?

1.2 Erfolgsmessung auf informationsorientierten Websites

Um einen Ansatzpunkt für eine Erfolgsmessung zu finden, der unabhängig von einer Transaktion ist, sollen die von Madeja [165] untersuchten Erfolgsfaktoren von E-Commerce betrachtet werden. In Abb. 1.7 werden die Faktoren und Stufen des Website-Erfolgs in ihre kausale Reihenfolge gebracht. Die Features einer Website bestimmen die Qualität einer Website, die sich wiederum auf die Usability auswirkt. Je besser diese ist, desto mehr Traffic zieht eine Website an und erhöht die User Akzeptanz. Der direkte Wirkungszusammenhang mit dem Unternehmenserfolg wird aber lediglich in einem Arbeitspapier der Washington University [199] belegt, das bei reinen Online-Firmen einen positiven Wirkungszusammenhang zwischen 50% der Website Features und dem Shareholder Value nachgewiesen hat. Diese Untersuchung gibt keinen Aufschluß darüber, worin der Zusammenhang besteht. Die von Madeja genannten

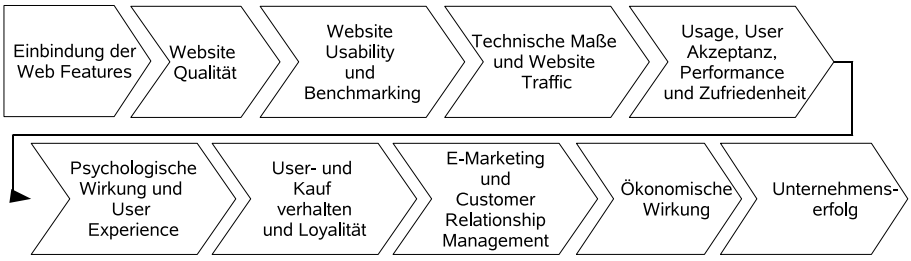


Abbildung 1.7: Wirkungskette des Website-Erfolgs, Madeja[165, Abb.:7.6]

Schritte auf dem Weg zum ökonomischen Website-Erfolg umfassen subjektive Kriterien der Website wie Qualität und Usability sowie User-Zufriedenheit und technische Maße. Der Fokus dieser Dissertation liegt auf Maßen, die auf der Seite des Website-Anbieters fortlaufend

erfaßt werden können. Auf der User- bzw. Client-Seite erhobene Daten stehen in der Regel nicht fortlaufend zur Verfügung und müssen durch Befragung erhoben werden.

Spiliopoulou et al. beschreiben in [225] einen Ansatz, den *Erfolg von nicht verkaufsorientierten Websites* zu messen. Sie teilen dazu die Webpages einer Website in Zielseiten und Aktionsseiten ein. Nach diesem Konzept hat ein User auf einer Aktionsseite sein Ziel noch nicht erreicht. Das Erreichen einer Zielseite ist mit dem Erreichen des Ziels einer Website gleichgesetzt und beschreibt den Erfolg dieser Website.

Sullivan [246] modelliert den Website-Erfolg im Verhältnis zu den von ihm beschriebenen drei Qualitätsmerkmalen einer Website: 1. Servicequalität wie Antwortzeiten, 2. Navigationsqualität und 3. Zugangsqualität zu der Website. Die Arbeit von Sullivan von 1997 ist noch von technischen Beschränkungen aus der Anfangszeit des E-Commerce geprägt.

Spiliopoulou et al. [225] gehen in drei Stufen vor: 1. Modellierung des Website-Inhalts hinsichtlich der Website-Ziele, 2. Kategorisierung der User-Aktivitäten anhand der User-Ziele, und 3. Definition des Website-Erfolgs als die Effizienz ihrer Komponenten, dem User bei der Erreichung seiner Ziele zu helfen. Zwar beziehen Spiliopoulou et al. alle Dienste einer Website mit ein, letztendlich beruht das Website-Ziel wiederum auf dem Verkauf der angebotenen Produkte. Auf die vorgestellten Grundideen von Ziel- und Aktionsseiten wird in Kapitel 5 bei der Entwicklung eigener Website-Maße zurückgegriffen.

Das E-Commerce-Qualitätsmodell in Abb. 1.8 ist eine Darstellung verschiedener Einflußfaktoren der vom Kunden wahrgenommenen Qualität einer Website. Im Mittelpunkt steht der Interaktionsprozeß des Users mit dem Unternehmen bzw. dessen Website. Aus dieser Interaktion bildet sich der User ein Qualitätsurteil, das aus der Abweichung der erwarteten Qualität von der wahrgenommenen Qualität resultiert. Der Interaktionsprozeß und das User-Qualitätsurteil sind detaillierter in der Wirkungskette in Abb. 1.7 dargestellt. An der Aussage des in Abb. 1.8 gezeigten Modells ändert es jedoch nichts.

Um nach Totz et al. [252] ein Qualitätsurteil über eine Website zu erhalten, muß man den Interaktionsprozeß des Users mit der Website verstehen. Daraus leitet der User sein Qualitätsurteil ab. Wie von Eighmey in [88] und Madeja in Abb. 1.7 beschrieben, ist das Qualitätsurteil für den ökonomischen Erfolg einer Website mitentscheidend.

Wie man bei der Beschreibung der zur Verfügung stehenden Daten in Kapitel 2 sehen wird, stehen für die Analysen in dieser Arbeit nur die unternehmensseitigen Daten zur Verfügung. Aus dieser Beschränkung ergibt sich die Hauptaufgabe, aus den verfügbaren Daten ein User-Feedback abzuleiten.

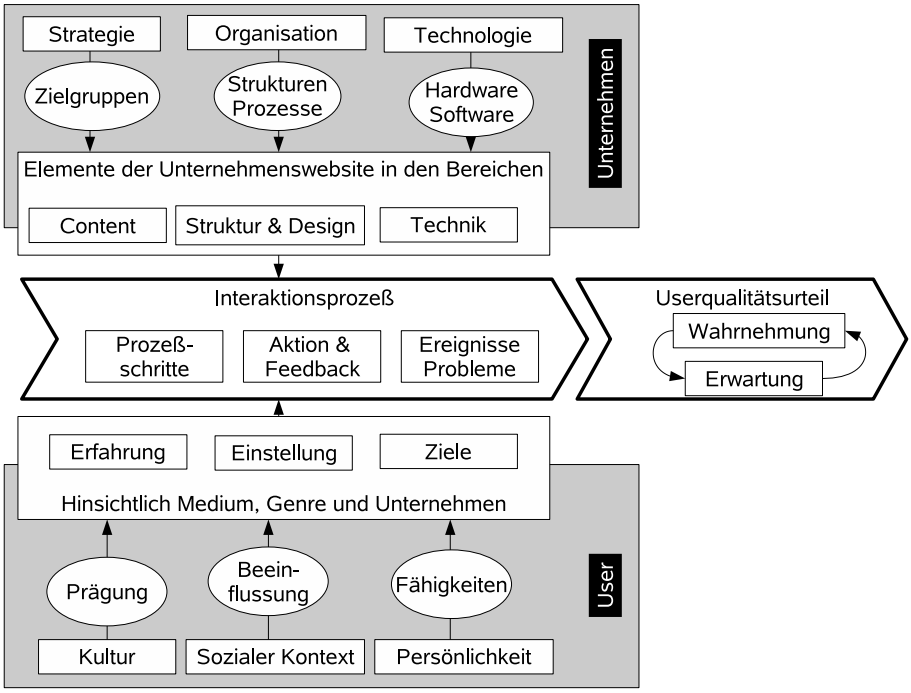


Abbildung 1.8: Das E-Commerce Qualitätsmodell, angelehnt an [252, Abb. 3],[201]

1.3 Gliederung der Dissertation

Der Aufbau dieser Arbeit orientiert sich am Web Mining-Prozeß in Abb. 1.9, der bei Wilde et al. in [124, S.8ff], Pyle [197, S.9-37] sowie Berry und Linoff [28, S.17-35] beschrieben wurde. Die Aufgabendefinition erfolgt in diesem Kapitel 1.

Die Datenauswahl, -aufbereitung und -integration werden in Kapitel 2 erklärt. Es wird auf die Besonderheiten von Daten, die im Internet anfallen, sowie auf deren Erfassung eingegangen.

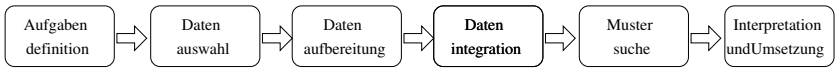


Abbildung 1.9: Web Mining-Prozeßschritte nach [173, 123]

Die Mustersuche und anschließende Interpretation der Ergebnisse erfolgt in den Kapiteln 3 und 4. In Kapitel 3 werden bekannte Web Metriken auf deren Einsatzfähigkeit zur Beurteilung des Erfolg informationsorientierter Websites hin untersucht.

Da Metriken und Maße zumeist ein sehr hohes Abstraktionsniveau besitzen, eignen sie sich nur bedingt zum Verständnis des einzelnen User-Verhaltens. Web Mining-Ansätze erheben

einen anderen Anspruch auf die Anwendbarkeit ihrer Ergebnisse als Web Metriken. Mit ihnen soll ein tieferes Verständnis des User-Verhaltens erreicht werden. Sie werden in Kapitel 4 vorgestellt und es werden verschiedene Methoden, Analysen und Algorithmen besprochen, die auf dem Weg zu einem Erfolgsmaß für informationsorientierte Websites hilfreich sind.

Im letzten Kapitel 5 wird mit Hilfe der erläuterten Web Mining-Methoden in einem eigenen Ansatz ein neues Maß zur Beurteilung des Erfolgs von informationsorientierten Websites erstellt und erläutert. In Abb. 1.10 ist die Vorgehensweise bei der Erstellung und Gliederung der Dissertation nochmals gezeigt.

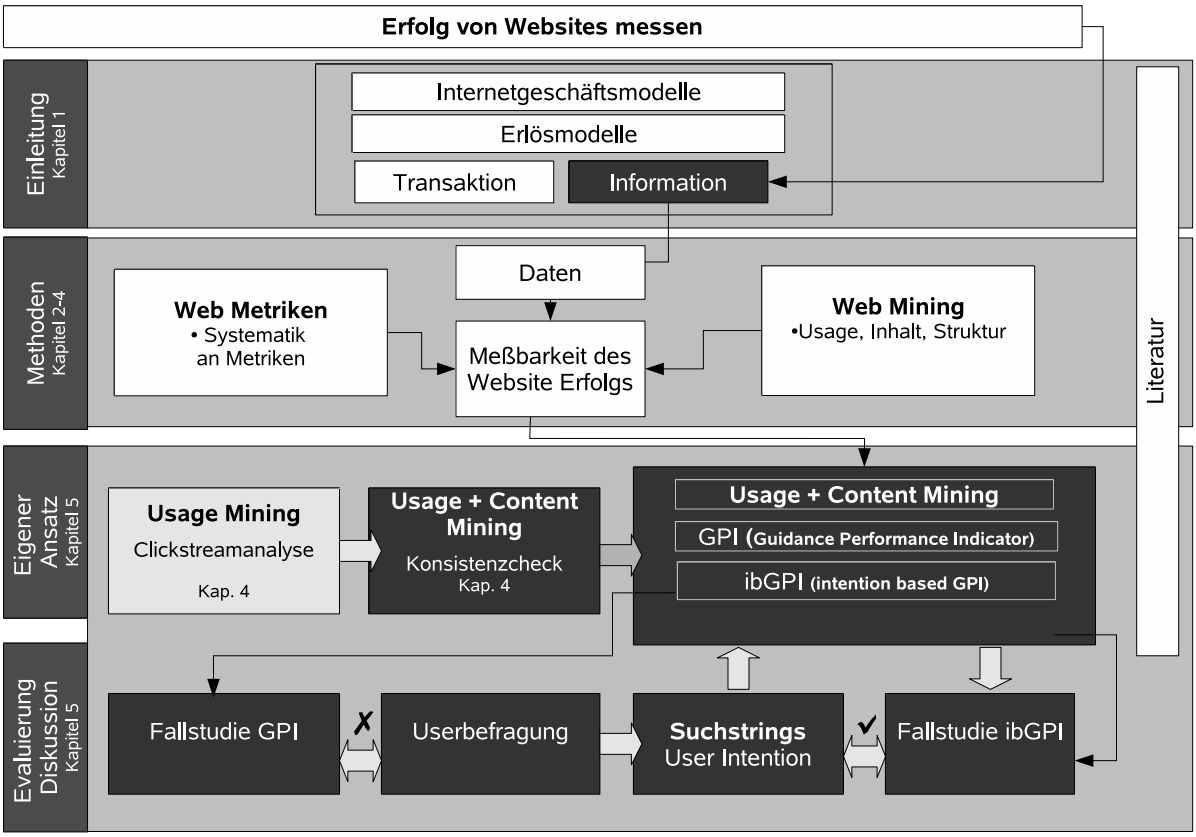


Abbildung 1.10: Aufbau der Dissertation

Aus einer Kombination des User-Verhaltens und des Inhalts der besuchten Webpages wird in Abschnitt 5.1 ein Erfolgsmaß, der Guidance Performance Indicator (GPI) abgeleitet. Der GPI beurteilt die Erfüllung des Ziels einer Website, den User zu Inhalten auf dieser Website zu leiten. Er beurteilt den Erfolg der Website aus Sicht des Unternehmens.

Um die vielfältigen Abhängigkeiten, die aus der Interaktion eines Users mit einer Website entstehen, überschauen zu können, wird ein formales Modell in Abschnitt 5.1.6 erstellt.

Der GPI wird in Abschnitt 5.2 auf mehrere Websites angewandt. Diese Bewertungen werden durch die Befragung von Usern in einer User-Studie in Abschnitt 5.3 überprüft.

Aus den Erkenntnissen der User-Studie ergibt sich die Notwendigkeit, die Website nicht nur aus Sicht des Unternehmens zu beurteilen, sondern auch die Sicht des Users zu berücksichtigen. Im Abschnitt 5.4 wird ein Ansatz vorgestellt, wie die User-Ziele durch die Analyse von Suchanfragen dieser User für eine Analyse verfügbar gemacht werden können. Das formale Modell wird für die neue Sichtweise erweitert. Auf formalem Weg wird ein neues Erfolgsmaß, der *intentionbased GPI* entwickelt, der die Zielerreichung der Website nun aus Sicht der User wiedergeben kann. Mit dem neuen Maß können die Diskrepanzen aus der User-Studie beseitigt werden.

Analog zu dem oben in Abb. 1.8 gezeigten Qualitätsmodell werden die zwei entwickelten Maße die Sicht des Unternehmens und die Sicht der User auf den Interaktionsprozeß einer Website beurteilen und zu einem Gesamturteil verdichten.

2 Daten

Unternehmen versuchen verstärkt, die Beziehung zu ihren ”virtuellen” Kunden im Internet zu intensivieren, wobei sich hierfür insbesondere das Customer Relationship Management (CRM) als tragfähiges Konzept erwiesen hat, wie Wilde et al. in [124] erläutern. Grundlage für ein erfolgreiches Management der Kundenbeziehungen ist das Wissen über den Kunden und seine Bedürfnisse. Wichtige Fragestellungen über die Eigenschaften der User und die Wirkung des Internet Auftritts eines Unternehmens bleiben unbeantwortet, wenn der Betreiber nur wenig Informationen über die User und die Website zur Verfügung hat. Erst mit diesem Wissen kann eine Website an die Interessen und Bedürfnisse ihrer Besucher angepasst werden und so die Zugriffszahlen erhöht werden, wie es Sane Solutions für ihre Logfile-Analyse beschreiben [205].

Für die Aussagefähigkeit der angewandten Web Mining-Algorithmen und Web Metriken spielt die Qualität und Zuverlässigkeit der gesammelten Daten eine entscheidende Rolle. Es bestehen unterschiedliche Technologien, mit denen das Benutzerverhalten im Internet aufgezeichnet werden kann und der Inhalt und die Struktur einer Website erfaßt werden können. In diesem Abschnitt werden die unterschiedlichen Datenquellen, verschiedene Datensammeltechnologien und die Aufbereitung der gesammelten Daten beschrieben. Es wird auf die Vor- und Nachteile der Datenerhebungstechniken und die sich daraus ergebenden Qualitätsunterschiede eingegangen.

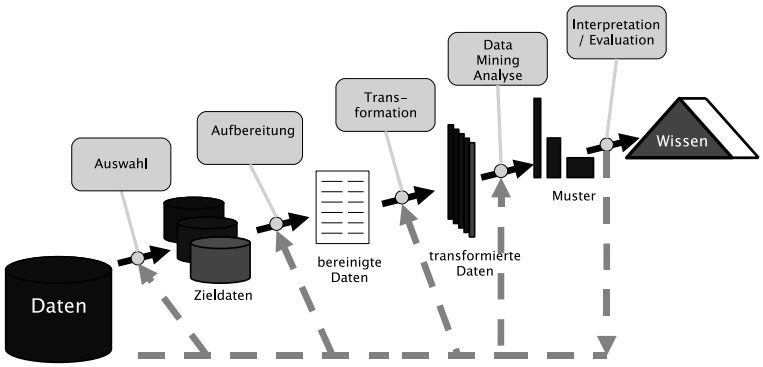


Abbildung 2.1: Der Data Mining-Prozeß [94]

Daten im Internet sind sehr heterogen und daher nicht einfach vergleichbar wie Daten aus anderen Unternehmensteilen, wie Berendt et al. in [22, S.2] ausführen. Sie reichen von freiem Text, semi-strukturierten Daten wie Listen oder Tabellen, Bildern, Hyperlinks bis hin zu den aufgezeichneten Usage-Daten, beispielweise aus Server-Logfiles oder einem Tracking-System.

Dieses Kapitel beschreibt die Prozeßschritte von Auswahl, Aufbereitung und Transformation der Daten, wie sie im Data Mining-Prozeß in Abb. 2.1 dargestellt sind. Das Ziel ist es, eine konsistente, hochqualitative Datenbasis für nachfolgende Analysen und Auswertungen bereitzustellen.

2.1 Datenquellen

Die für eine Analyse der Website notwendigen Daten stammen aus unterschiedlichen Quellen. Abbildung 2.2 zeigt die drei grundlegenden Dimensionen, in die sich die Daten einordnen lassen: Website Usage, Website Content und Website Structure.

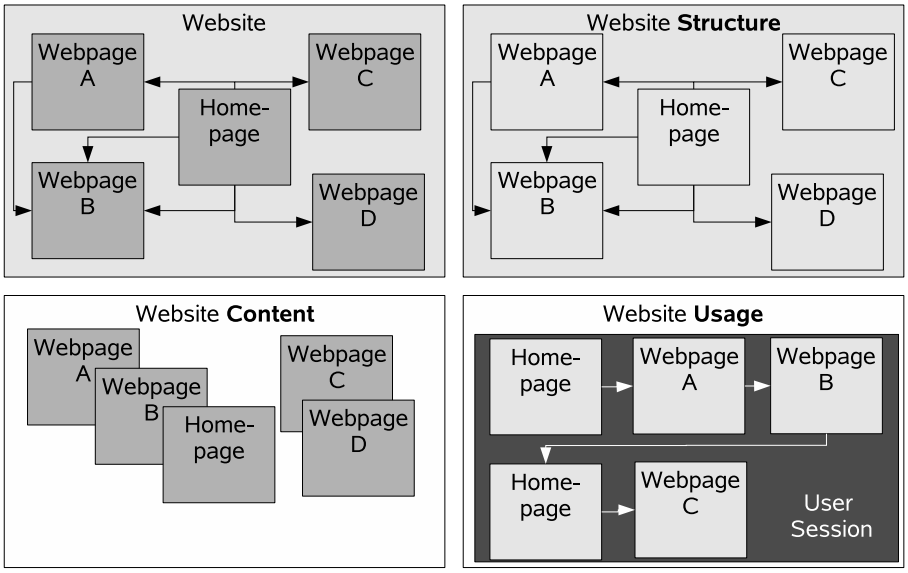


Abbildung 2.2: Datenquellen

Web User- und Usage-Daten Usage-Daten beschreiben das User-Verhalten, das durch die Bewegung des Users zwischen den Webpages einer Website entsteht [229]. User-Daten bezeichnen alle Daten, die den Internetnutzer beschreiben. Sofern er sich auf der Website

registriert, kann er als Einzelperson identifiziert werden. Einem User können dadurch über eine Session hinaus Eigenschaften und Interessen zugeordnet werden. Ohne Registrierung bleibt ein User anonym, kann aber durch den Einsatz persistenter Cookies als wiederkehrender User erkannt werden [175].

Web Content-Daten Der Content oder Inhalt einer Website wird durch dessen Autor festgelegt, um darin die Ziele der Website zu vermitteln. Der Content einer Website ist auf einer oder mehreren Webpages (Webseiten) angeordnet und kann aus beliebigen Formaten bestehen: z.B. Text, Bild, Grafik, Film, Ton. Die Inhalte einer Webpage können auch dynamisch erzeugt werden, sodaß mehreren Usern unterschiedliche Inhalte auf der gleichen Webpage zum gleichen Zeitpunkt präsentiert werden. Schließt man eine manuelle Änderung durch den Website-Autor aus, kann dem selben User zu unterschiedlichen, eng aufeinanderfolgenden Zeitpunkten auf einer dynamisch erzeugten Webpage ebenfalls unterschiedliche Inhalte dargestellt werden. Content-Daten beschreiben den Inhalt einer einzelnen Webpage und berücksichtigen nicht den Zusammenhang, in dem sie stehen. Dazu muß die Struktur der Website miteinbezogen werden.

Web Structure-Daten Die Struktur einer Website wird durch deren Autor festgelegt, indem er den Content nach seinen Vorstellungen anordnet und verlinkt. Die Verlinkung zwischen den Webpages bildet die Struktur der Website. Durch diese Links werden die einzelnen Webpages zu einer Website zusammengefaßt. Man kann die Struktur einer Website im mathematischen Sinne als Graphen betrachten, bei dem die Webpages die Knoten und die Links die Kanten bilden.

In den folgenden Abschnitten 2.2, 2.3 und 2.4 wird auf die Besonderheiten bei der Datenaufbereitung der jeweiligen Datentypen und auf die Anforderungen der späteren Analysen eingegangen.

2.2 Web User- und Usage-Daten

2.2.1 Datenerhebung

Daten über die Präferenzen und Absichten der User können nach Anand et al. [10, S.26] entweder **explizit** durch Befragung der User oder **implizit** durch Beobachtung des User-Verhaltens gesammelt werden. Je nach Geschäftsmodell fallen unterschiedliche Daten bei der Benutzung einer Website an. Zu diesen Daten gehören fast immer die Bewegungen der User

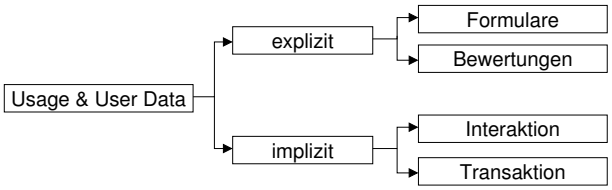


Abbildung 2.3: Datenerhebungsformen für User-Daten [10, nach Abb.2]

auf der Website und soweit vorhanden Registrierungsdaten oder Transaktionsdaten, wie sie bei E-Commerce Anwendungen anfallen.

Bei einer Registrierung gibt der User Informationen über sich explizit bekannt [207, S.112]. Diese Datenerhebungsform ist insbesondere bei transaktionsorientierten Websites häufig, da sich der Kunde identifizieren muß, um eine Waren- und Geldtransaktion durchführen zu können. Hierbei tritt der User aus seiner durch die Technik des Internets vorgegebenen Anonymität heraus. Er ist nun für das Unternehmen bei jedem weiteren Login-Vorgang auf der Website eindeutig identifizierbar. Der volle Nutzen der Login-Information eines Kunden wird durch ein CRM-System ausschöpfbar, da im besten Fall die komplette Kundenhistorie und Kundenwünsche abrufbar sind. Dadurch lassen sich die Inhalte einer Website für den User personalisieren und in die anderen Kundenkontaktkanäle integrieren. Der User muß allerdigns ein Interesse oder einen Anreiz haben, Informationen per Registrierung über sich preiszugeben.

In dieser Dissertation werden nur implizit erhobene Daten verwendet, da diese für die untersuchten Websites durch ein integriertes Web Reporting- & Mining-System automatisch und kontinuierlich zur Verfügung stehen. Explizit erhobene Daten, etwa bei Befragungen, sind nur zeitpunktbezogen und sind für eine dauerhafte Analyse und einen kontinuierlichen Verbesserungsprozeß nicht geeignet.

Anforderungen an die Usage-Daten

Unabhängig, ob die Erhebungsmethode bereits die gewünschten Datenformate und Datenqualität liefert, müssen mehrere Anforderungen erfüllt werden.

Aktionen, die durch den User ausgelöst wurden, müssen lt. Mobasher et al. [175, S.10] von Server-Aktivitäten und Besonderheiten des Seitenaufbaus getrennt werden (**Identifikation von Seitenaufrufen**). Einzelne User und deren Sessions müssen identifiziert werden (**Session- und User-Identifikation**), Weingärtner [222, S. 893ff].

Die einzelnen Aktionen eines Users innerhalb einer User-Sitzung müssen in eine zeitliche Reihenfolge gebracht (**Session Identifikation**) [223, S.495] und gegebenenfalls ergänzt werden (**Pfadvervollständigung**), siehe Spiliopoulou [223]. Dies ist nur notwendig im Fall von Caching durch den Client Browser oder durch einen dazwischenliegenden Proxy. Das Caching

kann auch umgangen werden, indem durch geeignete Befehle der Browser angewiesen wird, die Webpage jedesmal neu zu laden.

Aus umgekehrter Sicht sollen die User-Aktionen aus gegebenenfalls mehreren Sitzungen einem User zugeordnet werden können (**User Identifikation**)[223, S.495].

Außerdem müssen die Aktivitäten von künstlichen *Usern*, wie **Crawlern**, Robots oder Spidern erkannt und entfernt oder als solche markiert werden.

Es existieren verschiedene Techniken, den User zu beobachten, die sich in der Art der Datensammlung und dem Aufwand der Datenaufbereitung unterscheiden. Bei der Betrachtung der einzelnen Datenerhebungsformen in den folgenden Abschnitten werden deren Vor- und Nachteile bezüglich der Erfüllung dieser Anforderungen diskutiert. Die folgenden Abschnitte betrachten keine Websites mit Authentifizierungs- bzw. Registrierungsmöglichkeit, sodaß der User anonym bleibt [106, S. 43].

Web Server Logfiles

Seit Beginn der Analysen der Internet Nutzung wurden und werden immer noch Logfiles von Internet-Servern als Datenquelle herangezogen. Logfiles sind Textdateien, in denen automatisch bestimmte Abläufe der Rechnertätigkeit festgehalten werden [124, S.9].

Es gibt unterschiedliche Logfile-Formate. Hippner et al. beschreiben in [124, S.9f] die Access Logfiles, in denen Anfragen von Nutzern an den Server erfaßt werden. Außerdem beschreiben Cooley et. al [69, 67] Referrer und Agent Logfiles.

Die Logfiles von Web Servern werden primär aus technischen Gründen wie Betriebs- oder Leistungsüberwachung geführt. Es wird der gesamte Datenverkehr zwischen dem Web Server und dem Rechner des Users (Client) protokolliert [205]. Web Logfiles protokollieren alle Aktionen eines Web Servers auf der Ebene von Hits, das heißt jede Zeile eines Logfiles entspricht dem Zugriff auf ein Element, aus dem eine Webpage aufgebaut ist. Eine Webpage ist meist aus sehr vielen Elementen aufgebaut. Bereits kleinste übertragene Text- oder Grafikteile resultieren in einem Hit-Eintrag im Logfile. Auf die Berechnung von Hits, Page Views, Sessions und Usern wird in Kapitel 3 detaillierter eingegangen.

Die Anzahl der Logfile-Einträge (**Hits**) läßt jedoch nur indirekt auf die Anzahl der angeforderten Webpages schließen [124], da bei gleicher Zahl an Webpage-Aufrufen die Zahl der Hits durch unterschiedliche Anzahl an zu übertragenden Seitenelementen stark variieren kann [69]. Die Anzahl der Hits ist somit lediglich für die Server-Performanz entscheidend, jedoch nicht bei der Usage-Analyse. Es interessiert lediglich der Aufruf einer Webpage durch den User und nicht die technische Umsetzung, bedingt durch die Webpage-Konstruktion. Diese Logfile Einträge müßten aufwendig entfernt werden.

Der Vorteil der Logfiles liegt in ihrer meist kostenlosen und verbreiteten Verfügbarkeit. Das bedeutet, daß keine zusätzlichen Komponenten zur Datenerhebung benötigt werden. Allerdings ist der Datenaufbereitungsaufwand wesentlich größer als bei den weiter unten erläuterten Methoden. Außerdem ist die Identifikation von Usern in Logfiles problematisch und fehleranfällig, wie Wilde et al. in [124, S.11] feststellen, da kaum Identifikationsmerkmale des Users aufgezeichnet werden[106, S. 43]. Lediglich die IP-Adresse des Clients ist bekannt. Diese IP-Adresse ist meist nicht eindeutig einem einzelnen Client-PCs zuzuordnen, beispielsweise wenn zwischen Client und Server ein Proxy geschaltet ist. Das gleiche gilt, wenn der User über einen Internet Service Provider (ISP) Zugriff aufs Internet hat [124, S. 16f]. Dann erscheinen alle Browser, die über den ISP ins Internet gelangt sind, unter der gleichen IP-Adresse. Wird eine IP-Adresse dynamisch vergeben, kann ein wiederkehrender User nicht als solcher erkannt werden.

Durch einen Browser- und Proxycache kommt ein Seitenaufruf nicht unbedingt beim Web Server an und kann von ihm im Logfile nicht protokolliert werden. Ist die angeforderte Webpage noch im Zwischenspeicher (Cache) des Browsers, wird vom Browser keine Anfrage an den Server gesendet. User Sessions aus Logfiles werden dadurch unvollständig und der gesamte Website Verkehr wird unterschätzt.

Die Identifikation von Crawlern und Robots ist schwierig, solange diese sich nicht selbst in einem Agentlog identifizieren, z.B. indem die Browser Identifikation benutzt wird, um sich als Crawler auszuweisen. Es müssen dann die einzelnen Sessions mit Heuristiken analysiert werden, die nach crawlertypischem Verhalten suchen. Werden sehr viele Webpages innerhalb weniger Sekunden hintereinander aufgerufen, entspricht dies kaum dem Verhalten eines Menschen.

Die Integration zusätzlicher Datenquellen, wie Content- oder Structure-Daten, gestaltet sich schwierig, da nur die URL in den Logfiles gespeichert wird. Die URLs eines Content Management Systems (CMS) sind meist sehr komplex aufgebaut, sodaß ein aufwendiger Aufbereitungsprozeß notwendig ist.

Network Sniffer

Network Sniffer oder Netzwerkmonitore stellen eine ergänzende oder alternative Datenquelle zu Web Server Logfiles dar. Der Netzwerkmonitor sammelt die vom Web Server ausgesandten TCP/IP-Pakete und schreibt sie in eine Datei im Standard-Logfile-Format [205, S.4]. Im Normalfall liefern sie das gleiche Resultat wie Web Server Logs. Ihre Vor- und Nachteile stimmen mit denen der Web Server Logs weitgehend überein. Sane Solutions weisen in [205] darauf hin, daß der Hauptvorteil in der direkten Auswertbarkeit in Echtzeit liegt.

Beacon image

Bei der Technik des Beacon Image, Page Tagging oder Hidden Image wird durch ein in das HTML-Dokument eingebettetes, verstecktes Bild (Größe 1 Pixel) oder Javascript-Tag [205, S.4f] der Client Browser dazugebracht, Informationen an einen Tracking Server zu senden [106, S.50]. Die Informationen werden direkt vom Computer bzw. vom Browser des Users an einen vorzugsweise separaten Tracking Server gesendet. Dabei können zusätzliche Infomationen übertragen werden, die vorher in das HTML Dokument eingebettet wurden. So kann eine eindeutige Session ID zurückgesandt werden, die zusammengehörige Clicks identifiziert und so die Datenqualität lt. Säuberlich [207, S.114] erheblich verbessern kann. Dies ist besonders dann interessant, wenn die Website durch ein Content Management System erstellt wurde [124, S.14].

Neben dem Vorteil einer eindeutigen Session ID und zusätzlicher Informationen, liegt der Hauptvorteil eines Beacon Image im asynchronen Datenfluß. Dadurch wird die Performanz des Web Servers kaum beeinflußt und Wartezeiten bleiben für den User gering. Außerdem beeinflussen Crawler und Robots die Tracking-Resultate nicht, da sie Bilder meist nicht laden. Dadurch wird auch das Hidden Image nicht vom Server angefordert, und es findet keine Aufzeichnung statt. Sich identifizierende Crawler können auf anderem Wege in die Website-Statistik mitaufgenommen werden, wohingegen die meisten anonymen Crawler automatisch ausgeschlossen bleiben.

Caching kann dadurch umgangen werden [106, S.50], indem das Hidden Image als nicht cachable gekennzeichnet wird. So wird dem Browser mitgeteilt, das Hidden Image bei jedem Webpage Aufruf erneut zu laden, [106, S.45] [207, S.113]. Es muß nur das Hidden Image neu geladen werden und nicht die komplette Webpage, wodurch Ladezeiten und Übertragungskapazitäten gering gehalten werden.

Der Datenaufbereitungsaufwand ist beim Einsatz eines Beacon Image sehr gering, weil im Gegensatz zu Logfiles nur die User-Aktionen an den Tracking Server gemeldet werden und gleichzeitig keine User-Aktionen durch Caching fehlen. So entfällt der Schritt der Pfadvervollständigung völlig. Die gesamten überflüssigen Einträge, wie sie in Logfiles anfallen (Hits), entstehen bei Page Tagging erst gar nicht, sodaß die Daten fast direkt ohne großen Aufbereitungsaufwand in eine Datenbank geschrieben werden können.

Ein Nachteil besteht in der notwendigen Integration des Hidden Image bzw. Javascripts auf jeder Webpage. Wird dieses aber durch ein Content Management System erstellt, kann automatisch auf jeder Seite ein entsprechendes Hidden Image eingesetzt werden. Es können auch zusätzliche Informationen aus dem Content Management System über das Hidden Image an das Tracking System weitergeleitet werden.

Bei Formaten wie pdf-Dokumenten ist das Hidden Image schwieriger zu integrieren. Gaul et al. weisen in [106, S.50] darauf hin, daß selbst bei deaktiviertem Javascript auf ein statisches

Hidden Image referenziert werden kann und zumindest protokolliert wird, welche Webpages abgerufen wurden.

URL rewriting

Auf dem Weg zwischen Client und Server bearbeitet die URL-Rewriting-Anwendung sowohl die Anfragen des Client als auch die vom Server gesendeten HTML Dokumente [123, S.13]. Dabei wird der gesamte Datenverkehr durch diese Applikation geschleust, was zu Performanzproblemen führen kann. Außerdem stellt eine solche Anwendung im gesamten Datenverkehr einen für die Ausfallsicherheit kritischen Punkt dar. Das Problem von Proxies und Caching führt hier zu einer systematischen Unterschätzung der User-Zahlen. Roboter und Crawler lassen sich nicht leicht identifizieren. Die Integration in die bestehende Server-Architektur ist komplex und bedarf einer permanenten Pflege und Abstimmung mit dem Server-Betrieb.

2.2.2 Datenbereinigung

Crawler, Robots- und Spider-Zugriffe bereinigen

Die nicht von menschlichen Usern stammenden Zugriffe auf eine Website durch automatische Programme wie Crawler müssen aus den Daten entfernt, kenntlich gemacht oder gesondert betrachtet werden. Säuberlich schlägt in [207, S.115] vor, entsprechende Listen zu verwenden, die IP-Adressen der Crawler enthalten, damit deren Zugriffe entfernt werden können. Viele Crawler identifizieren sich aber nicht, bzw. wollen unerkannt bleiben. Da Crawler eine Website automatisch und möglichst schnell durchlaufen, können diese Sitzungen durch Erkennen von ungewöhnlichen Verhaltensmustern aus weiteren Analysen ausgeschlossen werden. So ist es für einen Menschen ungewöhnlich, mehrere hundert Seiten pro Minute zu öffnen.

Bereinigen von fehlgeschlagenen Zugriffen

Diesen Schritt beschreibt Säuberlich [207, S.117] insbesondere für Web Logfiles, weil hier auch fehlgeschlagene Anfragen von Clients an den Server protokolliert werden, z.B. auf nicht vorhandene Seiten. Da diese für die Analyse der vorhandenen Seiten nicht einsetzbar sind, müssen sie ausgeschlossen werden. Eine gesonderte Auswertung kann jedoch wichtige Hinweise liefern, wenn man erkennen kann, warum der User das gewünschte Ziel nicht erreicht hat.

Fehlgeschlagene Zugriffe können vom Beacon Image dadurch ausgeschlossen werden, daß das Beacon Image erst als letztes Element einer Webpage übertragen wird. So wird sichergestellt, daß nur komplett aufgebaute Webpages als Seitenzugriffe erfaßt werden.

2.2.3 Datenaufbereitung

Der Datenaufbereitungsprozeß von Web Usage-Daten wird in mehreren Arbeiten, wie der von Dai et al [76] sowie Mobasher et al. [174, 173] beschrieben. Sie gehen dabei insbesondere auf die Probleme der Session-Identifikation ein.

Identifikation einzelner Sessions und User

Eine große Herausforderung besteht insbesondere bei Logfiles in der Identifikation einzelner Sitzungen (Sessions) und wiederkehrender User [124, S.16ff][106, S.43]. Nur durch die Identifikation einzelner Sessions kann ein zusammenhängender Clickstream aus den zeitlich aufeinanderfolgenden User-Aktionen zusammengestellt werden.

Im Gegensatz zur Aussage des Whitepapers der Firma Sane Solutions [206, S.2ff] sind Session-Identifikationstechniken nicht auf Logfile Analysen beschränkt. Das Web Reporting- und Mining-System in Abschnitt 2.2.4 kombiniert Page Tagging, Javascript, Cookies, Session IDs und IP Adressen. Dadurch werden die zusätzlichen Informationen eingebunden, die durch Javascript erhoben werden können. Ist Javascript im Browser des Users nicht erlaubt, werden die übrigen Möglichkeiten wie Session IDs und Page Tags genutzt. Das Ergebnis ist in jedem Fall besser als Logfile-Analysen, wie sie Sane Solutions in [206] vorschlagen, die auf die Möglichkeiten von Javascript gänzlich verzichten.

IP Adressen sind nicht immer eindeutig einem User dauerhaft zuordenbar, da Internet Service Provider (ISP) eine begrenzte Anzahl von IP Adressen dynamisch an ihre Kunden vergeben. Dadurch lassen sich zwar einzelne Sessions unterscheiden, aber wiederkehrende User lassen sich so nicht erkennen [106, S. 42ff]. Ebenso werden einzelne Browser, die in einem Netzwerk hinter einem Proxy liegen, unter einer einzigen IP Adresse erkannt. Dadurch kann weder zwischen einzelnen Usern noch Sessions differenziert werden.

Einen möglichen Ansatz, verschiedene Nutzer mit gleicher IP Adresse voneinander zu unterscheiden, beschreiben Wilde et al. in [124, S.17f]. Dabei wird die IP Adresse mit der Browser-Kennung verknüpft, die Auskunft über die Browser Software und deren Versionsnummer gibt. Durch Kombination von IP Adresse und Browser-Kennung kann wesentlich genauer auf unterschiedliche User und damit Sessions geschlossen werden. Auf eine ähnliche Weise arbeitet der Ansatz von Cooley et al. in [69]. Allerdings ist die Browser-Kennung keine zuverlässige

Informationsquelle, da sie beliebig vom User abänderbar ist. So kann sich ein Firefox Browser als Internet Explorer ausgeben. Die Analyse der IP Adresse zur User Erkennung ist durch neue Methoden wie Page Tagging und Session IDs nicht mehr notwendig.

Wesentlich zuverlässiger als diese Heuristik sind die Verwendung von Cookies und Session IDs. Aus der IP Adresse lassen sich noch weitere Information gewinnen. Mittels **IP Adressen Übersetzung** (DNS Lookup) kann in vielen Fällen eine IP Adresse mit der Form *123.250.23.156* durch einen DNS Lookup einem Besitzer zugewiesen werden. Es kann nicht nur der Eigentümer, sondern auch oft der Ort (Land, Region und eventuell Stadt) ermittelt werden, sodaß ortsspezifische Auswertungen möglich werden. Diese Information ist allerdings mit Ungenauigkeiten verbunden, da nur der Ort registrierter Server bei einem DNS Lookup erkannt wird, nicht aber die dahinterliegende Netzwerklandschaft.

Cookies sind Textdateien, die vom Web Server erzeugt [205] und dauerhaft (persistente Cookies) oder für die Dauer einer User Session (temporäre Cookies) auf dem Client Computer gespeichert werden. Durch die Fähigkeit, über mehrere Sessions beim Client gespeichert zu bleiben, unterscheiden sich die Möglichkeiten eines Cookies lt. Säuberlich [207, S.111] von denen einer Session ID. Damit lassen sich User Sessions und User identifizieren [124][205, S.2f]. Bleibt das Cookie mehrere Sessions lang auf dem Client Computer gespeichert, können die einzelnen Sessions einem User zugeordnet werden, unter der Annahme, daß dieser Client Browser nur von einer Person genutzt wird [207, S.112]. Somit können wiederkehrende User identifiziert werden und deren Verhalten und Präferenzen wesentlich genauer analysiert werden. Ein Cookie kann jedoch jederzeit wieder gelöscht oder von vornherein abgelehnt werden.

Ein Cookie ist eine Ergänzung zu anderen Datenerfassungsmethoden. Die alleinige Nutzung von Cookies ist nicht ausreichend für eine Usage-Analyse. In Kombination mit anderen Verfahren, kann die Session- und User-Identifikation durch Cookies erheblich verbessert werden. Der Hauptvorteil von Cookies besteht darin, dass keinerlei Benutzer-Interaktion vonnöten ist [106, S.47], wie beispielsweise bei einer expliziten Registrierung.

Ein Nachteil von Cookies liegt in ihrem schlechten Image, das durch Verletzung der Privatsphäre in einigen Fällen verursacht wurde. Laut Säuberlich [207, S.113] hat dies zu einer vermehrten Deaktivierung der Cookie Annahme in Browsern geführt.

Session IDs dienen dazu, die einzelnen Aktionen eines Users einer Session zuzuordnen. Säuberlich beschreibt in [207, S.111], daß jedem Besucher mit der angeforderten Webpage vom Server eine eindeutige Kennung zugewiesen wird. Die Session ID kann zum Beispiel in der URL mitgeführt werden, sodaß Seitenaufrufe mit der gleichen Session ID einer Session zugeordnet werden können. Die Session ID wird für die Dauer einer Sitzung beibehalten. Dabei wird zur Abgrenzung von Sitzungen eine Zeitspanne festgelegt, beispielsweise 30 Minuten. Man unterstellt dabei, daß bei einer Unterbrechung von mehr als 30 Minuten der inhaltliche Zusammenhang nicht mehr gegeben ist und ein User einer neuen Aufgabe nachgeht.

Als Nachteil der Session ID nennt Säuberlich [207, S.111], daß die Session ID bei einem erneuten Besuch auf der Website verloren geht und auf diesem Wege wiederkehrende User nicht erkannt werden können. Dies kann aber wie beschrieben durch Cookies vermieden werden.

Pfadvervollständigung

Je nach der verwendeten Technologie bei der Datenerhebung von Usage-Daten liegt der Clickstream einer Session vollständig vor oder nicht. Durch Browser- oder Proxy Caches, den "Back"-Buttons eines Browsers, Bookmarks oder Direkteingabe einer URL, können Sprünge im Clickstream auftreten oder Webpages aufeinanderfolgen, die nicht durch Links erreichbar sind. Das Problem des Cachings kann wie oben in 2.2.1 beschrieben, vermieden werden.

Um die Clickstreams in die Seitenstruktur einfügen zu können, muß bei unvollständigen Clickstreams deren Pfad vervollständigt werden, wie Säuberlich in [207, S.118ff] beschreibt. Besonders Clickstreams aus Logfiles müssen vervollständigt werden. Abb. 2.4 zeigt, wie durch Caching der Clickstream nachträglich durch eine Heuristik vervollständigt werden kann. Kann man von Webpage l nicht direkt auf n gelangen, wird als Heuristik der kürzeste Weg von l über m nach n gewählt.

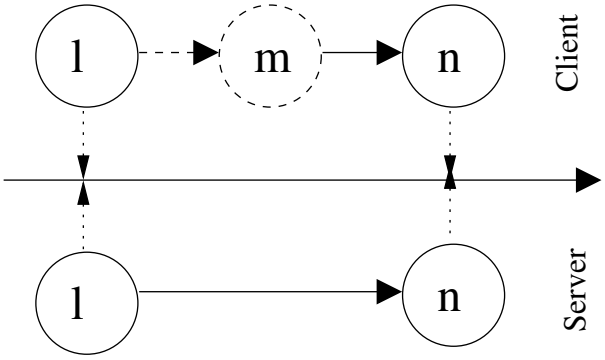


Abbildung 2.4: Pfadvervollständigung nach Caching

Zwei Arten der Pfadvervollständigung werden genannt [207, S.118ff]:

- Der Clickstream einer Session wird auf Sprünge untersucht. Diese Seiten werden entlang eines Linkgraphen der Website auf kürzestem Wege durch die dazwischenliegenden Seiten ergänzt.
- Unter der Annahme, daß der User mit dem Back Button seines Browsers den bisherigen Pfad rückwärts navigierte, wird im bisherigen Pfad rückwärts nach der Referrer-Seite des Zielpunktes gesucht und diese Seiten in der Session ergänzt.

Dabei tritt das Probleme auf, die Verweildauer des Users auf den ergänzten Seiten nicht feststellen zu können. Außerdem muß der Linkgraph der Website erstellt werden, um die kürzesten Wege berechnen und Sprünge im Clickstream feststellen zu können. Dies kann bei sich schnell ändernden oder dynamisch erzeugten Websites sehr zeitintensiv sein. Auch stellt die Pfadvervollständigung nur eine Heuristik dar, die durch neue Browsingtechniken zusätzlich erschwert wird, siehe Abschnitt 2.2.3.

Duration

Ist der Clickstream einer Session komplett, kann über die Zeitpunkte des Zugriffs auf die einzelnen Webpages die Dauer (Duration) berechnet werden, die der User auf der Seite verbracht hat. Dies wird in Abb. 2.5 veranschaulicht. Die Dauer ist ein wichtiger Hinweis auf die

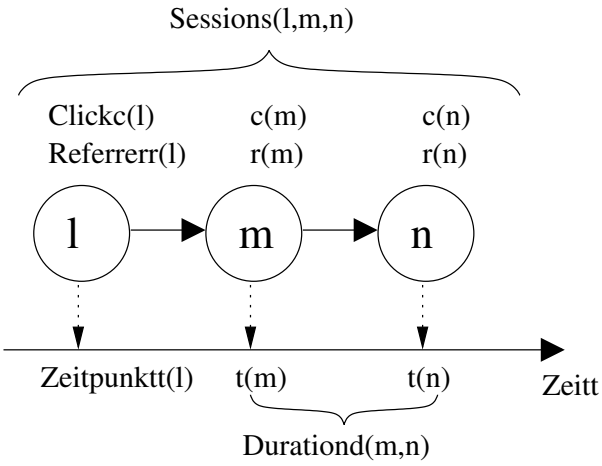


Abbildung 2.5: Duration-Berechnung

Aufmerksamkeit, die der User der Webpage entgegengebracht hat. Diese Interpretation muss jedoch eingeschränkt werden, da nicht bekannt ist, was der User während der Duration getan hat. Seine Aufmerksamkeit muß während der Duration nicht ausschließlich auf die Webpage gerichtet gewesen sein. Eine eingehendere Analyse der Duration findet sich in Kapitel 3.3.1.

Welche Komplikationen durch neue Browser-Techniken entstehen, die die Zuverlässigkeit der Kennzahl Duration beeinflussen, wird im folgenden Abschnitt beschrieben.

Exkurs: Tabbed Browsing

Die Notwendigkeit hochqualitativer Daten für die nachfolgenden Analysen zur Verfügung zu stellen, motivierte zu der Arbeit von Viermetz, Stolz et al. [254]. Darin werden die Unterschiede zwischen dem serverseitig beobachtbaren und dem tatsächlichen, clientseitigen Browsing-Verhalten untersucht.

Mit innovativen Browsern ist der User nicht mehr an ein simples Verfolgen von Links in einem Browserfenster gebunden. Im Gegensatz dazu kann ein User in mehreren Browserfenstern oder Tabs gleichzeitig parallele Clickstreams verfolgen. Wechselt ein User zwischen den Tabs hin und her, entsteht auf der Serverseite ein falsches Bild des User Verhaltens. Abb. 2.6 gibt hierfür ein Beispiel. Der User beginnt seine Session auf Webpage A und öffnet in einem neuen

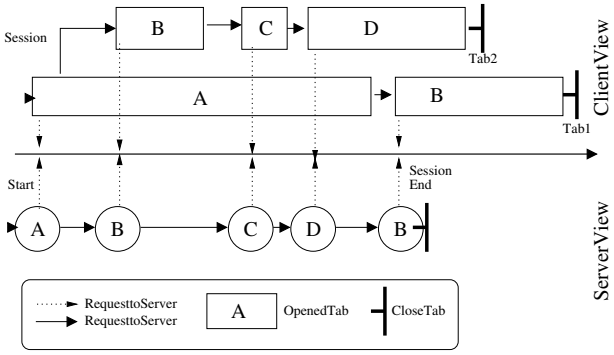


Abbildung 2.6: Paralleles Browsing-Verhalten

Tab Webpage B. Webpage A bleibt dabei geöffnet. Nach dem Aufruf von Webpage D in Tab 2, kehrt er in Tab 1 zurück und öffnet erneut Webpage B. Auch hier bleibt D geöffnet.

Wird die Duration wie in Abschnitt 2.2.3 berechnet, werden die offen gebliebenen Webpages A und D nicht berücksichtigt. Zum anderen wird nicht erkannt, von welcher Webpage aus der User die neuen Webpages aufgerufen hat. Auch werden Wechsel zwischen offenen Tabs nicht erkannt.

In [254] wird von Viermetz und Stolz et al. vorgeschlagen, an Stelle einzelner Webpage-Aufrufe zu Transitionen zwischen zwei Webpages unter Miteinbeziehung der Referrer-Information überzugehen. Der Referrer gibt Auskunft darüber, von wo die aktuelle Webpage aufgerufen wurde. So wird jeweils eine Transition von einer auf eine andere Webpage betrachtet.

Es wird ein Verfahren vorgeschlagen, das alle möglichen Clickstream-Kombinationen berechnet, die durch sequentielles oder paralleles Browsing Verhalten auftreten können. So kann bei serverseitiger Beobachtung in bestimmten Fällen Tabbed Browsing als solches erkannt werden.

An Stelle der serverseitigen Beobachtung kann der User auch von der Client-Seite aus beobachtet werden. Dieses Vorgehen wird von Hawkey [117] vorgeschlagen. Catledge et al. schlagen in [49] vor, den Browser des Clients zu modifizieren, sodaß der Browser alle User-Aktionen an einen Server meldet. Chan et al. [55] beschreiben das gleiche Vorgehen, stellen aber die datenschutzrechtliche Problematik und die Akzeptanzprobleme der User entgegen.

Nielson [179] führt 1993 eine frühe Untersuchung durch, bei der zwar Tabbed Browsing noch nicht existierte, aber durch parallele Browser-Fenster die gleiche Problematik parallelen Browsingverhaltens entsteht. Diese Arbeit und die von Claypool [64] gehen noch auf den Gebrauch des Back Buttons in einem Browser ein, der an sich kein paralleles Verhalten verursacht, bei dessen Gebrauch der User aber auch nicht der Linkstruktur folgt.

Das Problem parallelen Browsing-Verhaltens ist noch keineswegs erforscht oder bis auf [254] überhaupt problematisiert worden. Daher geht diese Dissertation von der traditionellen Sichtweise, dem sequentiellen Browsing-Verhalten, aus.

Datenintegration

Viele große Websites werden durch ein Content Management System verwaltet und erstellt. In diesen Systemen sind zusätzliche Informationen der einzelnen Webpages und ihres Inhalts hinterlegt. Werden die Usage-Daten auf solche Weise erfaßt, daß eine Verknüpfung zu den Datenbeständen des CMS möglich ist, können beide Datenbestände während des Datenaufbereitungsschritts zusammengeführt werden. Je nach Art der Daten des CMS, läßt die so entstandene Datenbasis Auswertungen der Benutzung im semantischen Zusammenhang zu. Der folgende Abschnitt 2.2.4 beschreibt ein solches System, das für diese Dissertation benutzt wurde.

2.2.4 Integriertes Web Reporting and Mining System

Die Daten, auf denen diese Arbeit basiert, wurden auf den Websites eines großen deutschen Unternehmens durch ein Web Reporting & Mining System erfasst und aufbereitet. Abbildung 2.7 zeigt schematisch den Prozeß, beginnend mit der Bereitstellung der Internetseiten durch eine Content Delivery Application (CDA) aus einem Content Management System (CMS). Die auf die Internetseiten zugreifenden User werden von der Tracking Engine mittels Beacon Image erfasst und in einer operationalen Datenbank gepeichert. Der ETL-Prozeß (Extraction,

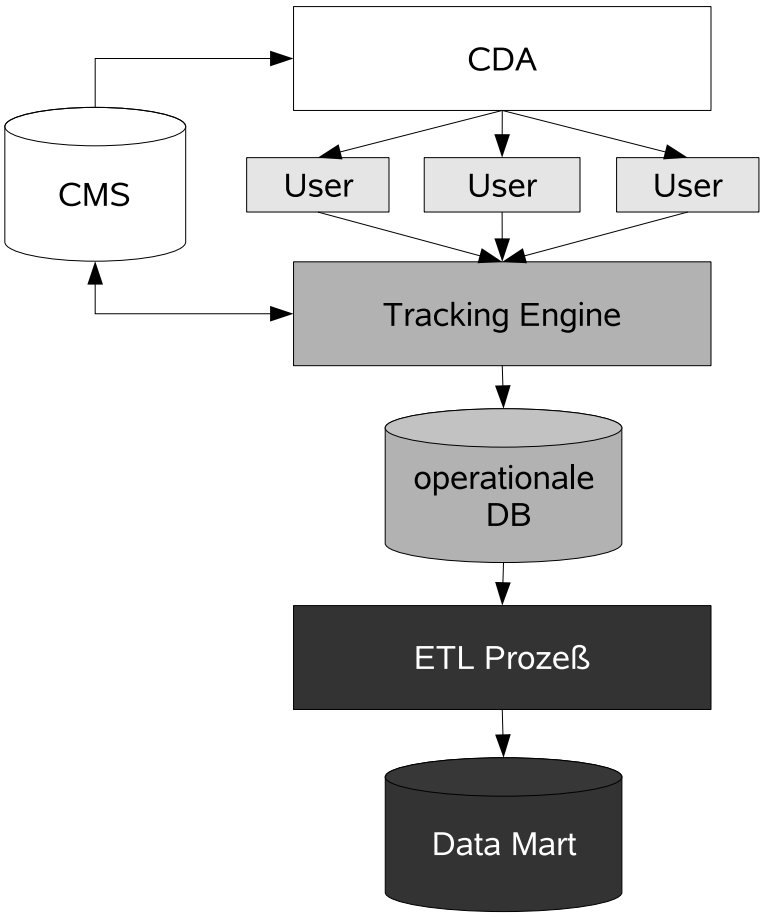


Abbildung 2.7: Integriertes Web Reporting und Mining System

Transformation, Load) ist eine Bezeichnung für den Datenaufbereitungs- und Transformationsschritt im Data Mining-Prozeß aus Abb. 2.1. Diese Daten werden täglich im ETL-Prozess aufbereitet. Dabei werden Sessions gebildet, die Durations berechnet, IP Adressen übersetzt und seitenspezifische Informationen mit dem CMS abgeglichen und zusätzlich für den schnelleren Zugriff aggregiert. Alle Daten werden in einer relationalen Datenbank (Data Mart) gespeichert. Auf dieser Datenbasis können deskriptive Statistiken erstellt (Web Reporting) und tiefergehende Analysen im Rahmen des Web Mining vorgenommen werden.

Die zusätzlichen Informationen, die durch die Anbindung des CMS erreicht wurden, umfassen: Angaben über den Autor, Erstellungsdatum des Inhalts, Titel des Inhalts sowie die genaue Zugehörigkeit einer Webpage zu den jeweiligen Unternehmensteilen. Das CMS dient außerdem dazu, die Webpages und deren Sprachklone als eine Webpage zu identifizieren. Dadurch ist eine konsistente Datenbasis ohne redundante Informationen gesichert.

Genaue Informationen über den Inhalt der Webpages liegen zwar als Metainformationen vor,

wurden aber nicht gepflegt. Sie sind somit nicht nutzbar. Inhaltliche Information über die Website müssen mit der im nächsten Abschnitt 2.3 beschriebenen Methode gesammelt werden.

2.3 Web Content-Daten

Der Content (Inhalt) einer Website besteht aus mehreren Ebenen. Da ein Inhalt jedem Betrachter unterschiedlich erscheint, muß zwischen den faktisch vorhandenen Daten und deren Bedeutung für den User und den Autor unterschieden werden. Bei der Erfassung der Daten wird nur die faktische Ebene berücksichtigt.

Der Content einer Webpage kann aus Text, Grafik, bewegten Bildern, Ton und Farben bestehen. Im weiteren werden nur Informationen berücksichtigt, die in Textform vorliegen. Zum einen wegen der ausgereiften Bearbeitungs- und Analysemöglichkeiten für textuelle Informationen, zum anderen, weil auf den hier untersuchten Websites der Großteil der Informationen als Text vorhanden ist.

2.3.1 Datenerhebung

Der Text auf den Webpages einer Website wird durch einen Crawler eingesammelt. Huang [135] und Gudivada et al. [110] geben eine technische Beschreibung, wie ein Crawler eingesetzt werden kann.

Die Sammlung der Daten erfolgt, wie in Abb. 4.5 und bei Aggarwal et al. [3] beschrieben. Ein Parser extrahiert aus den HTML-Dateien die reine Textinformation. Anschließend findet das sogenannte **Tokenizing** statt [12], bei dem die Worte aus dem Gesamttext einer Webpage als einzelne Elemente erkannt werden. Außerdem können Groß- und Kleinschreibung vereinheitlicht werden.

Die Aufbereitung und Analyse der gesammelten Daten wird ausführlich in Kapitel 4.3 erläutert.

2.3.2 Datenaufbereitung

Stopwords

Als nächstes werden die Daten bereinigt, indem, wie Fox [100] beschreibt, bestimmte Wörter, die man in einer Stopword-Liste festgelegt hat, entfernt werden. Dazu gehören Bindewör-

ter aus dem Satzbau wie *und, oder, aber, weil*, Eigennamen, Monats- und Tagesnamen. Eine Stopword-Liste kann individuell für eine Website angepaßt werden, um beispielsweise in dieser Arbeit den Firmennamen der Website zu entfernen. Da dieser auf jeder Webpage einer Website vorkommt, trägt er nichts zum Inhalt innerhalb dieser Website bei.

Stemming

Eine weitere Form der Aufbereitung von textuellen Daten ist das Stemming, bei dem Worte auf ihren Wortstamm zurückgeführt werden, sodaß Verben, Adjektive und Substantive des gleichen Wortstamms auf selbigen zurückgeführt werden. Der verbreitetste Stemming-Algorithmus stammt von Porter [195]. Der Porter-Algorithmus wurde für die englische Sprache entwickelt und läßt sich an einem bei Kosala et al. [151, S.5] gegebenen Beispiel verdeutlichen: *informative*, *information*, *informer*, *informed* werden auf deren gemeinsamen Wortstamm *inform* zurückgeführt. Dadurch erreicht man eine starke Vereinheitlichung und Dimensionsreduktion. Da die Daten der analysierten Websites auf Englisch vorliegen, kann der Porter Stemming-Algorithmus angewandt werden.

Synonyme

Wünschenswert bei der inhaltlichen Textanalyse ist außerdem die Erkennung von Synonymen und deren Konsolidierung auf jeweils ein Wort. Eine lexikalische Datenbank mit Synonymsuche bietet das Wordnet [172], das von Fellbaum in [97] beschrieben wird. Die Synonymerkennung wurde in dieser Arbeit aufgrund des hohen technischen Aufwands nicht berücksichtigt. Außerdem umgeht der später beschriebene PLSA-Algorithmus in Kap. 4.3.3 die Synonymproblematik.

Metainformationen

Mit Hilfe von Metainformationen ist es möglich, den Content jeder einzelnen Webpage zu beschreiben und in den Gesamtzusammenhang der Website semantisch einzuordnen. Mit Ontologien und Taxonomien können einzelne Begriffe und der komplette Content in die richtige Semantik eingeordnet werden. Leider sind für die hier analysierten Websites keine Semantiken im Sinne des Semantic Web, wie von Berendt [21] beschrieben, verfügbar.

Neben einem Semantic Web können durch den Autor der Website Metainformationen hinterlegt sein, die den Content beschreiben. Wie wir in [239] untersucht haben, leiden je nach Website Metainformationen unter mangelnder Pflege und sind daher nicht immer zuverlässig. Daher stützen wir uns auf den reinen Text einer Webpage.

2.3.3 Datenintegration

Zur Integration der Content-Daten in die Usage-Daten müssen beide Datenbasen die gleichen Identifikatoren für die Webpages besitzen. Das kann zum einen die URL sein oder die im Content Management System hinterlegten Content IDs. Es hat sich herausgestellt, daß in den URLs andere Content IDs mitgeführt werden, als durch das CMS und CDA im Tracking Code der Webpages hinterlegt wurden. Außerdem existierten mehrere Sprachklone für eine Webpage. Der Aufwand der Datenintegration war sehr groß. Letztendlich konnte aber eine konsistente Datenbasis erreicht werden. Dies zeigt, wie wichtig eine integrative Lösung bei derart heterogenen Systemen und Datenquellen ist.

2.4 Web Structure-Daten

Die Struktur einer Website ergibt sich durch die Verlinkung der Webpages. Da das Benutzerverhalten analysiert werden soll, sind diejenigen Links von Interesse, denen der User folgt bzw. die man beobachten kann. Abbildung 2.8 zeigt die unterschiedlichen Linktypen. Die Links innerhalb einer Webpage (Intrapage Links) tauchen im Clickstream nicht auf und werden im folgenden nicht weiter betrachtet. Der User folgt auf seinem Weg durch die Website

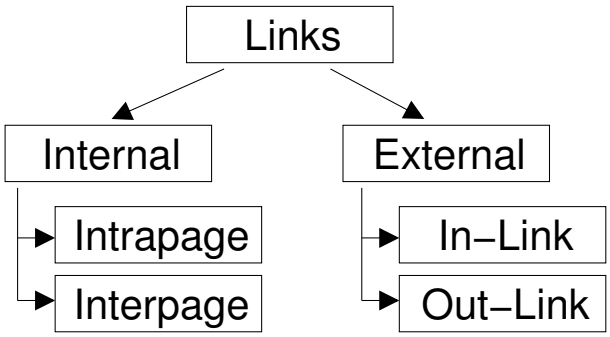


Abbildung 2.8: Link-Typen

den Interpage Links zwischen den Webpages. Diesen Links gilt das besonderes Interesse dieser Arbeit. Aus den Interpage Links der Website läßt sich ein gerichteter Graph der Website erstellen.

Hyperlinks, die von der betrachteten Website auf andere Sites verweisen, werden als Outlinks bezeichnet. Im Moment der Erstellung dieser Dissertation wurden die Outlinks noch nicht vom WRM-System aufgezeichnet, sodaß sie in dieser Arbeit nicht berücksichtigt werden konnten. Die durchgeführten Analysen erlauben jedoch auch die Einbindung dieser Links.

Die Link Struktur wird ebenfalls durch den Crawler zusammengetragen, der neben den textuellen Informationen auch die Links extrahiert. Die Datenintegration gestaltet sich analog zur Integration der Content-Daten sehr aufwendig, da auch hier Content IDs und URLs konsolidiert werden müssen.

Nach erfolgter Sammlung und Aufbereitung stehen die Daten für weitere Analysen zur Verfügung. In den folgenden Kapiteln 3 und 4 werden zwei Kategorien an Analysemöglichkeiten auf ihre Fähigkeiten hin untersucht, den Erfolg informationsorientierter Websites messen zu können. In Kapitel 3 werden bestehende Kennzahlen und Kennzahlensysteme beschrieben und beurteilt. Web Mining Analysen in Kapitel 4 versuchen die Zusammenhänge der Fakten sichtbar zu machen. Dadurch soll ein Verständnis des Userverhaltens erreicht werden. Im Verständnis der in Kapitel 1 beschriebenen Website Erfolgs- und Qualitätsmodelle soll daraus ein User-Qualitätsurteil erstellt werden können.

3 Web Metriken und Maßzahlen

3.1 Kapitelüberblick

Nach der Aufgabendefinition und der Bereitstellung der Daten kann nun mit deren Auswertung begonnen werden. Den ersten Überblick erhält man durch deskriptive Statistiken, auf denen weitere Kennzahlensysteme aufbauen. Solche Kennzahlen für Websites werden in diesem Kapitel beschrieben und es wird beurteilt, inwieweit sie sich zur Erfolgsmessung informationsorientierter Websites eignen.

Abb. 3.1 zeigt den Aufbau dieses Kapitels, der die zugrundeliegenden Kennzahlen nach ihrem Erhebungsort unterscheidet. Diese können entweder auf Seiten des Servers durch implizites Beobachten des Users und Crawlen der Website oder auf der Client Seite durch explizites Erfragen beim User gesammelt werden. Außerdem besteht serverseitig die Möglichkeit der inhaltlichen und strukturellen Analyse einer Website. Die Möglichkeiten zur Beobachtung des Users von der Serverseite aus sind eingeschränkt. Darauf wurde in Kap. 2.2.3 bezüglich Tabbed Browsing bereits hingewiesen. Vollständige Informationen über den User sind auf der Clientseite verfügbar. Da diese Arbeit auf serverseitigen Erkenntnissen beruhen soll, werden die Möglichkeiten auf der Clientseite nur kurz beschrieben.

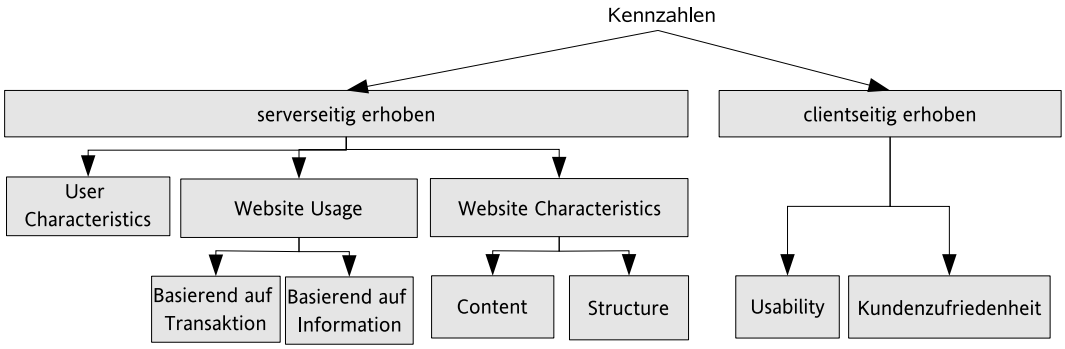


Abbildung 3.1: Kapitelüberblick

3.1.1 Dimensionen der Web Evaluation

Internet-Auftritte sind nach Riemer et al. [201] für die Marketingaktivitäten und den Kundenkontakt eines Unternehmens auch noch im Jahr 2006 von zunehmender Bedeutung. Websites stellen eine Schnittstelle zum Kunden dar, deren Erfolg durch die wachsenden Online-Umsätze und die steigende Zahl von Internet-Nutzern verstärkten Einfluß auf den gesamten Geschäftserfolg eines Unternehmens haben kann. Riemer et al. fordern ein gezieltes Qualitätsmanagement, um die angestrebte Güte der Website als Kundenschnittstelle sicherzustellen. Sie begründen dies mit einer Studie aus dem Jahr 2004 [82], die für die Hälfte von rund 200.000 Web Angeboten ein erhebliches Verbesserungspotential festgestellt hat.

Web-Evaluations . . .					
. . . Objekte	. . . Kriterien	. . . Subjekte	. . . Modus	. . . Horizont	. . . Zeitpunkt
Interne: - Technische Systeme - Prozesse Externe: - Eigene Website - Wettbewerberwebsites - Branche	Objektive: - Effizienz: z.B. Zeit - Vorhandensein von Funktionen - Anzahl Fehler Subjektive: - Effektivität - Qualität - Image - Gefallen	- Kunden - Neutrale Personen - Experten - Mitarbeiter - Technische Systeme (autom. Evaluation)	- Fragebogen, Offline-Interview - Laborexperiment - Automat. Evaluation: Sammlung technischer Daten	- einmalig - langfristig, kontinuierlich	- Neupositionierung der Werbeaktivitäten - vor Markteintritt - Anforderungsanalyse - Qualitäts-Check vor Start - Monitoring während Betrieb

Tabelle 3.1: Dimensionen der Web Evaluation, Riemer et al.[201], Totz et al.[252]

Der Forschungsansatz dieser Dissertation bezieht sich auf die grau hinterlegten Bereiche in Tabelle 3.1. Das Objekt der Untersuchung ist die **Eigene Website** eines Unternehmens, deren **Effektivität** und **Effizienz** gemessen werden soll. Eine **automatisierte Evaluation** soll ein **kontinuierliches Monitoring** und damit einen fortlaufenden Verbesserungsprozeß ermöglichen.

Um den Erfolg von Unternehmen im allgemeinen zu messen, stehen eine Reihe von Maßzahlen zur Verfügung. Die Kennzahlen spiegeln in aggregierter Form oder im Detail den Geschäftsverlauf wider und erlauben so als Steuerinstrument, Stärken und Schwächen zu identifzieren und damit steuernd einzugreifen. Auf welche Weise ein positiver Geschäftswertbeitrag einer Website entsteht, hängt, wie wir in Abschnitt 1.1.4 beschrieben haben, vom Internet Geschäftsmodell ab.

Den Erfolgsbeitrag einer Website zum Ergebnis eines Unternehmens zu bemessen, erhöht laut Kaufman et al. [142] und Kleist [149] die Komplexität der unternehmensweiten Erfolgsmessung. Die Firma DuPont hat ein bekanntes Kennzahlensystem entwickelt, das aus Daten der Bilanz- und Erfolgsrechnung als oberste Aggregationsstufe den Return on Investment (ROI) als Spitzenkennzahl berechnet [168].

Im Gegensatz zu Rechnungsgrößen, wie dem ROI oder Return on Equity (ROE), sind bei E-Commerce Projekten finanzielle Maße selten verfügbar. Anand et al. stellen in [10, S.24] das Problem der Berechnung des Return on Investment für Web Analysen heraus. Die Kosten sind recht genau ermittelbar, die Erlöse lassen sich dagegen ursächlich kaum den Web Aktivitäten zurechnen. Dieses Messungs- und Zuordnungsproblem wird bei Unternehmen mit mehreren, parallelen Vertriebswegen noch verschärft, wie Wade et al. in [256] erläutern.

Zu den Dimensionen, die E-Commerce Metriken abdecken sollen, zählen Zhu et al. [275] Information, Transaktion, Kundenorientierung und Anbindung der Lieferverbindungen. Die Arbeit von Zhu et al. untersucht mittels einer Befragung von 260 Firmen die Fähigkeit von Web Metriken, Effekte des E-Commerce auf das Unternehmensergebnis aufzeigen zu können. Sie fanden heraus, daß neben den direkten Online Erlösen traditionelle Industrieunternehmen E-Commerce eher als Kostenfaktor sehen, wohingegen IT-Unternehmen sich der Ertragschancen des Vertriebs- und Informationskanals Internet stärker bewußt sind. Dies ist ein Beleg dafür, daß die Erfolgsmessung auch bei informationsorientierten Websites relevant ist, da hier kein direkt meßbarer Nutzen durch den Verkauf von Produkten oder Dienstleistungen erzielt wird. Nur nachweisbarer Erfolg kann die Investitionen rechtfertigen.

3.1.2 Systematik serverseitiger Web Metriken und Maße

Kennzahlen verdichten Informationen und können komplexe, quantitative Sachverhalte in konzentrierter Form abbilden [213, S.3]. Schmitt [209] weist darauf hin, daß Kennzahlen im Internet, insbesondere zu Beginn der Internet-Ära, nicht einheitlich definiert waren. Laut einer Forrester-Studie [209] sollen die gleichen Kennzahlen, die mit unterschiedlicher Auswertungssoftware berechnet wurden, Abweichungen von 100 % auf der gleichen Datenbasis aufgewiesen haben. Daher wird in diesem Kapitel genau festgelegt, wie die einzelnen Kennzahlen berechnet werden.

Schwickert et al. beschreiben in [213, S.3] das Problem des nicht vorhandenen mathematischen Zusammenhangs zwischen vielen Web basierten Kennzahlen. Dadurch konnte ein hierarchisches Kennzahlensystem, wie das von DuPont, bislang nicht erstellt werden. Schwickert et al. [213] systematisieren erstens nach absoluten Zahlen, die sich direkt aus den Rohdaten ermitteln lassen, zweitens nach Kombinationen dieser Werte als Verhältniszahlen und drittens nach der Einbeziehung einer zeitlichen Komponente.

In Abb. 3.2 wurden Web Kennzahlen anhand deren Anwendungsgebiete zu einer Systematik zusammengefaßt.

Die Maßzahlen werden, wie in Abb. 3.2 zu sehen ist, in Maße zur Beurteilung von Usern, in Maße der Benutzung der Website sowie in Maße über die Eigenschaften einer Website unterteilt. Die User-Kennzahlen werden unterteilt in: 1. Kennzahlen, die auf alle Arten von Websites anwendbar sind (Kapitel 3.3.1), 2. Kennzahlen speziell für transaktionsorientierte Websites (Kapitel 3.3.2) und 3. Kennzahlen für informationsorientierte Websites (Kapitel 3.3.3).

Es werden Kennzahlen zur Beurteilung der Struktur (Kapitel 3.4.3), sowie des Inhalts (Kapitel 3.4.2) beschrieben. Danach werden zusammengesetzte Kennzahlen erläutert, die Usage-, Content- und Structure-Maße kombinieren. Abschließend wird die Einsatzfähigkeit der Kennzahlen zur Erfolgsmessung informationsorientierter Websites beurteilt.

3.2 Kennzahlen über die Benutzer einer Website

Wie in Abb. 2.3 in Kapitel 2.2.1 bereits erläutert, unterscheiden Anand et al. [10, S.26] zwischen **explizit** und **implizit** erhobenen Daten. Es muß auch unterschieden werden, ob ein User als wiederkehrender User erkannt werden kann oder ob jede Session losgelöst voneinander betrachtet werden muß. Die durch Beobachtung gewonnenen, d.h. implizit erhobenen Daten sind unabhänging von einer Registrierung für alle User verfügbar.

3.2.1 Implizite Erhebung, ohne Registrierung

Ein User benutzt eine Website zunächst anonym, wie in Kapitel 2 beschrieben. Je nach den Sicherheitseinstellungen des Client Browsers, können Daten über dessen Hard- und Softwarekonfiguration per Javascript gesammelt werden, wie:

- Hersteller und Version der Browser Software
- Verwendetes Betriebssystem
- Akzeptanz von Cookies
- Aktivierung von JavaScript
- Referrer-Information. Der Referrer eines Webpage-Aufrufs gibt Auskunft über die Webpage, von der aus auf die aufgerufene Webpage verlinkt wurde. Durch den Referrer kann beispielsweise die Effektivität von Banner-Werbung oder Suchmaschinenergebnisse überprüft werden. Man kann durch den Referrer feststellen, welches Banner ein User ausgewählt hat, durch das er auf die Website geleitet wurde.

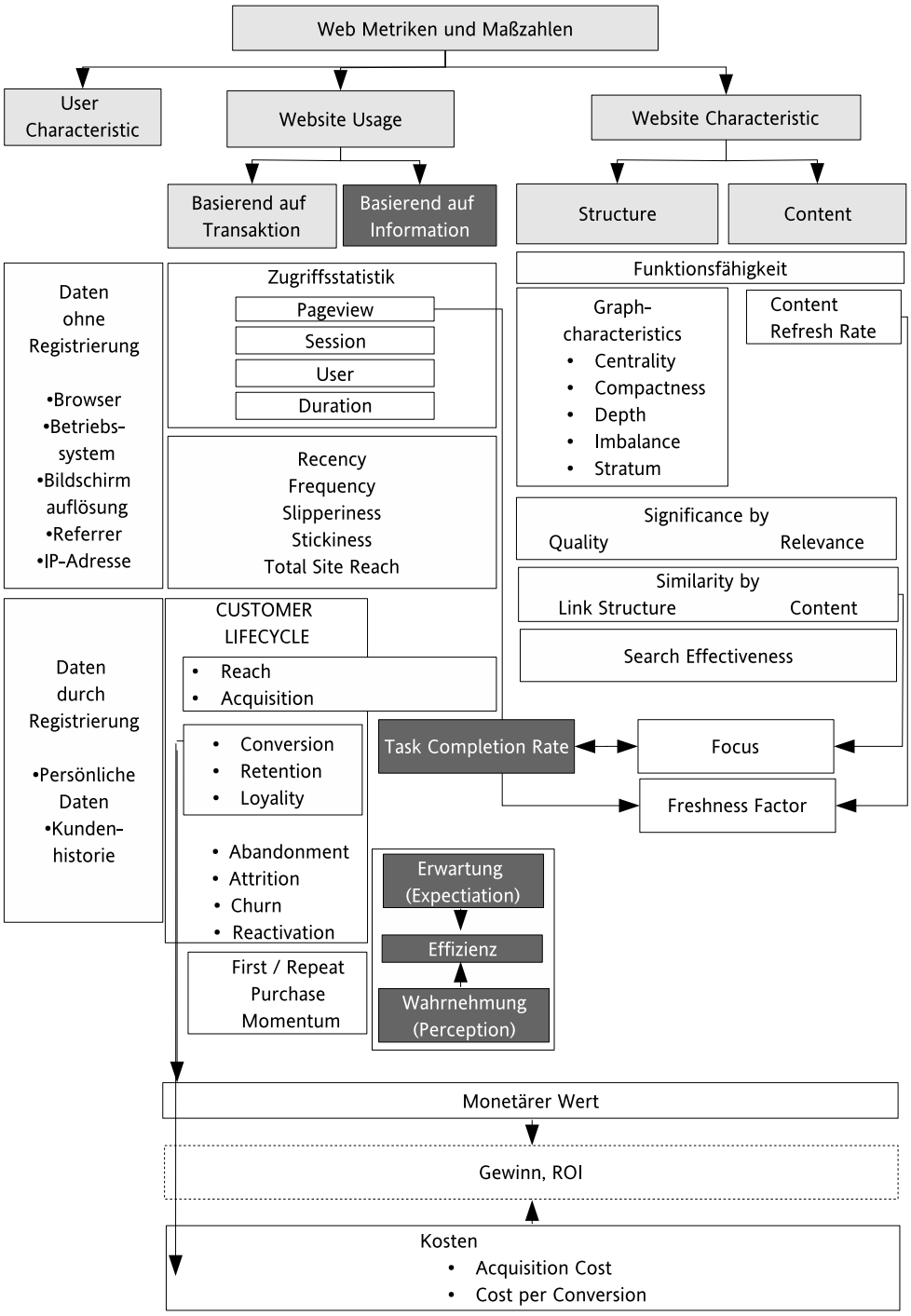


Abbildung 3.2: Web Metriken-Systematik, angelehnt an [19, 47, 74, 84, 107, 213]

- IP-Adresse. Durch die IP-Adresse und deren Rückübersetzung in sprechende Namen kann festgestellt werden, wie der User auf das Internet zugreift: beispielsweise über den Internetzugang eines Unternehmens oder als Privatnutzer über einen Internet Service Provider (ISP). In vielen Fällen kann über den DNS Eintrag (Domain Name Service) auch das Land, die Region oder sogar die Stadt, von wo aus der Zugriff auf das Internet stattgefunden hat, ermittelt werden.

Aggregiert man diese Informationen, können Kennzahlen über die Zusammensetzung der User erstellt und mit anderen Kennzahlen kombiniert werden.

3.2.2 Explizite Datenerhebung, mit Registrierung

Registriert sich der User, gibt er meist persönliche Informationen preis und kann bei erneutem Login eindeutig identifiziert werden. Die Kennzahl *Unique User* ist bei sich registrierenden Usern sehr präzise. Dagegen wird ohne Registrierung beispielsweise beim Einsatz von Cookies die Kennzahl *Unique Users* konsequent überschätzt. Sobald Cookies gelöscht werden, kann ein eigentlich wiederkehrender User nicht als solcher erkannt werden und wird erneut als Unique User gezählt. Die Kennzahl *Users* bzw. *Unique Users* wird in Kapitel 3.3.1 genauer beschrieben.

Der Umfang und die Qualität der verfügbaren Informationen hängt von den durch die Website abgefragten Daten, der Auskunftsbereitschaft und Ehrlichkeit der User ab. Zu den möglichen Daten gehören zum Beispiel:

- Persönliche Angaben: Name, Adresse
- Interessen

Über eine längeren Zeitraum kann bei erneut eingeloggten Usern eine Kundenhistorie erstellt werden. Durch die Identifizierbarkeit der registrierten User ist es möglich, die User- und Usage-Daten mit externen Daten z.B. aus einem CRM-System oder Geoinformationen zu kombinieren, wie Wilde et al. in [124, S.14f] beschreiben.

Abgeleitete Kennzahlen werden von Schwickert et al. in [213, S.11] aufgelistet. Unter anderem:

- Anteil der registrierten User an der Gesamtzahl der User

Hier muß man sich wiederum bewußt sein, daß die Zahl an nicht registrierten Usern überschätzt wird, während die Zahl der registrierten User präzise ist.

3.3 Kennzahlen über die Benutzung einer Website

Im Gegensatz zu Schwickert et al. [213, S.9] und Bensberg [19] werden hier Clickstreams nicht den Kennzahlen bzw. Web Metriken zugeordnet, sondern als Teil des Web Mining in Kapitel 4 betrachtet.

3.3.1 Allgemeine Web Usage-Kennzahlen

Mit den ersten Websites wurden Zähler eingeführt, welche die Anzahl der beim Server angefragten Webpages gemessen haben. Wie Anand et al. in [10, S.27f] weiter zusammenfassen, waren die zweite Generation an Meßinstrumenten Werkzeuge, die die Logfiles der Server ausgewertet haben und dabei häufig auf Hits basierten. In Kapitel 2 wurde die Datengewinnung aus Logfiles und anderen Quellen beschrieben. Darauf baut dieses Kapitel auf. Die nach Säuberlich [207, S.110] verbreitetsten Maße zum Vergleich von Websites sind die Anzahl an **Page Views**.

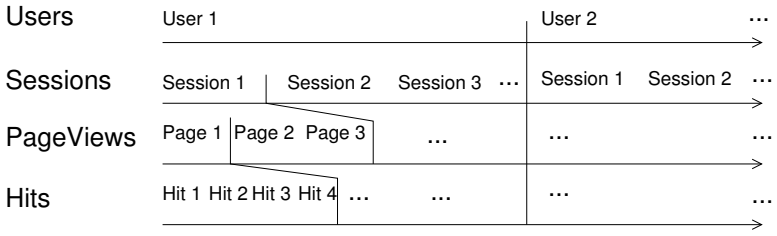


Abbildung 3.3: Aggregationsstufen von Hits zu Usern; nach Säuberlich [207, S.109]

User, Session, Page View, Hit Anand et al. unterscheiden in [10, S.28f] wie folgt:

- **Unique Users**: Die Gesamtzahl an unterschiedlichen User-Identifikationen. Mehrfach vorkommende User IDs werden nur einfach gezählt. So werden wiederkehrende User identifiziert. Die einzige Möglichkeit, User zuverlässig als einzelne, reale Personen zu identifizieren, bietet eine Registrierung. Persistente Cookies stellen nur eine Heuristik dar, wenn man unterstellt, daß an einem Browser immer der gleiche User sitzt und dieser die Cookies nicht löscht.
- **Sessions**: sind inhaltlich und zeitlich zusammenhängende Nutzungsvorgänge, d.h. Page Views, die einem User zugeordnet werden können. Zwischen den Page Views dürfen keine längeren Pausen liegen, sonst wird nach einem festgelegten Zeitraum (z.B. 30 Minuten) die alte Session als abgeschlossen betrachtet und eine neue Session angefangen.
- **Page Impression**: Ein User bzw. dessen Browser fordert vom Webserver z.B. per Click auf einen Link eine Webpage an. Diese User-Aktion wird als *Page View*, *Page Impression*, *Click* oder *Event* bezeichnet.
- **Hits**: Hits sind die einzelnen Dateien, die ein Webserver aufgrund der Anfrage eines Clients an diesen User sendet. Die vom User angeforderte Webpage besteht aus mehreren Elementen, die jeweils einen Eintrag im Logfile des Webservers verursachen, sogenannte Hits. Da jede Webpage aus einer unterschiedlichen Anzahl an Elementen wie Texten oder Grafiken besteht, unterscheidet sich dadurch auch die Zahl der Hits. Der Kritik von Säuberlich [207, S.109] an Hit-basierten Statistiken folgend, werden Hits in dieser Arbeit als Kennzahl nicht weiter berücksichtigt.

Wie man in Abbildung 3.3 sieht, kann einem User eine oder mehrere Sessions zugeordnet werden, die jeweils mindestens einen Page View besitzen, der wiederum so viele Hits generiert, wie die jeweilige Webpage an Elementen besitzt.

Duration Neben Page Views, Sessions und Usern, stellt die Dauer zwischen zwei User-Aktionen, die **Duration**, die vierte Standardkennzahl dar. Ist eine Session eindeutig identifizierbar, kann der zeitliche Abstand zwischen zwei Page Views gemessen werden. Dieser zeitliche Abstand wird als Duration bezeichnet. Siehe dazu Abb. 2.5 in Abschnitt 2.2.3.

Die Duration auf der zuletzt aufgerufenen Webpage einer Session kann nicht gemessen werden, da nach dem letzten Page View keine User-Aktion mehr aufgezeichnet wird. Nach dem letzten aufgezeichneten Page View bleibt es dem Tracking-Mechanismus unbekannt, was der User getan hat.

Folgende User-Aktionen sind denkbar:

- Der User hat über einen externen Link die Website verlassen. Es ist technisch möglich, die Benutzung externer Links mitzuverfolgen. Auf diese Weise kann der letzte Page View innerhalb der Website gemessen werden.
- Die Session des Users endet wegen Überschreitung der maximalen Duration zwischen zwei Page Views, sodaß eine neue Session gestartet wird. Es bleibt unbekannt, was der User mit der zuletzt aufgerufenen Webpage gemacht hat - er kann sie vollständig gelesen haben oder sofort geschlossen haben. Um die letzte Aktion eines Users dennoch nicht zu verlieren, wird die vorsichtige, systematisch zu niedrige Schätzung von 1 Sekunde angenommen - die minimale Duration. Trotz dieser Problematik ist der letzte Click innerhalb einer Session sehr wichtig, da hier der Grund liegt, warum der User seine Session nicht fortgesetzt hat. Entweder hat er gefunden, wonach er gesucht hat und verläßt zufrieden die Website, oder das genaue Gegenteil ist der Fall.
- Der User gibt manuell eine andere Website vor, durch Direkteingabe der URL oder Bookmarks. Hier gibt es keine Möglichkeit, diese Aktion zu erfassen. Auch hier wird für die zuletzt aufgerufene Webpage eine minimale Duration von 1 Sekunde angenommen.

Die vier Standardkennzahlen, Page View (Click), Session, User und Duration können für jede Webpage oder die gesamte Website aggregiert werden. Genauso können Durchschnitte jeder Standardkennzahl auf der Basis der anderen erstellt werden, wie von Pitkow et al. in [194] vorgeschlagen, beispielsweise:

- Page Views per Session
- Sessions per User
- Duration per Session
- Duration per User

Page View, Session, User und Duration bilden die unverzichtbare Grundlage jeder Website-Statistik. Sie sind der erste Indikator für die Nutzung einer Website. Aus deren zeitlichem Verlauf können Effekte einzelner Kampagnen oder Website-Änderungen erkannt, aber nicht unbedingt der jeweiligen Aktion ursächlich zugeordnet werden. Will man die Ursachen des User-Verhaltens verstehen, sind diese Kennzahlen nur ein erster aber nicht ausreichender Schritt. Zusammenhänge und Ursachen des User-Verhaltens, sowie Struktur und Inhalt der Website, bleiben unbeachtet. Diese Kombination ist nur durch Web Mining-Analysen zu erreichen.

Recency beschreibt den Zeitraum seit der letzten Session eines Users. Die Firma NetGenesis hat in einer Übersicht von Web Metriken [74, S.36] Recency folgendermaßen definiert: ”Recency beschreibt, wie lange es her ist, seit ein User das letzte Mal die Website aufgesucht hat.” Daraus kann auf die Attraktivität der Website und ihrer Inhalte geschlossen werden. Allerdings mußten die individuellen Präferenzen eines Users bekannt sein, um beurteilen zu können, inwieweit die Website diesen entsprechen konnte und ob Potential zu Verbesserungen besteht. So werden Nachrichten-Websites regelmäßig besucht, Websites des Prüfungsausschusses einer Universität dagegen nur bei akuten Fragen.

Frequency beschreibt, wie häufig User innerhalb eines festen Zeitraums die Website besucht haben.

$$Frequency = \frac{Sessions(Visits) Visits\ innerhalb\ Zeitperiode\ t}{Unique\ Users\ innerhalb\ Zeitperiode\ t}$$

Frequency und Recency ergänzen sich bei der Beurteilung der Attraktivität einer Website und ihrer Webpages. Die Loyalität ist bei häufig wiederkehrenden Besucher besonders hoch. Auf transaktionsorientierten Websites in Kap. 3.3.2 wird dafür die spezielle Kennzahl *Loyality* benutzt.

Stickiness bemißt die Fähigkeit eines Webpage-Inhalts, die User und im positiven Fall auch deren Aufmerksamkeit auf der Webpage zu binden [74, S.47].

$$Stickiness = Frequency \quad Duration \quad Total\ Site\ Reach$$

oder vereinfacht:

$$Stickiness(Webpage\ p) = \frac{Gesamtduration\ auf\ Webpage\ p}{Zahl\ Unique\ Users\ auf\ Webpage\ p}$$

Slipperiness bewertet das umgekehrte Verhalten wie Stickiness und wird auf die gleiche Art berechnet. Bei manchen Seiten wird eine geringe Stickiness angestrebt, sodaß User möglichst wenig Zeit auf einer Seite verbringen und zum Beispiel schnell weitergeleitet werden sollen. Der Bestell- und Bezahlvorgang soll beispielsweise möglichst einfach und schnell durchlaufen werden, damit die Abandonment-Rate möglichst gering bleibt. Slipperiness ist ein möglichst geringer Stickiness-Wert [107, S.243].

Beide Kennzahlen können nützlich sein, wenn man Eigenschaften von Webpages unterschiedlichen Typs beurteilt. So ist es für eine informationsorientierte Website wünschenswert, den User durch Navigationsseiten schnell zum Ziel, d.h. zu Contentseiten zu leiten. Also kann das

Ziel einer informationsorientierten Website in einer hohen Slipperiness von Navigationsseiten und einer hohen Stickiness von Contentseiten liegen.

Acquisition mißt die Zahl der User, die an den Angeboten des Unternehmens Interesse zeigen, z.B. indem sie einen Newsletter abonnieren, an einer Befragung teilnehmen, sich für ein Whitepaper registrieren, oder einen Katalog anfordern [74, S.28].

Acquisition ist Teil des Kundenbeziehungslebenszyklus, der in 3.3.2 beschrieben wird. In diesem Sinne ist Acquisition eine transaktionsorientierte Kennzahl, die in ihrer Aussagefähigkeit als Kennzahl nicht auf eine später stattfindende Transaktion angewiesen ist.

Site Penetration Rate wird von Gaul et al. [107, S.242] wie folgt definiert:

$$Site\ Penetration\ Rate = \frac{Click\ Throughs\ to\ First\ Interior\ Page}{PageViews\ auf\ Homepage}$$

Leider beschreiben Gaul et al. nicht, was eine Interior Page auszeichnet. Ohne Berücksichtigung der Linkstruktur einer Website hat das Verhältnis einer Webpage zur Homepage wenig Aussagekraft. Eine Clickstreamanalyse (siehe Kap. 3.3.2) scheint hier geeigneter.

Click-Through-Rate ist eine Kennzahl über die Benutzung eines bestimmten Elements, z.B. eines Banners, durch das die User zu einer anderen Webpage geleitet werden sollen. Click-Through zählt beispielsweise die Clicks auf ein Banner [213, S. 10f] und setzt sie ins Verhältnis zur Größe der Zielgruppe.

$$ClickThroughRate = \frac{Click \quad Throughs}{Reach}$$

Diese Kennzahl mißt den Erfolg des Banners, aber nicht den Erfolg der Website, auf die der Banner verlinkt.

Reach beschreibt das Potential an Usern und Kunden, die man zu erreichen versucht - die Zielgruppe. Die maximale Höhe von Reach besteht in der Zahl aller User und Kunden, die technisch in der Lage sind, über das Internet mit dem Unternehmen in Kontakt zu treten [74, S.27]. Im Falle eines maximalen Reachs sind das alle Menschen, die einen Internetanschluß zur Verfügung haben. Verwandt zu Reach ist die Kennzahl *Total Site Reach*

Total Site Reach wird von NetGenesis [74, S.48] sehr speziell formuliert: indem der Anteil der User in einem gewissen Zeitraum t im Verhältnis zur Gesamtzahl der User gesetzt wird:

$$Total\ Site\ Reach = \frac{Unique\ Users\ im\ Zeitraum\ t}{Gesamtzahl\ Unique\ Users}$$

Bei dieser Zahl wird vorausgesetzt, daß die User über einen längeren Zeitraum voneinander abgrenzbar sind. Ohne Registrierung ist dies nur möglich, wenn permanente Cookies nicht zwischenzeitlich gelöscht wurden.

Um das Potential an möglichen Usern festzulegen, bedarf es eher traditioneller Marktforschung. Auch die Unsicherheit in Bezug auf den langfristigen Gebrauch permanenter Cookies trägt nicht zur Zuverlässigkeit beider Kennzahlen bei. Beide Kennzahlen können die Zahl der angelockten User beschreiben, sagen aber nichts über deren Zufriedenheit aus.

Entry Exit Page ist im eigentlichen Sinne keine Kennzahl, sie beschreibt, welche Seite am Anfang einer Session steht (Entry Page) und welche Seite die letzte Webpage war (Exit Page), bevor der User die Website verlassen hat. Webpages werden zu Einstiegsseiten durch Links von anderen Websites, Bookmarks, Direkteingabe der URL oder Fortsetzung einer durch Timeout beendeten vorherigen Session.

Die Exit Page ist von großem Interesse, da der User danach seine Session beendet. Der Grund für das Session-Ende liegt mit großer Wahrscheinlichkeit auf der Exit Page. Entweder in einem externen Link, sodaß der User die Website verläßt, um dort seine Ziele weiterzuverfolgen, oder er hat gefunden, wonach er gesucht hat und beendet die Session. Oder der User kann sein Ziel nicht erreichen und hat die Website unzufrieden verlassen.

3.3.2 Transaktionsorientierte Web Usage-Maßzahlen

Customer Lifecycle

Transaktionsorientierte Websites zielen auf den Abschluß einer Transaktion, in der Produkte oder Dienstleistungen dem User bzw. Kunden verkauft werden. Mit Abschluß der Transaktion

gibt der User durch Registrierung seine Anonymität auf. Dadurch kann der Erfolg und dessen Höhe gemessen und einem Kunden zugeordnet werden. Abweichungen können entstehen, wenn ein Kunde sich mehrfach registriert und nicht mehr als eine physische Person erkannt werden kann. Auch bleibt unbekannt, ob mehrere Personen einen Zugang bzw. Registrierung gemeinsam nutzen.

Um das eventuelle Zustandekommen einer Transaktion lassen sich mehrere Kennzahlen erstellen, die sowohl den kompletten Kaufprozeß als auch die Entwicklung des Kunden im Kundenbeziehungslebenszyklus nach Stauss [233] beschreiben. Es kann sowohl der Erfolg als auch Mißerfolg analysiert und deren jeweilige Höhe, Abbruchzeitpunkt und -grund untersucht werden.

In Abb. 3.4 wird der Zusammenhang mehrerer transaktionsorientierter Kennzahlen des Kaufprozesses verdeutlicht, die gemeinsam einen Kundenlebenszyklus beschreiben. Die Kennzah-

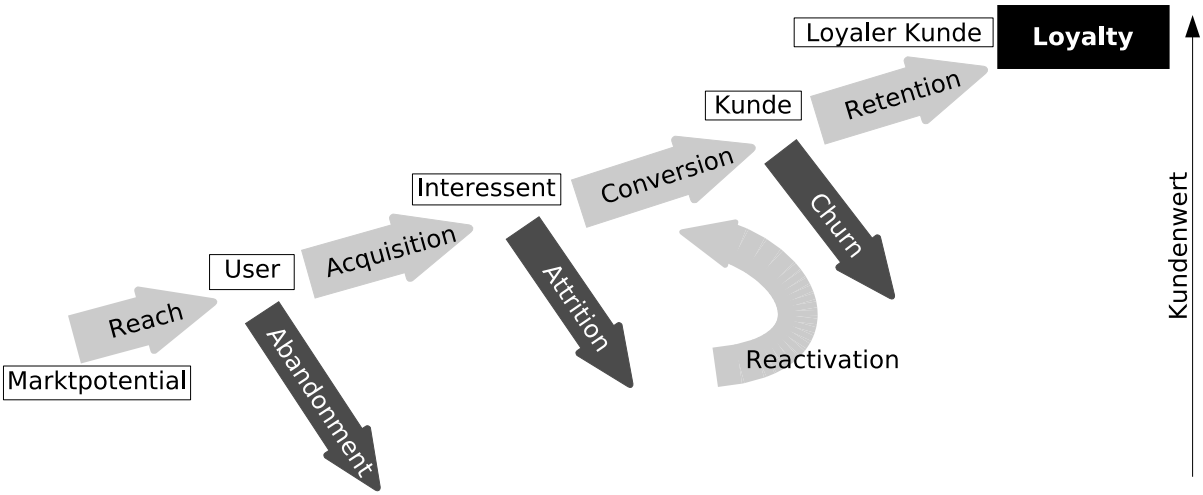


Abbildung 3.4: Kundenbeziehungslebenszyklus; nach [74, Abb.:14] [233]

len beschreiben, wie aus dem anonymen User einer Website durch **Acquisition** ein *Interessent* wird. Durch eine erfolgreiche Transaktion (**Conversion**) wird der *Interessent* zum *Kunden*. Bei regelmäßiger Wiederholung (**Retention**) von Transaktionen auf der Website wird aus einem *Kunden* ein *loyaler Kunde*.

Reach und **Acquisition** bilden den Teil des Customer Lifecycle, der nicht von einer Transaktion abhängig ist und bereits oben in 3.3.1 beschrieben wurde.

Conversion Wie NetGenesis in [74] hervorhebt, ist der Begriff *Conversion* sehr heterogen besetzt. Conversion bezeichnet den Zeitpunkt, zu dem aus einem Interessenten ein Kunde

geworden ist. Dazu muß, je nach Auffassung, ein Interessent nicht unbedingt ein Produkt kaufen. Der Zeitpunkt der Conversion kann auch die Erstregistrierung, die Einrichtung eines persönlichen Profils oder die Benutzung eines Produktkonfigurators sein. In jedem Fall bedarf es einer expliziten User-Aktion - Kauf oder Registrierung - damit der Interessent als Kunde gewertet werden kann.

Retention bezeichnet nach [74, S.30] den Zeitpunkt eines wiederholten Kaufs durch eine Kunden. Die Anzahl der notwendigen Wiederholungskäufe, bis ein Kunde als "retained" angesehen werden kann, hängt von der jeweiligen Marktsituation des Internet-Anbieters ab. Der Kunde erreicht dann den Status eines loyalen Kunden.

Loyality Loyale Kunden haben über die normale Geschäftsbeziehung hinaus eine intellektuelle oder emotionale Bindung an das Unternehmen bzw. dessen Website [232]. Die Loyalität drückt sich in einer hohen *Frequency* und kurzen *Recency* dieser User aus. Ein loyaler Kunde kann, wie bei Stauss [232, 233] beschrieben, auch wieder auf den Status eines normalen Kunden zurückfallen oder komplett abwandern. Die nachfolgenden Kennzahlen beschreiben dieses Abwanderungsverhalten.

Abandonment beschreibt das Abbrechen eines Kauf- oder Registrierungsvorgangs. Dadurch wird verhindert, daß aus einem User ein Interessent wird. Typisch für Internetseiten ist laut [74, S.34] das Füllen eines Warenkorbs und der spätere Abbruch des Kaufvorgangs. Dieser Vorgang ist außerhalb des Internets, z.B. in einem Supermarkt, sehr ungewöhnlich. Der Grund kann in einer schlechten Navigierbarkeit, dem Design der Website oder in einem zu langwierigen, komplizierten Bestell- und Bezahlvorgang liegen.

Attrition ist der Anteil der Käufer, die keine Wiederholungskäufe mehr tätigen oder anderweitig den Kontakt zu dem Unternehmen eingestellt haben. Diese ehemaligen Käufer können durch geeignete Maßnahmen reaktiviert werden (Reactivation).

Churn mißt den Anteil der abgewanderten Kunden im Verhältnis zur Gesamtzahl der Kunden.

$$Churn\ Rate = \frac{attrided\ customers}{total\ number\ of\ customers}$$

Monetärer Wert Nach NetGenesis [74, S.38] ist es nur möglich, einen monetären Wert zu bestimmen, wenn ein Kauf - sprich eine Transaktion - durchgeführt wurde. Dies entspricht der hier getroffenen Unterscheidung zwischen informationsorientierten und transaktionsorientierten Websites in Kapitel 1.

Acquisition Cost

$$Acquisition\ Cost = \frac{Kosten\ für\ Marketing\ und\ Kampagne}{Anzahl\ an\ Click \quad Throughs}$$

Cost per Conversion

$$Cost\ per\ Conversion = \frac{Kosten\ für\ Marketing\ und\ Kampagne}{Anzahl\ an\ Verkäufen}$$

First Purchase Momentum

$$First\ Purchase\ Momentum = \frac{Minimum\ an\ Clicks\ bis\ zum\ ersten\ Kauf}{tatsächliche\ Clicks\ bis\ zum\ ersten\ Kauf}$$

Repeat Purchase Momentum

$$Repeat\ Purchase\ Momentum = \frac{Minimum\ an\ Clicks\ bis\ zum\ wiederholten\ Kauf}{tatsächliche\ Clicks\ bis\ zum\ wiederholten\ Kauf}$$

Teltzrow und Berendt [248, 249] unterscheiden bei den klassischen Conversion Rates eine Makro- und eine Mikroebene. Die Makroebene entspricht dem Kundenbeziehungslebenszyklus , wie er von Stauss in [233, 232] beschrieben und bereits von Cutler et al. in [74] ausführlich als Web Metrik beschrieben wurde. Eine Conversion Rate auf Mikroebene wurde auch von Lee et al. [159] vorgeschlagen, die den Kundenkaufprozeß [134] beschreibt. Die Mikro-Übergangsraten nach Lee et al.[159] und Schonberg et al.[210] sind:

- Look-to-Click
- Click-To-Basket
- Basket-To-Buy
- Look-to-Buy

Weitere Kennzahlen, die sich auf die Kauffrequenz, die Ordergröße und den Kaufvorgang beziehen, werden bei Gaul et al. in [107, S.241ff] genannt.

- Sales per Visitor
- Order Rate of Repeat Customer
- Average Order Size

Die transaktionsorientierten Kennzahlen beschränken sich in ihrer Anwendbarkeit auf transaktionsorientierte Websites. Auf informationsorientierten Websites stehen ihrer Anwendung zwei Grundprobleme entgegen:

- Es gibt keine Transaktionen, an denen sich Erfolgskennzahlen orientierten könnten und durch die sich die Höhe des Erfolgs quantifizieren ließe.
- Es besteht kein Zwang oder kein Anreiz zur Registrierung. Dadurch bleiben User meist anonym. Eine Registrierung ist nur dann durchsetzbar, wenn dem User durch sie ein Nutzen erwächst. Bei Unternehmenswebsites, die Informationen frei zugänglich zur Verfügung stellen, ist es kaum möglich, den User zu einer dauerhaften Registrierung zu bewegen. Dies wäre bei speziellen Inhalten für einen ausgesuchten User-Kreis denkbar, beispielsweise bei Informationen für Kooperationspartner eines Unternehmens.

3.3.3 Informationsorientierte Web Usage-Maßzahlen

Über die allgemein anwendbaren Kennzahlen zur Beschreibung der Benutzung von Websites hinaus gibt es nur wenige Kennzahlen, die speziell für informationsorientierte Websites erstellt wurden.

Task Completion Rate vorgeschlagen von Brinck et al. [38], basiert auf der Identifikation bestimmter Aufgaben und Ziele, die ein User auf einer Website verfolgen kann. Der Clickstream eines Users wird dabei mit dessen verfolgtem Ziel abgeglichen. Das Ziel wurde bei Brinck et al. zu Testzwecken vorgegeben und daraufhin der User beobachtet.

Die vorgeschlagene Kenngröße ist gut geeignet, den Erfolg von informationsorientierten Websites aus User-Sicht zu beschreiben. Man steht allerdings vor dem Problem, daß man auf der Serverseite nicht weiß, welche Ziele ein User verfolgt [210]. Im nächsten Kapitel 4 soll deshalb untersucht werden, ob das User-Interesse durch Web Mining-Methoden erkannt werden kann. Auf das Konzept der Task Completion Rate wird bei der späteren Erstellung eines Erfolgsmaßes in Kapitel 5 in Abschnitt 5.4 zurückgegriffen. Auch die User-Studie in Kapitel 5.3 berechnet eine Task Completion Rate, indem der User unter Vorgabe eines Ziels beobachtet wurde.

Web Service Quality Pather et al. beschreiben in [188, S.143ff] die Verbindung zwischen der Zufriedenheit der User und der Service Qualität, die eine Website als integraler Bestandteil für ein Unternehmen und die Organisation der Geschäftsprozesse spielt.
Pather et al. schlagen vor, sowohl die User Zufriedenheit als auch die Service Qualität zu messen, indem die **Erwartungen** der User vor der Benutzung der Website mit der **Wahrnehmung** der Servicequalität der Website nach dem Besuch verglichen werden, siehe Abb. 3.5. Die Lücke zwischen beiden bewertet die Service-Qualität.

Pather et al. passen die Instrumente zur Bestimmung der User-Zufriedenheit den Gegebenheiten des Internets an. Die Zufriedenheit der User wird als Maß für die Effektivität eines

Informationssystems genutzt. Pather et al. geben einen Überblick [188] von Studien zum Thema User-Zufriedenheit, die meist durch explizite Befragung der User erhoben wurden.

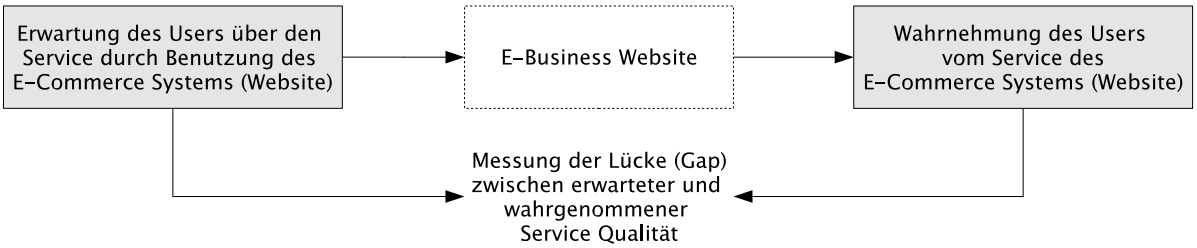


Abbildung 3.5: Effektivitätsmessung von E-Commerce Websites [188, Fig.3]

Als Dimensionen der traditionellen Messung von Service-Qualität nennen Pather et al. [188, S.150]: Tangibles, Reliability, Responsiveness, Assurance und Empathy, Trust, Content-Qualität. Die Relevanz dieser Dimensionen überläßt Pather weiteren empirischen Untersuchungen. Zuvor hat bereits Grönroos 1982 [109] ein Kundenqualitätsurteil aus dem Vergleich der Kundenerwartung mit der Kundenwahrnehmung beschrieben.

Web Service Quality fügt der Task Completion Rate die qualitative Dimension der User Wahrnehmung hinzu. Es wird weniger die Effektivität und Effizienz der Website bewertet, also die Erfüllung der User-Ziele und auf welche Art und Weise die Website den Vorstellungen des Users entsprochen hat. Die empfundene Qualität kann vom User nur explizit erfragt werden. Es darf bezweifelt werden, ob aus dem beobachtbaren Verhalten des Users auf der Server-Seite auf die von ihm wahrgenommene Qualität der Website geschlossen werden kann. Die Task Completion Rate dagegen verspricht einen präziser umzusetzenden Ansatz.

3.4 Kennzahlen über Charakteristika einer Website

3.4.1 Technische Funktionsfähigkeit der Website

Technische Maßzahlen, die die Funktionsfähigkeit und Verfügbarkeit einer Website beschreiben, beschäftigen sich mit der Überwachung der Web Server, die die Internetseiten bereitstellen. Freshwater [220, S.4] zählt die dazu notwendigen Funktionen einer Web Server-Überwachung auf:

- Sicherstellen der Funktionsfähigkeit aller Komponenten einer Website
- Schnelle und zielsichere Diagnose von auftretenden Problemen
- Performanzmessung zur Kapazitätenplanung [213]

- Mögliche Probleme und Kapazitätsengpässe

Beispiele findet man bei Pitkow in [194], bei Bensberg in [18, S.250ff] und Schwickert et al. [213, S.14f].

Bensberg et al. [18] zählen Maße über die Benutzerfreundlichkeit einer Website (Usability) zu den technischen Kennzahlen. Für den Fall von Übertragungsgeschwindigkeit und Dauer des Seitenaufbaus, wird dies hier ebenfalls so gesehen. Andere Usability-Kennzahlen, die die Präsentation des Inhalts und deren Struktur betreffen, werden hier nicht zu den technischen Kennzahlen gezählt.

Sind Teile oder die komplette Website nicht erreichbar, muß auch der Erfolg einer Website nicht weiter gemessen werden. Für unsere Untersuchung ist der Fall der Nichtverfügbarkeit von Websites oder einzelner Webpages nicht relevant, da hier mit großer Wahrscheinlichkeit keine User Aktionen stattfinden, die analysiert werden können. Nicht verlinkte Webpages können dann entweder nur durch Kenntnis der URL, gespeicherte Bookmarks oder Browser-Caches erreicht werden. Die Fähigkeit einer Website, den User durch eine intuitive Navigation zu Inhalten zu leiten, kann in diesem Fall nicht bewertet werden. Daher wird von einer technisch funktionsfähigen Website und von per Link erreichbaren Webpages ausgegangen.

3.4.2 Kennzahlen über den Inhalt einer Website

Mit Kennzahlen über den puren Inhalt einer Webpage oder der gesamten Website können keine Aussagen über den Erfolg der Website gemacht werden. In Kombination mit Usage-Daten kann aber der Einfluß des Inhalts auf das User-Verhalten deutlich gemacht werden.

Focus zählt die Page Views eines Users innerhalb eines bestimmten Bereiches einer Website. Dadurch wird gemessen, inwieweit die User einem Themenbereich besondere Aufmerksamkeit schenken.

Um diese Kennzahl berechnen zu können, müssen die Themenbereiche einer Website bekannt sein. In Kap. 4.3 wird auf die Themenidentifikation mittels Web Mining-Algorithmen eingegangen. Der Focus eines Users kann dazu benutzt werden, die Task Completion Rate zu berechnen und ist damit ein relevantes Maß zur Erfolgsbestimmung informationsorientierter Websites.

Freshness Factor mißt die Auswirkung der Erneuerung von Content auf die Attraktivität beim User.

$$Freshness\ Factor = \frac{durchschnittliche\ Content\ddot{U}berarbeitungsrate}{Frequency}$$

Da die konstante Erneuerung und Überarbeitung von Content kostenintensiv ist, mißt der Freshness Factor, inwieweit sich der Aufwand, Content aktuell zu halten, in einem gesteigerten User-Interesse widerspiegelt.

Ähnlichkeits- und Signifikanzmaße

Die Effektivität von Suchanfragen läßt sich durch zwei Maße beschreiben: **Precision** und **Recall**. Sei N die Zahl der gefundenen Dokumente in einer Suche nach einer Anfrage aus Keywords Q. Aus dieser Suche sind N' Dokumente relevant zu Q. R ist die Gesamtzahl der existierenden relevanten Dokumente, die die Suchmaschine theoretisch finden kann.

Precision ist der Anteil an Dokumenten, die die Suchanfrage zurückliefert, die relevant sind. $Precision = \frac{N'}{N}$

Relevance ist der Anteil der gefundenen relevanten Dokumente an der Gesamtzahl der relevanten Dokumente überhaupt. $Recall = \frac{N'}{R}$

Relevanzmaße werden nicht alleine auf dem Inhalt einer Webpage berechnet, es wird auch deren Umfeld in Form der verlinkten Webpages einbezogen und damit die Struktur der Website. Die Relevanz gibt die Übereinstimmung des Inhalts einer Webpage mit einer Suchanfrage in Keyword-Form, sprich einer Folge einzelner Wörter, an.

Most Cited weist einer Webpage p einen Relevanzwert zu, der der aufsummierten Anzahl der Suchwörter auf den Webpages entspricht, die auf p verlinken (citing). Dieser Algorithmus [84, S.477] weist referenzierten Dokumenten höhere Relevanzwerte zu als den referenzierenden Dokumenten.
Boolean Spread Activation basiert wie *Most Cited* auf der Hyperlink-Struktur ohne Einbeziehung der Auftrittshäufigkeit der Suchwörter. Es wird gemessen, ob ein Suchwort in der Webpage p auftritt. Außerdem wird das Vorkommen der Suchwörter in den p verlinkenden Webpages und den von p bereits verlinkten Webpages untersucht. Der Beitrag zum Relevanzmaß von benachbarten Seiten wird durch eine Konstante bestimmt.
TFxIDF und *Vector Spread Activation* basieren nach [84, S.476ff] auf einem *Vector Space Model* [12, S.82ff], das die Dokumente und Suchanfragen in Form von Wort-Vektoren repräsentiert. Siehe Kap.4.3.2.
Vector Spread Activation berücksichtigt keine Hyperlink-Informationen. Wie bei Boolean Spread Activation wird der Beitrag der benachbarten Dokumente durch Propagation der Relevanzwerte berücksichtigt. Die Relevanzwerte für eine Webpage p ergeben sich aus den propagierten Relevanzwerten der Umgebung und der wie in TFxIDF berechneten Relevanzwerte für p.

Ähnlichkeitsmaße dienen zum Vergleich von Webpages anhand ihres Inhalts untereinander oder im Verhältnis zu einer Suchanfrage. Sie eignen sich zur Beurteilung von Suchmaschinen und zur Evaluation der Suchalgorithmen. Geht man von einem bekannten User-Interesse aus, wäre es hilfreich, durch Ähnlichkeitsmaße die inhaltlich geeignetsten Webpages zu finden bzw. die Session danach zu beurteilen. Hierfür reicht ein Maß nicht aus. Es müßte für jeden Click separat berechnet werden und gegebenenfalls zu einem neuen Maß aggregiert werden, um die Übersichtlichkeit zu wahren.

3.4.3 Kennzahlen über die Struktur einer Website

Graphen sind mathematische Modelle für netzartige Strukturen in Natur und Technik, [250, S.11]. Eine Website kann als Graph, bestehend aus Webpages als Knoten, und Hyperlinks als Kanten, angesehen werden [12, S.51]. Unter dieser Annahme können Grapheigenschaften gemessen werden. Diese Maße dienen dazu, die Navigation zu verbessern, und eine intuitiv verständliche und benutzbare Website zu erstellen. Wie bei den Content-Maßen ist auch aus der reinen Betrachtung eines Graphen keine Aussage über den Erfolg einer Website möglich. In Kombination mit den Usage-Daten können aber Gründe für das User-Verhalten sichtbar werden, sofern sie in der Struktur der Website begründet liegen.

Globale Grapheigenschaften

Als Grundlage der folgenden Maßzahlen wird eine $N \quad N$ Distanzmatrix C_{ij} berechnet, die die kürzesten Wege innerhalb einer Website von der Webpage i zur Webpage j angibt. Bei Nichterreichbarkeit von zwei Webpages zueinander wird ein Wert K vorgegeben, der um 1 größer als die maximale Distanz in C ist.

Centrality beschreibt nach Dhyani et al. [84, S.471f] den Grad der Verbindung einer Webpage zu den anderen Webpages einer Website. Die Homepage einer Website sollte einen, wenn nicht den höchsten Centrality-Wert besitzen. Um diesen Wert zu berechnen, zählt man in der Distanzmatrix C für jede Webpage i die Gesamtdistanz der ausgehenden Links zu allen direkt oder indirekt verbundenen Webpages,

$$OD_i = \sum_j C_{ij}$$

Außerdem wird die Distanz von allen anderen Webpages zu dieser Webpage berechnet,

$$ID_i = \sum_j C_{ji}$$

Um die Größen von Graphen unterschiedlicher Websites vergleichen zu können, wird die Summe aller paarweisen Distanzen zwischen den Webpages summiert. Anschließend dividiert man durch die jeweiligen OD_i und ID_i und erhält so
Relative **I**n **Centrality**

$$RIC_i = \frac{\sum_i \sum_j C_{ij}}{\sum_j C_{ji}}$$

und die **R**elative **O**ut **Centrality**

$$ROC_i = \frac{\sum_i \sum_j C_{ij}}{\sum_j C_{ij}}$$

Eine Webpage mit zentraler Bedeutung sollte sowohl einen hohen RIC als auch ROC aufweisen.

Compactness beurteilt die Querverweise zwischen Webpages und damit die Erreichbarkeiten innerhalb der Website. Compactness C_p variiert zwischen 0 und 1. Ein völlig unverlinkter Graph hat Compactness von 0. Nach Dhyani et al. in [84, S.472] berechnet sich Compactness folgendermaßen:

$$C_p = \frac{Max \quad \sum_i \sum_j C_{ij}}{Max \quad Min}$$

wobei Min und Max den minimalen und maximalen Werten der normalisierten Centrality entsprechen.

Stratum ist lt. Dhyani et al. [84, S.472f] als soziometrisches Maß definiert, genannt **Prestige**. Es misst die lineare Ordnung des Webgraphen. Hochlineare Websites haben eine einfache Struktur, können aber umständlich zu benutzen sein. Je höher das Stratum, desto linearer ist der Webgraph. Stratum (Prestige) ist die Differenz zwischen *Status* und **Contrastatus**. Dabei ist Status die Summe der Distanz zu allen Webpages (Zeilensumme der Distanzmatrix) und Contrastatus die Summe aller endlichen Distanzen von allen Webpages auf die jeweilige Webpage (Spaltensumme der Distanzmatrix). Die benutzte Distanzmatrix unterscheidet sich von der obigen C_ij dadurch, daß sie bei Nichterreichbarkeit zweier Webpages nicht K sondern ∞ annimmt. Stratum:

$$S = \frac{\sum_i (|\sum_j C_{ij} \quad \sum_j C_{ji}|)}{LAP}$$

wobei LAP (Lineare Absolute Prestige) sich wie folgt berechnet: $\frac{N^3}{4}$ für eine gerade Anzahl an Knoten (Webpages) und $\frac{N^3 \quad N}{4}$ für eine ungerade Anzahl.

Lokale Grapheigenschaften

Depth ist der Abstand einer Webpage vom Root-Dokument, sprich der Homepage. Mit Depth kann man eine Aussage über die Leichtigkeit machen, wie eine Webpage zu erreichen ist. Damit steigt oder sinkt die Wahrscheinlichkeit, daß die Seite gelesen wird. Mit zunehmender Tiefe (Depth) sollte der Detaillierungsgrad der Informationen zunehmen.

Imbalance beschreibt die Ausgeglichenheit eines Webgraphen. Die Idee von Dhyani et al. [84, S.474f] baut darauf, daß jede Webpage ein Thema bearbeitet und jeder Link von diesem Knoten eine Weiterentwicklung dieses Themas darstellt. Inwieweit eine Website ausbalanciert sein muß und inwieweit ein Thema durch mehrere Webpages ausgearbeitet werden soll, muß vom Autor der Website eingeschätzt werden. Botafogo et al. [35] schlagen zwei Maße vor, um Imbalance zu quantifizieren:
Absolute Depth Imbalance Man berechnet für alle Subpages (Subknoten), die von einer Webpage p ausgehen, deren maximale Tiefe (Depth) und erhält den Depthvektor D_p mit den Maximalen Tiefen der Subknoten von p. Die Absolute Depth Imbalance wird als Standardabweichung der Distanzen berechnet, die man zurücklegen muß, wenn man allen Subknoten von p folgt. **Absolute Child Imbalance** zählt die Subknoten einer Webpage p, also die Anzahl an Knoten des unter p liegenden Subgraphen. Daraus wird die Standardabweichung berechnet, die die Absolute Child Imbalance widergibt.

Significance Maße einer Webpage haben nach Dhyani et al. [84, S.476] zwei Perspektiven: 1. die absolute Qualität ungeachtet von User Bedürfnissen und 2. die Relevanz einer

Webpage für das Informationsbedürfnis eines Users, zum Beispiel bei einer Suchanfrage, siehe Abschnitt 3.4.2. Die Qualität einer Webpage wird von deren Verlinkung abhängig gemacht. Dieses Maß ist nicht erfolgsrelevant und wird nur zur Vollständigkeit angegeben.

Quality Der **Hits-Algorithmus** wurde von Kleinberg [148] eingeführt. Kleinberg unterscheidet zwischen *Authorities* Webpages, die eine Information besitzen, und *Hubs* - Webpages, die auf viele Authorities verlinken. Der Hub-Wert y_i einer Webpage p_i ergibt sich aus der Summe aller Authority-Werte der Webpages, die von p_i verlinkt sind.

Der Authority-Wert einer Webpage x_i ergibt sich aus der Summe aller Hub-Werte der Webpages, die auf p_i verlinken. Diese Werte werden in einem iterativen Algorithmus, dem *Hypertext Induced Topic Selection* Algorithmus (HITS), berechnet.

$$x_i = \sum_{j:e_{ij} \in E} y_i, \; y_i = \sum_{j:e_{ij} \in E} x_i \tag{3.1}$$

Diese Gewichtung wird in mehreren Iterationsschritten für alle Webpages durchgeführt und nach jedem Iterationsschritt werden die Hub- und Authority Ratings normalisiert. Kleinberg weißt darauf hin, daß eine gute Authority von vielen hoch bewerteten Hubs referenziert wird und ein guter Hub auf viele hoch bewertete Authorities verweist. Die Hub- und Authority Scores werden durch iterative Neuberechnung der Bewertungen jeder einzelnen Webpage berechnet. Die Iteration wird entweder abgebrochen, wenn die Hub- und Authority-Werte jeder Seite konvergieren oder nach einer vorbestimmen Anzahl an Iterationsschritten. Die Webpages werden anhand ihrer Ratings als Authority oder Hub klassifiziert. Ding et al. [85] und Baldi et al. [12, S.131ff] gehen näher auf die Anwendungen des Hits-Algorithmus ein, beispielsweise als Ranking von Webpages für Suchmaschinen.

PageRank von Brin und Page [184] ist der Ursprung des Such- und Bewertungsalgorithmus von Google zur Erstellung eines Rankings. PageRank wird auch als die Wahrscheinlichkeit angesehen, daß ein zufälliger User (random surfer) die jeweilige Webpage besucht. Je mehr Links auf eine Webpage verweisen, desto höher ist das Gewicht dieser Webpage. Je höher das Gewicht der verweisenden Seite ist, desto größer ist der Einfluß dieses Verweises. Das bedeutet, daß eine hoch bewertete Webpage, hoch angesehene Verlinkungen zu anderen Webpages besitzt. Der Page-Rank-Algorithmus bildet die Wahrscheinlichkeit eines Users nach, auf eine Website zu stoßen, wenn dieser sich per Zufall entlang der Links durch das Internet bewegt.

3.5 Clientseitig erhobene Kennzahlen

Die bisher beschriebenen Kennzahlen beruhen alle auf Daten, die serverseitig erhoben werden können. Im folgenden Kapitel werden Arbeiten vorgestellt, die sich mit Eigenschaften einer Website befassen, die vom User wahrgenommen werden. Im Vergleich zu den oben beschriebenen serverseitigen Maßen, sind sie vom Charakter her mehr Website-Eigenschaften, die ein User einer Website zuordnet, als Maßzahlen.

3.5.1 Usability

Spool et al. geben in [228] eine kurze Übersicht von Empfehlungen für den Aufbau einer leicht benutzbaren Website. Eine Untersuchung über drei Webstudien wurde von Palmer in [185] durchgeführt. Der Erfolg einer Website ist demnach von der

- **Navigierbarkeit** (Organisation, Anordnung und Layout),
- **Inhalt** (Umfang und Vielfalt von Produktinformationen),
- **Interaktivität** sowie der
- **Fähigkeit auf Fragen der User zu antworten** (Feedback Option und FAQs)

abhängig. Neben diesem Katalog an Website-Merkmalen, gibt Palmer [185] keine Antwort, wie der Erfolg letztendlich gemessen werden kann.

Aladwani et al. untersuchen in [8] die Wahrnehmung der User von der inhaltlichen Qualität einer Website. Zviran et al. [277] und Yeung et al. [268] bauen auf dieser Studie auf und untersuchen den Erfolg transaktionsorientierter Websites, wie Online Shopping, Customer Self Service, Online Trading und Publikationsdienste. Sie führen die Qualitätsbewertung mittels Fragebogen durch und erheben Bewertungen zu Website-Merkmalen und der Wahrnehmung der Website durch den befragten User. Ein kompletter Prozeß zur Evaluierung der Website-Qualität wird von Olsina et al. in [181, 182] entwickelt und beschrieben. Alle drei Ansätze zur Bewertung einer Website beruhen auf statischen Eigenschaften einer Website. Sie beruhen auf direkter Erhebung der Daten durch Befragung. Der mit dieser Dissertation verfolgte Ansatz beruht dagegen auf den Beobachtungen der User und der Website aufgrund der Daten, die serverseitig zur Verfügung stehen.

Fleming nennt in [99] folgende Prinzipien eines erfolgreichen Designs der Website-Navigation: leichte Erlernbarkeit, Konsistenz, Verfügbarkeit von Feedback, klare visuelle Botschaften und die Unterstützung des User-Verhaltens und seiner Ziele. Ivory et al. fassen diese Prinzipien zusammen und leiten daraus in [138] Leitlinien für das Website Design ab. Diese Metriken beziehen sich lediglich auf eine graphisch ansprechende Gestaltung einer Website und sind für die Erfolgsermittlung nicht geeignet.

Studie / Quelle	Maß	Maßparameter	Erhebungsmethode
Abbott et al. [1]	-	Zugang und Verfügbarkeit an Informationen, Sicherheit, Atmosphäre, Personalisierung	konzeptionelle Studie
Cho and Park [63]	Website Gestaltung	Produktinformationen, Website Design, Zahlungsmethoden, Kaufprozeß	Befragung
Eroglu et al. [91]	User Wahrnehmung	Freude, Vergnügen	Befragung
Ho et al. [129]	-	Gestaltung der Homepage, logistische Unterstützung, technologische und Produktcharakteristiken	Befragung
Kim et al. [146]	Inhalt	Informationsbreite, Aktualität, Website Konstruktion, Unterhaltung, Werbung, leichte Bedienbarkeit	Befragung
Kholi et al. [144]	-	Zeitersparnis, Kostenersparnis für den Kunden	Befragung
Lederer et al. [157]	Strategischer Vorteil	Webauftritt: Informationsqualität und -zugang [...]	Befragung
Agarwal et al. [2]	Website Usability	Mensch-Computer-Interaktion, Usability Richtlinien	Expertenbefragung
Kim et al.[145]	Kundenloyalität	Qualität der E-Commerce Architektur: Beständigkeit, Bequemlichkeit, Vergnügen	Befragung
Torkzadeh et al. [251]	E-Commerce Success	Erreichung der Projekt- und Unternehmensziele	Befragung

Tabelle 3.2: Studien zur Kundenzufriedenheit [59] und Erfolgsmessung [256]

Devaraj et al. untersuchen in [83] E-Commerce-Metriken, die durch ein Technology Acceptance Model, Transaction Cost Analysis und Service Quality Model evaluiert werden. Die analysierten Metriken und deren Evaluation bezieht sich allerdings nur auf direkt beim User erhobene Daten, sodaß eine Anwendung für serverseitige Beobachtung entfällt.

Eine Untersuchung der Zufriedenheit von Website-Besuchern wurde von Buys et al. [45] durchgeführt. Hierzu wurden fünf Faktoren untersucht: Kundensupport, Sicherheit, Benutzbarkeit, Transaktion und Bezahlung sowie Information, Inhalt und Innovation. Der Faktor Information wird weiter konkretisiert durch Up-to-Date-Informationen und innovative Produkte und Dienstleistungen. Da sich die Arbeit von Buys et al. auf transaktionsorientierte Banken-Websites bezieht, scheidet auch deren Anwendung aus.

3.5.2 Kundenzufriedenheit

In den unten beschriebenen Arbeiten wird die Kundenzufriedenheit untersucht. Dabei werden unterschiedliche Parameter berücksichtigt, die in Tab. 3.2 aus Studien von Wade et al. [256] und Cheung et al. [59] zusammengefaßt sind.

Cheung et al. bieten einen umfassenden Überblick an Untersuchungen aus den Jahren 1999 bis 2004. Wade et al. bezeichnen in [256] *"weiche"* Erfolgsmaße als Ergänzung zu den oben beschriebenen *"harten"* Maßen.

Außerdem nennen Cheung et al. [59] vier Dimensionen der Informationsqualität einer Website:

- Präzision: die Seriosität der Informationen beeinflußt die Meinung des Kunden über die Website

- Inhalt: relevante und vollständige Informationen ermöglichen dem Kunden eine kompetente Entscheidung
- Format: die Art der Präsentation der Information, z.B. angereichert durch Graphiken
- Aktualität: regelmäßige Überarbeitung der Informationen

Da alle genannten Arbeiten zur Erhebung dieses User Feedbacks auf Fragebögen angewiesen sind, kommen diese Kenngrößen zur Erfolgsmessung informationsorientierter Websites in dieser Dissertation nicht in Frage.

Maßparameter wie Zugänglichkeit und Gestaltung der Website können durch Strukturmaße (Kap. 3.4.3) beschrieben werden. Zum Beispiel muß das Strukturmaß Centrality durch User-Studien kalibriert werden, je nachdem, welche Web Struktur von den Usern gewünscht wird. Allerdings können die Arbeiten zur Festlegung der mit der Website verfolgten Ziele genutzt werden. Wie im E-Commerce-Qualitätsmodell von Riemer et al. [201] in Abb. 1.8 gezeigt, soll der Erfolg einer Website sowohl von Unternehmenszielen als auch von den Zielen des Users abhängen. Zur Bestimmung dieser "weichen" Ziele sind explizite User-Befragungen geeignet. Zur kontinuierlichen Messung des Website-Erfolgs eignen sich solche Untersuchungen nicht.

3.6 Zusammenfassung

Sterne [234] zählt den Mangel an Maßzahlen zu den größten Hemmnissen zur Entwicklung des E-Business.

Straub et al. geben in [243, 244] einen Überblick der existierenden Web Metriken, auf dem Teltzrow und Berendt in [248, 249] aufsetzen. Darin befassen sich Teltzrow et al. mit Erfolgsmetriken für E-Commerce Websites. Sie begründen in [248, S.17] die Notwendigkeit neuer, benutzungsbezogener Metriken mit der eingeschränkten Interpretierbarkeit der bekannten Statistiken und Metriken, wie Conversion Rates.

Da sie eine E-Commerce Website im Zusammenhang anderer Vertriebskanäle eines Unternehmens betrachten, führen Teltzrow et al. Multichannel-Metriken ein, wie beispielsweise eine Offline Payment Rate, Payment Migration Rate, Deliveries-to-Store Rate und eine Delivery Migration Rate. Zusätzlich zu diesen Metriken, die sich auf die Web Usage-Daten beziehen, eine Concept Conversion Rate und eine Offline Conversion Rate ein. Dazu werden die Webpages identifiziert, die Informationen zu anderen Vertriebskanälen bereithalten, wie ein Store Locator für den nächstliegenden Bezugspunkt.

Teltzrow et al. [248] beurteilen die Konzentration der Conversion Rates auf den Kaufzeitpunkt als Einschränkung und Problem bei einer Multichannel-Betrachtung. Hierfür sei eine

Einbeziehung des Informationsverhaltens der User notwendig. Teltzrow et al. schlagen daher eine detailliertere Betrachtung der Informationsphase im Kundenkaufprozeß vor, in: Service, Offline Information, Katalog Information, und Produkt Information. Auch Teltzrow et al. [248, 249] bleiben bei der Bemessung des Erfolgs von Websites auf transaktionsorientierte Websites beschränkt, da sie sich an Conversion Rates orientieren.

Allerdings unterstreicht das Urteil von Teltzrow et al. den Bedarf nach neuen Erfolgsmaßen, die mit dieser Dissertation für informationsorientierte Websites untersucht und erstellt werden sollen.

Als Ergebnis dieses Überblicks an Web Metriken und Web Maßen bleibt festzuhalten, daß zur Erfolgsmessung transaktionsorientierter Websites eine große und gut erforschte Zahl an Kennzahlen existiert - wie Maße über den Kaufprozeß und den Kundenbeziehungslebenszyklus. Mit ihnen läßt sich der Erfolg nicht nur feststellen, sondern auch quantifizieren.

Zur Erfolgsmessung informationsorientierter Websites stehen direkt keine Maße zur Verfügung. In Kapitel 5 werden die Basiskennzahlen mit den Ansätzen wie der Task Completion Rate und Focus zu einer neuen Kennzahl kombiniert.

4 Web Mining für informationsorientierte Websites

Kapitel 3 hat gezeigt, wie man die anfallenden Datenmengen im Internet zu Kennzahlen verdichten kann. Kennzahlensysteme und Maße zur Erfolgsmessung transaktionsorientierter Websites sind umfangreich beschrieben und finden bereits Anwendung. Für informationsorientierte Websites ist dies nicht der Fall. Maße wie Page Views, Sessions, User, Duration, Focus oder die Task Completion Rate können nur einzelne Aspekte des Erfolgs abdecken. Es ist ein tieferes Verständnis des Benutzerverhaltens, der Wirkungen der Website, ihres Aufbaus und Inhalts, sowie der Reaktionen des Kunden auf Aktionen auf der Website erforderlich.

Dieses Kapitel untersucht, inwieweit Web Mining-Analysen geeignet sind, Erkenntnisse über den Erfolg von informationsorientierten Websites zu erlangen und Ansätze zu finden, aus denen in Kombination mit Kennzahlen ein Erfolgsmaß erstellt werden kann.

Hierzu werden zunächst der Begriff des Web Mining und dessen Teilgebiete erläutert. Anschließend werden einzelne Web Mining Arbeiten, Verfahren und Methoden auf deren Eignung zur Beantwortung der Fragestellung dieser Dissertation untersucht. Data Mining und Web Mining werden unter anderem von Wilde et al. in [124, S.8ff], Pyle [197, S.9-37] und Berry und Linoff [28, S.17-35], Mena [170] oder Küppers [153] in ihren Grundlagen und Konzeptionen ausführlich beschrieben. Daher konzentriert sich diese Arbeit auf die Anwendbarkeit und Bewertung von Web Mining-Verfahren bezüglich der Erfolgsmessung informationsorientierter Websites.

Die Begriffe *Data Mining* und *Web Mining* werden von Berry und Linoff [28], Fayyad et al. [94] und Wrobel et al. [265] wie folgt definiert: "*Wissensentdeckung in Datenbanken ist der nichttriviale Prozeß der Identifikation gültiger, neuer, potentiell nüztlicher und schlußendlich verständlicher Muster in großen Datenbeständen.*" Kosala erweitert die eigentliche Web Mining-Analyse um die Datenextraktion, Datenaufbereitung und Interpretation der Ergebnisse zum Web Mining-Prozeß in [151], siehe Abb. 1.9.

4.1 Web Mining-Überblick

Gemäß Hippner et al. [123, S.5] bestehen für die Betreiber von Websites verschiedene Möglichkeiten, Wissen über Nutzung und Nutzer ihrer Website zu erlangen. Diese Informationen beinhalten ein großes Potential zur Anpassung des Internet Auftrittes an individuelle Kundenbedürfnisse und damit zur Umsetzung eines individualisierten CRMs im Internet [125].

Für Chakrabati [50] liegt das Ziel von Web Mining-Analysen darin, dem User neben statischen Hyperlinks und generischen Suchanfragen an Suchmaschinen verbesserte Navigationsmöglichkeiten und effizientere Informationszugänge zu ermöglichen. Berendt. et al. sehen in [22, S.1] den Zweck von Web Mining darin, Methoden und Systeme zur Entdeckung und Beschreibung von Prozessen innerhalb des World Wide Webs zu entwickeln. Sie beschreiben Web Mining als Teilgebiet und Synthese aus den Gebieten: Data Mining, Machine Learning, Knowledge Discovery in Databases, Business Intelligence (Chamoni et al. in [54]), Internettechnologie und Semantic Web, Statistik, Verarbeitung natürlicher Sprachen (natural language processing) und Informatik. Die Betriebswirtschaftslehre wird zur Definition der Analyseziele und Interpretation, Umsetzung und Kontrolle der Ergebnisse, von Kosala in [151] als notwendig angesehen und er schlägt eine Einbindung der Web Mining-Analysen in die bestehenden Geschäftsprozesse vor.

Als Hauptmethoden von Web Mining werden bei Berendt et al. [22, S.6ff] Klassifikation, Clustering, Regelidentifikation und Sequenzanalyse genannt.

Kosala et al. [151] und Etzioni [92] geben folgende Problemstellungen an, mit denen sich Data Mining befaßt, die von Srivastava et al. in [229] speziell für Web Usage Mining konkretisiert werden:

- Auffinden relevanter Informationen
- Business Intelligence: Neues Wissen aus vorhandenen Informationen generieren
- Personalisierung des Informationsangebots
- Usage Charakterisierung und Wissen über den Konsumenten und User sammeln
- Verbesserung der Website und des Webangebots

Anand et al. nennen in [10] als Grund für den noch immer zögerlichen Einsatz von Web Mining-Analysen, daß der Wert, der durch diese Analysen geschaffen wird, meist unbekannt bleibt. Dieses Problem stellt den Ausgangspunkt für das Vorhaben der Dissertation dar, den Erfolg informationsorientierter Websites meßbar zu machen.

Abb. 4.1 gibt einen Überblick der Teilgebiete von Web Mining und der daraus resultierenden Anwendungsmöglichkeiten. Wie von Kosala et al. [151, S.2] und Berendt et al. in [22, S.1] beschrieben, ist Web Mining eine Unterkategorie des Data Mining. Web Mining läßt sich analog zu Srivastava et al. [229] und Cooley et al. [68, Abb.1] nach den Untersuchungsobjekten weiter untergliedern in: Web Usage- (Abschnitt 4.2), Web Content- (Abschnitt 4.3) und

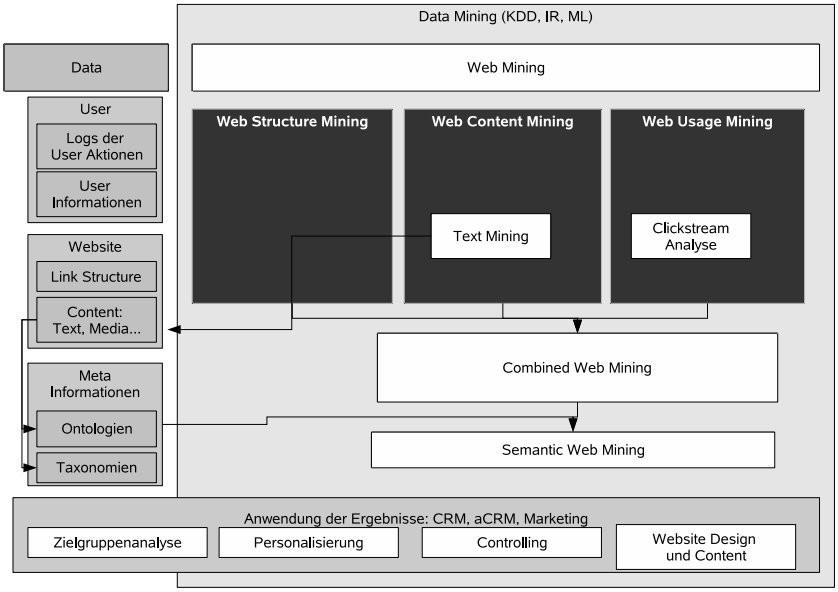


Abbildung 4.1: Themenüberblick Data Mining

Web Structure Mining (Abschnitt 4.4). Im Verlaufe dieses Kapitels werden Ansätze beschrieben und bewertet, die Web Structure, Content und Usage Mining miteinander kombinieren. Wie in Abb. 4.1 gezeigt, wird die Einbeziehung von Zusatzinformationen, wie Ontologien, zu Semantic Web Analysen und deren Anwendungsmöglichkeiten angesprochen.

Wie in Kapitel 1.3 beschrieben, orientiert sich der Aufbau dieser Arbeit an den Schritten des Web Mining-Prozesses nach Fayyad et al. [94], Wilde et al. [127] und Srivastava et al. [229], der in Abb. 1.9 dargestellt ist. Die dort gezeigten einzelne Schritte haben Chapman et al. zu einem geschlossenen Prozeß angeordnet, der um einen ersten Prozeßschritt ergänzt wurde. Den Ausgangspunkt des Data Mining-Prozesses bildet die betriebswirtschaftliche Analyse der Aufgabenstellung. Dieser Prozeß ist in Abb.4.2 dargestellt. Küsters weist in [154, S.97] darauf hin, daß die Idee eines automatisierten Data Mining Prozesses, wie von Fayyad et al. [94] vorgeschlagen, eine idealtypische Vorstellung wäre, von der man noch weit entfernt sei, sodaß die menschliche Expertise nach wie vor notwendig bleibe.

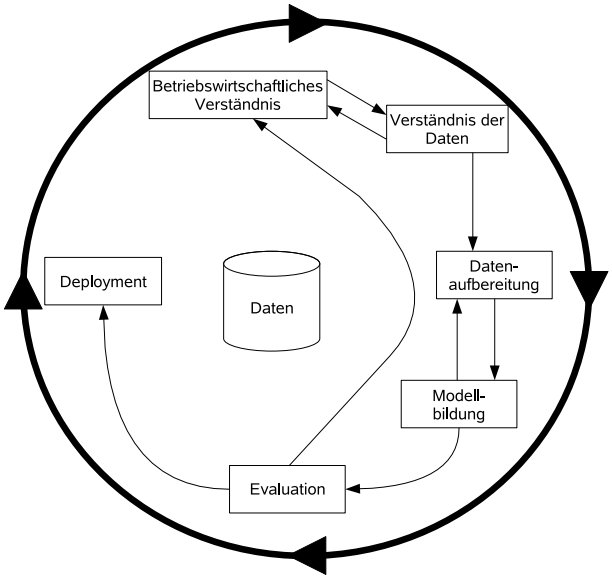


Abbildung 4.2: Der Data Mining-Prozeß, nach [56, 57]

4.2 Web Usage Mining

4.2.1 Überblick Web Usage Mining

Kosala et al. unterteilen in [151] Web Usage Mining in die Erstellung von User-Profilen und die Analyse der Navigationsmuster von Usern. Anand et al. unterscheiden in [10, S.24ff] zwei Hauptanwendungen von Web Usage Mining:
1. Web Measurement, um den Wert zu erfassen, der durch die Website erzeugt wird, zum Beispiel durch Messung und Verfolgung von Marketingkampagnen und
2. die Analyse von Conversion Rates (siehe Web Metriken, Kap. 3.3.2) und die Hindernisse bei der Umwandlung eines Users in einen Kunden. Den Einsatz von Conversion Rates zur Beurteilung von Websites, beschreibt Berthon in [29].

Srivastava et al. unterteilen in [229, S.18] Web Usage Mining in fünf Anwendungsgebiete, die von Arndt et al. [11, S.90f] noch genauer erläutert werden:

- **Personalisierung** versucht aus dem Nutzerverhalten im Internet Rückschlüsse zu ziehen und mit dieser Information einen persönlichen Service für jeden einzelnen User zu erstellen.
- **Systemverbesserung** ist die leistungs- und kostenoptimale Ausgestaltung der eigenen Website in Bezug auf Hard- und Software.

- **Website-Modifizierung** nutzt die Datenauswertungen zur Erhöhung der Attraktivität der eigenen Internet-Präsenz.
- **Business Intelligence** setzt sich mit der Gewinnung strategischer Marketing-Informationen aus dem Internet auseinander.
- **Web Usage-Charakterisierung** ist fokussiert auf die Untersuchung genereller Verhaltensmuster der User.

Nach Wilde et al.[124] ist die Grundlage eines erfolgreichen Kundenbeziehungsmanagements "...das Wissen über den Kunden und seine Bedürfnisse". Im folgenden Abschnitt wird das Kundenverhalten auf einer Website auf typische Verhaltensweisen hin untersucht. Im Gegensatz zu einzelnen Kennzahlen, werden mit der Clickstream-Analyse Zusammenhänge zwischen einzelnen User-Aktionen und dadurch auch Zusammenhänge zwischen Webpages deutlich.

4.2.2 Clickstream-Analysen

Für die Beurteilung des Erfolgs einer Website ist das Verständnis der User-Aktivitäten ein wichtiger Ausgangspunkt. Reine Zugriffsstatistiken, wie sie im Abschnitt 3.3.1 beschrieben werden, ermöglichen nach Spiliopoulou et al. [224, S.1][221] nur eine rudimentäre Annäherung an ein User Feedback. Sie schlagen stattdessen vor, das Navigationsverhalten der User als Ganzes zu untersuchen [226, S.1], indem die Clickstreams der User ausgewertet werden.

Bekannte Anwendungen der Clickstream-Analyse, auch Pfadanalyse genannt, beschreiben Berendt et al. in [23]. Dazu zählen beispielsweise die Personalisierung einer Website durch Prognose des nächsten Clicks und entsprechender Anpassung der Website an den User.

Es bieten sich hierzu Assoziationsanalysen an, die gemeinsam auftretende Ereignisse extrahieren können und daraus Regeln ableiten. Diese Data Mining-Methode wird außerhalb des Internets beispielsweise zur Warenkorbanalyse eingesetzt, bei der gemeinsam gekaufte Produkte gesucht werden. Die Assoziationsanalyse eignet sich durch Einbeziehung des Zeitpunktes eines Clicks auch als Clickstream-Analyse, um Sequenzen von Webpages innerhalb von Sessions zu analysieren und daraus "Wenn-Dann-Regeln" abzuleiten [23, S.145]. Durch diese Regeln erfährt man, welche Pfade durch die Website von den Usern häufig benutzt werden.

Zunächst wird beschrieben, wie eine Clickstream-Analyse durchgeführt wird. Dies soll auf die hier zu analysierenden Websites angewandt werden. Diese Untersuchung soll als Grundlage dienen, aus der Anforderungen für ein mögliches Erfolgsmaß erstellt werden.

Die grundlegende Arbeit zur Assoziationsanalyse ist laut Berendt und Spiliopoulou [23] die Arbeit von Agrawal et al. [4], die von ihm in [5, 7] weiter vertieft wird. Darin werden Algorithmen beschrieben, um eine Clickstream-Analyse effizient durchführen zu können. Bevor

auf diesen Algorithmus eingegangen wird, soll der Gegenstand einer Assoziationsanalyse erläutert werden. Denn daraus ergibt sich die Notwendigkeit, effiziente Algorithmen und Maße zur Erkennung interessanter Regeln einzusetzen.

Gegenstand der Assoziationsanalysen

Itemsets ohne Berücksichtigung der Reihenfolge Wird die Assoziationsanalyse auf Clickstreams von einer Website angewandt, ist man an häufigen Clicks bzw. Click-Sequenzen interessiert. Der erste Schritt besteht in der Suche nach häufig gemeinsam besuchten Webpages, also häufigen Webpage-Sequenzen, sogenannten **Itemsets**, wie Bollinger in [32] beschreibt: ein Item ist der Click bzw. der Besuch einer Webpage, und ein Itemset besteht aus mehreren Items.

Hettich et al. [120] beschreiben eine Assoziationsregel wie folgt: $A \rightarrow B$ bestehe aus einer Itemmenge A im Regelrumpf und einer Itemmenge B im Regelkopf, wobei A und B disjunkt sein müssen. Eine Assoziationsregel läßt sich umgangssprachlich als Wenn-Dann-Regel formulieren: $A \rightarrow B$; wenn Webpage A aufgerufen wurde, wird in x Fällen auch Webpage B aufgerufen.

Itemsets mit Berücksichtigung der Reihenfolge Bei Warenkorbanalysen spielt die Reihenfolge der Itemsets keine Rolle, sodaß $A \rightarrow B$ die gleiche Bedeutung beigemessen wird wie $B \rightarrow A$, [23]. Spielt der Transaktionszeitpunkt und damit die Reihenfolge der Items aber eine Rolle, wie hier bei der Reihenfolge der Webpage-Aufrufe, analysiert man Sequenzen von Items und spricht von einer **Sequenzanalyse**. Diese Sonderform der Assoziationsanalyse ist lt. Agrawal et al. [7] insbesondere für die Analyse von Internetzugriffen geeignet, da die Bewegungsmöglichkeiten der User auf Websites durch deren Verlinkung in ihrer Reihenfolge beschränkt sind.

Die Sequenzanalyse geht von einem sequentiellen Browsingverhalten aus. Diese Annahme ist, wie in Kapitel 2.2.3 über paralleles Browsingverhalten gezeigt, nicht immer gegeben - siehe auch Viermetz und Stolz et al. [254]. Da dieses Browsingverhalten noch nicht ausreichend untersucht ist, wird hier von einem sequentiellen, nicht parallelen Clickstream ausgegangen.

Da in der zeitlich geordneten Sequenz der Clicks die Gründe eines Users liegen, von einer Webpage auf eine andere zu wechseln, werden ungeordnete Itemsets hier nicht untersucht. Ansonsten würden Informationen, wie der Clickzeitpunkt und die Benutzung bestimmter Hyperlinks, verloren gehen. Es wird also eine Sequenzanalyse durchgeführt. Andere Fragestellungen können genau diese Bindung des User-Verhaltens an die Link-Struktur stören. So können häufig benutzte Wege auf einer Website durch die Verlinkung erzwungen werden, die aber vom User nicht gewünscht sind. In dieser Arbeit soll das beobachtete Verhalten der User zu

einer Bewertung der Website herangezogen werden. Eine noch zu findende Bewertungsfunktion soll zwischen gewünschtem, zielgerichtetem Verhalten und nicht gewünschtem Verhalten unterscheiden.

Komplexe Itemsets Komplexe Assoziationsmuster entstehen, wie bei Gaul et al. in [104] beschrieben, wenn der Weg von Webpage *A* auf *B* untersucht werden soll, dieser aber nicht zwingend direkt, sondern möglicherweise über weitere Seiten führt. Spiliopoulou beschreibt in [227] sogenannte Platzhalter (wildcards), sodaß beliebig viele Webpages zwischen A und B $A \rightarrow \quad \rightarrow B$ liegen können. Der Ereignisraum möglicher Regeln wird dadurch enorm vergrößert. Diese Art von Itemsets wird daher hier nicht weiterverfolgt. Um die Zahl möglicher Regeln abschätzen zu können, wird im folgenden auf die Kombinatorik eingegangen.

Häufigkeiten der Itemsets bestimmen Zur Bestimmung der Häufigkeiten von Itemsets wird jede Session in darin enthaltene Itemsets zerlegt und deren Häufigkeiten abgezählt. Die Regelrümpfe und Regelköpfe umfassen ein oder mehrere Items. Gegeben sei die Session $A \rightarrow B \rightarrow C \rightarrow A \rightarrow B$, die sich in folgende Itemsets aufspalten lassen:

- Itemmenge 2: 2 $A \rightarrow B$; $B \rightarrow C$; $C \rightarrow A$;
- Itemmenge 3: $A \rightarrow B \rightarrow C$; $B \rightarrow C \rightarrow A$; $C \rightarrow A \rightarrow B$
- Itemmenge 4: $A \rightarrow B \rightarrow C \rightarrow A$; $B \rightarrow C \rightarrow A \rightarrow B$
- Itemmenge 5: $A \rightarrow B \rightarrow C \rightarrow A \rightarrow B$

Es entstehen bei einer Session mit 5 Clicks 10 mögliche Itemsets, unter Berücksichtigung der Reihenfolge. Probleme bestehen allerdings in der kombinatorischen Explosion des Lösungsraums [154], d.h. der möglichen Itemsets. Daraus ergibt sich die Notwendigkeit, intelligente Algorithmen zu entwickeln, um die Zahl möglicher Itemsets zu verringern.

Komplexitäten Die theoretisch mögliche Zahl an Webpage-Kombinationen, die bei der Benutzung einer Webpage entstehen können, ist wie beim obigen Beispiel überschaubar, muß aber bei steigender Webpage-Anzahl und Itemset-Länge untersucht werden.

Man spricht von Kombinationen (*C*), wenn die Reihenfolge innerhalb der Itemsets keine Rolle spielt, und von Variationen (*V*), wenn die Reihenfolge der Itemsets berücksichtigt wird. Kann sich außerdem noch jedes Element beliebig oft wiederholen, wie zum Beispiel bei Webpages, die innerhalb einer Session beliebig oft besucht werden können, steigt die theoretisch denkbare Zahl der Kombinationsmöglichkeiten gemäß $V_N^K = N^K$.

Diese theoretische Zahl wird durch die tatsächlich besuchten Webpages und deren Verlinkung stark reduziert.

Daher benötigt man gemäß Dong et al. [86, 87] intelligente Verfahren, die die Kombinationsmöglichkeiten auf die tatsächlich vorkommenden Möglichkeiten beschränken. Da dies zur Reduktion der Ergebnismenge nocht nicht ausreichend ist, sollen Kombinationen, die nicht in ausreichendem Maße auftreten, ausgeschlossen werden. Außerdem ist man nur an potentiell interessanten, neuen Regeln interessiert. Offensichtliche Regeln interessieren nicht. Zu diesem Zweck werden die beobachteten Itemsets mit Interessantheitsmaßen beurteilt.

Interessantheitsmaße

Agrawal [4] benutzt die Interessantheitsmaße Support und Confidence, um die Datenmenge zu beschränken und effizient berechenbar zu machen. Interessantheitsmaße werden außerdem von Dehaspe et al. [78] und Mannila et al. [167] beschrieben:

Der **Support** einer Regel ist die Häufigkeit des gemeinsamen Auftretens der entsprechenden Items. Der Support liegt in einem Intervall zwischen 0 und 1 und gibt Auskunft darüber, welcher Anteil an allen Transaktionen *D* diese Regel erfüllt. Dieses Maß gibt jedoch keine Auskunft darüber, inwieweit die gefundene Regel neue oder bekannte Tatsachen widerspiegelt [120].

$$support(A \rightarrow B) = \frac{|t \in D|\{(A \cup B) \quad t\}|}{|D|} \tag{4.1}$$

Die **Confidence** beschreibt die Stärke der Korrelation der Items. Sie beschreibt den Anteil der Transaktionen, die *A* und *B* beinhalten, an der Menge der Transaktionen, die *A* beinhalten.

$$Confidence(A \rightarrow B) = \frac{|t \in D|\{(A \cup B) \quad t\}|}{|\{t \in D|A \quad t\}|} \tag{4.2}$$

Der **Lift** [137] wird durch das Verhältnis aus der Confidence der Regel zur erwarteten Confidence (entspricht dem support von B) gebildet [120]. Der Lift gibt an, um wieviel häufiger B in allen Transaktionen mit A vorkommt als in der Grundgesamtheit. In anderen Worten, der Lift beschreibt den Erkenntnisgewinn: wenn man weiß, daß A vorliegt, kommt B x-mal häufiger vor als wenn man nicht wüßte, daß A vorliegt.

$$lift(A \rightarrow B) = \frac{conf(A \rightarrow B)}{sup(B)} \tag{4.3}$$

Diese Interessantheitsmaße werden von Algorithmen eingesetzt, um die Analyse- und Ergebnismengen zu reduzieren. Als Beispiel soll der Apriori-Algorithmus von Agrawal et al. [4] beschrieben werden.

Apriori-Algorithmen

Die Idee von Agrawal ist, nicht alle theoretisch möglichen Itemsets zu erstellen und zu testen. Es werden nur diejenigen Itemsets berücksichtigt, die einen Mindestsupport erfüllen. Der Apriori Algorithmus besteht aus zwei aufeinanderfolgenden, sich wiederholenden Schritten.

1. Aus der Datenbasis D_i werden Itemsets mit i Elementen gebildet, wobei $i = 1 \ldots k$ und k z.B. die maximale Sessionlänge ist. Diese Itemsets der Länge i stellen die Kandidatenmenge C_i dar. Dann werden die jeweiligen Häufigkeiten dieser Itemsets festgestellt.
2. Im nächsten Schritt werden die Itemsets bestimmt, die ein Mindestinteressantheitsmaß erfüllen, zum Beispiel einen Mindestsupport oder eine Mindestconfidence. Sie bilden die Ergebnismenge L_i.

L_i bildet die neue Datenbasis D_{i+1} aus der die Itemsets mit $i + 1$ Elementen gebildet werden, die wiederum die Kandidatenmenge C_{i+1} bilden. Dies entspricht wiederum Schritt 1. Dabei werden in der $i + 1$-ten Schleife nur noch diejenigen Variationen als potentielle häufige Itemsets untersucht, die aus der neuen, reduzierten Kandidatenmenge gebildet werden können. Der Prozess [6][120] wird solange fortgesetzt, bis C_i und L_i, mit $i = 1...k$ eine leere Menge bilden.

Aus den gefundenen Itemsets werden Regeln generiert. Dabei werden die n Elemente eines Itemsets so aufgespalten, daß $n \quad p$ Elemente mit $p = 1 \ldots n \quad 1$ einen Regelrumpf bilden und ein Element den Regelkopf. Die n Elemente werden mit allen Kombinationsmöglichkeiten auf ihre Interessantheit überprüft. In der Sequenzanalyse, die bei der Clickstream-Analyse angewandt wird, entsprechen die Itemsets bereits den Regeln, denn die Reihenfolge, in der die Elemente eines Itemsets auftauchen, ist relevant.

Der Effizienzgewinn durch den Apriori-Algorithmus liegt darin, daß sukzessive die zu durchsuchende Datenmenge reduziert wird. Die Itemsets höherer Dimension können nur aus der Menge an interessanten Itemsets einer niedrigeren Dimension gebildet werden. Daher reduziert sich die Anzahl der Variationen deutlich.

Graphische Darstellung der Ergebnisse

Mit einer Clickstream-Analyse [235] wurde für diese Dissertation der Apriori-Algorithmus in der Skriptsprache R umgesetzt und auf die zu analysierenden, informationsorientierten Websites eines großen, deutschen Unternehmens angewandt. Die gefundenen Ergebnisse werden hier ausschnittsweise präsentiert.

Die wichtigsten gefundenen 2-er Itemsets auf einer der analysierten Websites sind als gerichteter Graph in Abbildung 4.3 dargestellt. Ein Graph wird schnell unübersichtlich, sodaß nicht alle Itemsets im Graphen dargestellt sind. Der Graph zeigt die nach dem Support stärksten

Clicks zwischen zwei Content IDs, die sich teilweise zu einem Graphen ergänzen. Mit den reinen IDs ist eine solche Darstellung nur für Kenner der Website interpretierbar. Eine weite-

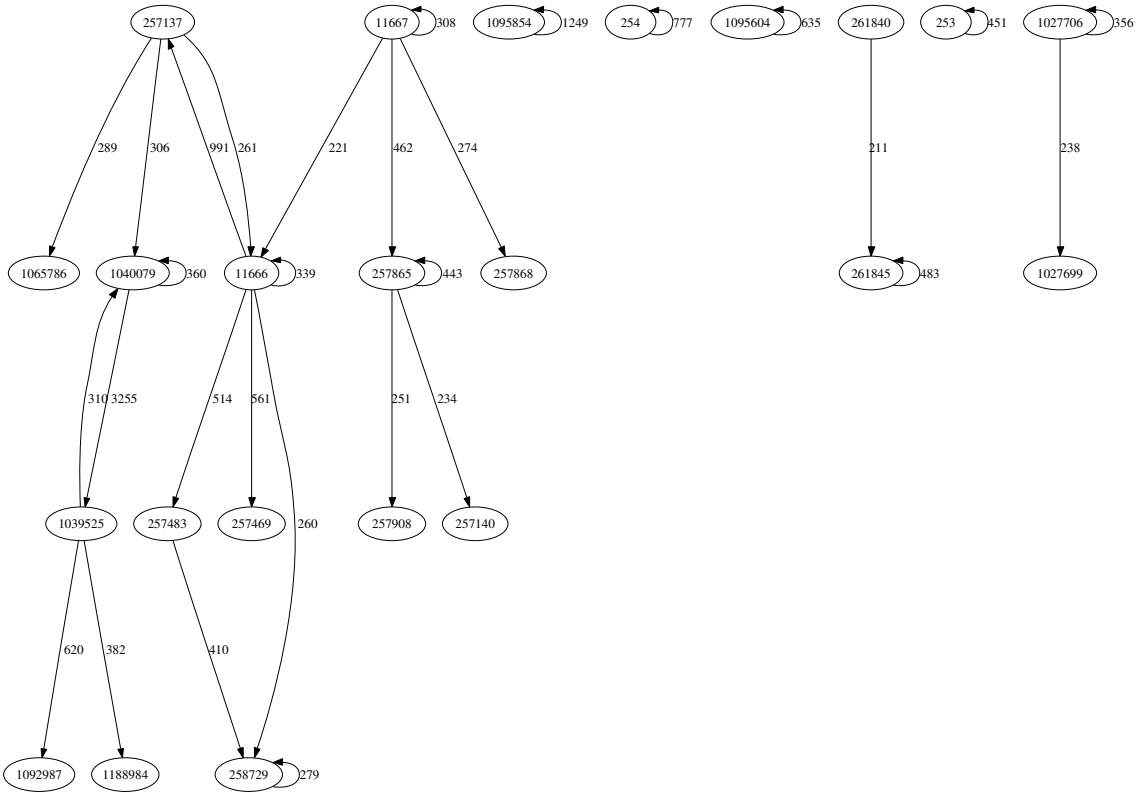


Abbildung 4.3: Graph mit Content IDs

re Beobachtung zeigt, daß die Benutzer der Suchfunktion (ID 253) innerhalb dieser bleiben und sich von dort nicht zu Inhaltsseiten weiterbewegen. Da nicht alle Clickstreams dargestellt sind, muß weiter untersucht werden, ob die Suchfunktion suboptimal ist und die Benutzer tatsächlich nicht zum Inhalt gelangten oder ob nur nicht alle Clicks bzw. Itemsets dargestellt sind. Daher sollte die Länge der untersuchten Clickstreams vergrößert werden. Dazu wurde Abbildung 4.4 erstellt. In dieser Abbildung wurden die kryptischen Content IDs durch die Webpage-Typen ersetzt. So wurden alle Content-Seiten zu einem Knoten im Graphen zusammengefaßt. In Bezug auf die Suchseite erkennt man durch die auf sich selbst gerichtete Kante, daß die User die Suchseite von dieser erneut aufrufen, z.B. um in den Ergebnissen weiterzusuchen. Dies kann durch langes Blättern in den Suchergebnissen verursacht werden. Positiv zu werten ist dagegen, daß sehr viele User zwischen Content-Seiten navigieren. Da hier die Informationen präsentiert werden, die die Website vermitteln will, ist dies ein positiver Indikator für den Erfolg dieser Website. Wünschenswert wäre jedoch eine session-individuelle Bewertung, sodaß aus dieser Bewertung die Schwachpunkte einer Website ersichtlich werden, bzw. der Website-Erfolg gemessen werden kann.

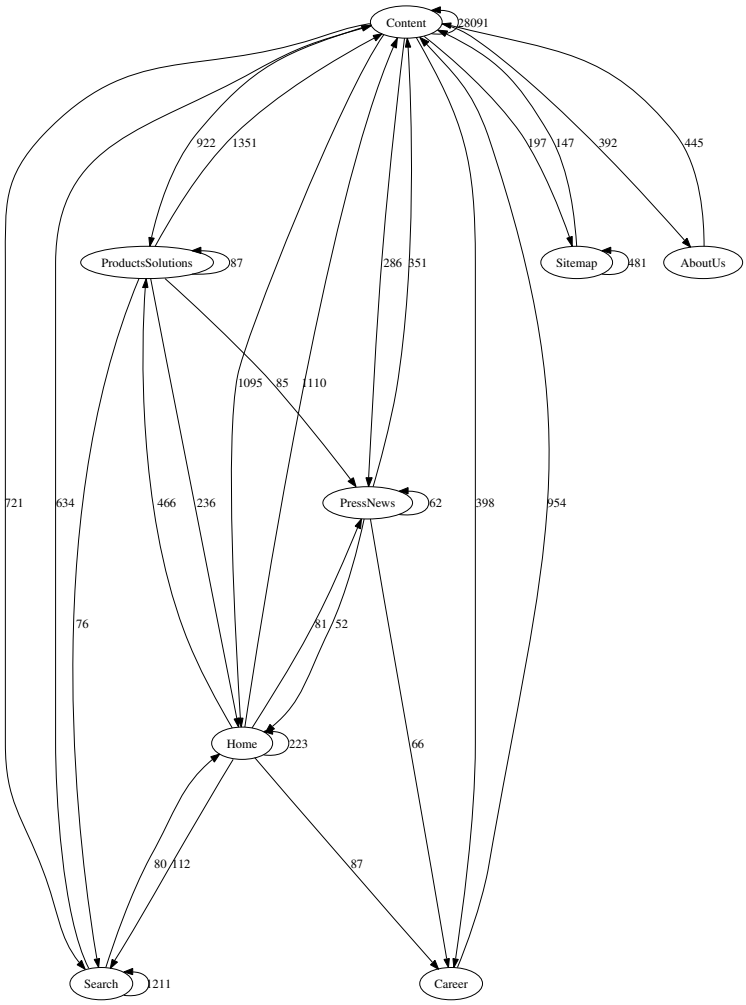


Abbildung 4.4: Graph mit Content IDs und Metainformationen, 2 Clicks je Itemset

Die Darstellung als Graph vermittelt einen ersten Eindruck von den wichtigen User-Pfaden auf einer Website. Für eine genauere Analyse müssen die einzelnen Regeln selbst betrachtet werden. Ein Hinweis auf den Erfolg einer Website ist so allerdings auch nicht zu erreichen.

Der folgende Abschnitt beschreibt Anwendungen der Usage-Analyse, wobei wiederum auf die Erfolgsmessung als Anwendungsfall geachtet wird.

Anwendungen der Clickstream-Analyse

Mobasher et al. beschreiben in [176] die Personalisierung von Websites aufgrund der automatischen Analysen der Benutzung der Website. Hierzu verwenden sie Assoziationsalgorithmen,

wie den Apriori-Algorithmus von Agrawal et al. [4]. Weiterhin beschreiben Mobasher et al. Cluster-Techniken, die eine User Session als Vektor der besuchten Webpages ansehen und über Distanzmaße ähnliche Sessions gemeinsamen Clustern zuweisen. Sowohl aus den gefundenen Regeln als auch den identifizierten Clustern lassen sich für aktuelle Sessions Empfehlungen je nach Übereinstimmung mit erkannten Regeln ableiten. Diese Empfehlungen können einem User in Form von personalisierten Links angeboten werden.

Eine Umsetzung des Apriori-Algorithmus in der Programmiersprache *C* findet man bei Borgelt [33] mit der Besonderheit, daß die Itemsets in einer Baumstruktur angeordnet werden, um effizient auf sie zugreifen zu können. Boulicaut und Jeudy stellen in [36] einen Ansatz der Assoziationsanalyse vor, in dem sie *freie* und *geschlossene* Itemsets benutzen. Geschlossene Itemsets liegen dann vor, wenn es keine Untermenge dieser Itemsets gibt, deren Frequenz größer ist. Dadurch kann die Menge an gefundenen Regeln verringert werden, ohne daß Informationen verloren gehen. Einen anderen Weg zur Reduzierung der Ergebnismenge beschreiben Li und Zhang in [161]. Sie bestimmen die Regeln ohne Berücksichtigung des Supports und wählen ausschließlich Regeln mit 100 % Confidence aus.

Boulicaut und Jeudy schlagen in [36] Effizienzverbesserungen für diese Assoziationsanalysen vor, indem sowohl ein Cache für gefundene Regeln vergangener Analysen zur Verfügung gestellt wird, als auch eine kondensierte Darstellung der Regeln durch sog. Supersets.

Gaul et al. präsentieren in [105] ein formales Gerüst für navigationsbasierte **Recommender-Systeme**. Unter einem Recommender-System verstehen Gaul et al. eine Software, die Informationen über die Besucher einer Website sammelt, aggregiert und aus deren Verhalten Empfehlungen für alle User generiert. Recommendersysteme werden von Resnick et al. [200] beschrieben, die für solche Systeme die Bezeichnung *Collaborative Filtering* benutzen. Das Ziel von Collaborative Filtering sei es lt. Zeng et al. in [271], die Präferenzen eines aktiven Users zu einem Zielobjekt auf Basis anderer User Präferenzen vorauszusagen.

Borges et al. stellen einen Ansatz zur Modellierung von häufigen User-Pfaden durch eine N-Grammatik aus der Linguistik vor, bei der die letzten n besuchten Webpages die Wahrscheinlichkeit beeinflussen, eine bestimmte Webpage im Anschluß zu wählen. [34]. Für die gefundenen Itemsets setzen sie wiederum die bekannten Maße *Support* und *Confidence* ein. Weitere Maße zur Beurteilung der User-Pfade werden nicht verwendet.

Diese Art einer topologischen Clickstream-Analyse verwenden auch Büchner et al. in [44, 16]. Sie suchen nach zielorientierten Navigationspfaden. Dabei analysieren sie Clickstreams von transaktionsorientierten Websites. Sie unterscheiden zwischen Referrerinformationen, Einstiegsseiten, Homepages und Zielseiten, auf denen Transaktionen durchgeführt werden können. Auch wird das Wissen um Marketingkampagnen ausgewertet. Allerdings ist deren Zielorientierung an transaktionsorientierte Websites gebunden.

Shahabi et al. analysieren in [214] die User-Pfade und leiten aus der Sammlung der Webpages je Session ein User-Interesse ab. Allerdings erscheint es recht unsicher, lediglich aus den User-Pfaden direkt auf ein Interesse zu schließen und keine weiteren Informationen heranzuziehen.

Sinnvoll wäre hier eine Kombination mit dern Kennzahlen *Focus* und *Task Completion Rate*. Hierzu wäre eine inhaltliche Analyse der kompletten Website erforderlich. Damit könnte man das User-Interesse besser approximieren und dessen Erfüllung durch die Website beurteilen.

Oberle et al. [180] beschreiben die Anreicherung von Web Usage-Analysen mit Ontologien und Taxonomien. Mit dieser semantischen Anreicherung der Usage-Daten könnte das User-Interesse besser beschrieben werden. Anwendungen hierfür werden vorgeschlagen, aber nicht beschrieben. Einen früheren Ansatz dieser Idee beschreiben Mobasher et al. [176], die aus dem Text einer Webpage eine Menge an Konzepten extrahiert und diese den Zugriffen auf die jeweilige Webpage zuordnet.

Tseng et al. schlagen in [253] die Berücksichtigung der zeitlichen Entwicklung von Clicksequenzen vor, um die Prognosegenauigkeit zu erhöhen. Sie passen dabei die Maße Support und Confidence an, stellen aber keine Erfolgsmaße im Sinne dieser Arbeit zur Verfügung. Der Gedanke, daß unterschiedliche Zeitpunkte zu unterschiedlichem Browsingverhalten führen, haben bereits vorher Nasraoui et al. in ihrem Ansatz [178] erläutert. Darin beschreiben sie die sehr dynamische Natur von Webzugriffsmustern, die aus den häufigen Änderungen der Website-Inhalte, Website-Struktur und den sich schnell ändernden User-Interessen resultieren. Dabei weisen sie auf die Herausforderungen hin, die an Web Mining-Analysen gestellt werden, sich auf diese starke Dynamik einstellen zu müssen. Mit dieser Dynamik muß auch ein Erfolgsmaß umgehen können, damit eine kontinuierliche Überprüfung der Website ermöglicht wird. Den Zusammenhang zwischen Navigationsmustern und der Website-Struktur beschreiben auch Eirinaki et al. [89, 90] und weisen auf die Problematik dynamischer Link-Generierung hin.

4.2.3 Markov Modelle

Eine Alternative im Web Usage Mining zu Assoziationsanalysen bieten Markov Modelle. Darin werden Übergangswahrscheinlichkeiten zwischen Webpages in einer Übergangsmatrix abgebildet. Der Übergang zwischen zwei Webpages wird als Transition bezeichnet. Markov Ketten erster Ordnung beschreiben die Wahrscheinlichkeit für einen Zustand nur aus der Betrachtung des unmittelbar vorherigen Zustands [12, S.184]. Will man Übergangswahrscheinlichkeiten anhand längerer Clicksequenzen berechnen, muß deren Dimensionalität angepaßt werden. Dieses Vorgehen entspricht der Berechnung von Itemsets höherer Dimension. Für jede Tupeldimension muß eine eigene Übergangsmatrix erstellt werden[84]. Eine Reduzierung der Itemsets wie beim Apriori Algorithmus findet zunächst nicht statt.

Wie Markov Modelle zur Clickstream-Analyse eingesetzt werden können, zeigen Deshpande und Karypis in [81]. Ihre Anwendung besteht in der Vorhersage des nächsten Clicks in einer Session. Ein Großteil der Arbeit widmet sich den Möglichkeiten, die Markov Modelle in jeder Dimension im Umfang zu reduzieren, sogenannte *Selektive Markov Modelle*. Dies wird

erreicht, indem Markov Modelle bezüglich der Confidence bereinigt werden - analog zum Apriori-Algorithmus.

Um den zeitlichen Verlauf mit Markov Ketten abzubilden, schlagen Baldi et al. in [12, S.184ff] vor, eine Markov Kette mit diskreten Zeitabschnitten, beispielsweise für jede Sekunde, zu erstellen. Wie dies bei einer großen, stark frequentierten Website im kontinuierlichen Betrieb machbar ist, wird von Baldi nicht untersucht.

Eirinaki et al. beschreiben in [90] Markov Modelle höherer Ordnung, um eine höhere Prognosegenauigkeit der nächsten Webpage zu erreichen. Sie weisen auf den höheren Aufwand bei diesen Markov Modellen hin, die nicht nur Transitionen zwischen zwei Webpages berücksichtigen, sondern längere Pfade zur Modellbildung heranziehen. Dadurch erhöht sich nicht nur gemäß der Kombinatorik (4.2.2) die Zahl der zu berücksichtigenden Pfade, sondern die Trainingsmenge muß auch entsprechend erhöht werden.

Markov Modelle haben das gleiche Anwendungsfeld wie der Apriori Algorithmus. Sie können zwar das User-Verhalten modellieren, sagen aber nichts darüber aus, welches User-Verhalten zu den Zielen einer Website beiträgt und welches nicht.

4.2.4 Web Usage Mining zur Erfolgsmessung von Websites

Viele der Web Usage Mining-Analysen konzentrieren sich auf die Beschreibung des User Verhaltens durch User-Pfade. Hierzu werden beispielsweise Assoziationsregeln oder Transitionswahrscheinlichkeiten durch Markov Modelle beschrieben. Allerdings fehlt eine Analyse dieser User-Pfade und deren Prognosen im Hinblick auf deren Beitrag zu den Zielen einer Website oder den Zielen der User. Die meisten Arbeiten versuchen entweder die Effizienz der Algorithmen oder die Prognosegenauigkeit für das User-Verhalten zu verbessern.

Eine Bewertung der Clicks und User-Pfade findet nicht statt. Häufige User-Pfade sind noch kein Indikator für deren Zielerreichungsbeitrag aus Sicht der Website oder der User. User Pfade hängen außerdem, wie von Eirinaki et al. [90] beschrieben, stark von der Website Struktur ab.

Da neben den Zielen der Website-Autoren auch die Ziele der User für den Erfolg einer Website entscheidend sind, beschreiben Chi et al. in [61, 60, 62] einen Ansatz, der aus den User-Aktionen auf die User-Ziele schließt. Chi et al. sehen dabei, wie diese Dissertation, die User-Ziele darin, Informationen zu erreichen. Ausgehend von einem gegebenen Informationsbedarf und einer Startseite soll das erwartete User-Verhalten vorhergesagt werden. Sie greifen dabei auf die Arbeiten von Pirolli et al.[191, 190, 193] und Olston et al. [183] zurück, die die Analogie einer “Futtersuche“ mit der Informationssuche ziehen. Die User sollen demnach einer “Geruchsspur“ nach den gesuchten Informationen folgen. Eine Bewertung der Clickstreams je nach Erfolg der Informationssuche wurde in den Arbeiten von Chi et al. nicht angedacht. Bei

der Erstellung von Erfolgsmaßen für informationsorientierte Websites in dieser Dissertation wird das Ziel der User als *Suche nach Informationen* analog zu Chi et al. [60, 62] angenommen.

Wie man sieht, reicht Web Usage Mining alleine nicht aus, um das User-Verhalten zu verstehen. Chi et al. [61] kombinieren Usage mit Content Mining, um die Informationssuche beschreiben zu können. Cooley bestätigt in [67, S. 93], daß Web Usage Mining alleine nur geringe Erfolgschancen hat und ein Verständnis der Inhalte und der Struktur einer Website notwendig sind. In einer Reihe von Arbeiten [68, 69, 70] entwickeln Cooley et al. ein integriertes Website-Informationssystem, das Web Usage-, Structure- und Content Mining zusammenführt. Ihre daraus abgeleiteten Untersuchungen und Ergebnisse konzentrieren sich auf die Personalisierung von Websites und stellen eine eher akademische Arbeit zu den grundlegenden Techniken des Web Mining dar. Dieser Idee folgend, gehen die Kapitel 4.3 und 4.4 auf die Analyse der inhaltlichen und strukturellen Aspekte einer Website näher ein.

4.3 Web Content und Text Mining

Web Usage Mining befaßt sich mit strukturierten Daten, die zum Beispiel als Clickstream im Logfile-Format oder aufbereitet in einer Datenbank vorliegen. Inhaltliche bzw. Content Daten liegen aber oft als unstrukturierte oder semi-strukturierte Daten vor [217], beispielsweise als freier Text. Mit der inhaltlichen Analyse anderer Formate wie Multimedia-Daten befaßt sich diese Arbeit nicht. Einen Überblick zum Multimedia Mining gibt Zaiane et al. [269] und zahlreiche andere Autoren [116, 189, 245].

Einen Überblick zu den theoretischen Aspekten von Text Mining und dessen Anwendung geben [9, 98, 101, 150, 163, 171, 270].

Felden nennt in [95, 96] drei Anwendungsgebiete von Text Mining: Klassifikation, Clustering und Abstraktion, also die Zusammenfassung von Texten.

Weiss et al. [258] sehen zusätzliche Anwendungsfelder von Text Mining in der Organisation von Dokumenten, im Information Retrieval und Information Extraction, Evaluation, sowie in der Vorhersage von Trends.

Lu et al. [164, S.175f] und Zhong et al. [273] sehen die Anwendung von Web Content Mining in der Unterstützung des Users beim Auffinden von Informationen über Suchmaschinen, Personalisierung und Empfehlungen über Recommender-Systeme.

Bergmark beschreibt in [24], wie ein Crawler mit Hilfe von Text Mining Hyperlinks nach inhaltlichen Gesichtspunkten auswählt und somit große Mengen an Dokumenten zielgerichtet durchsuchen kann.

Ein umfassender Überblick verschiedener Text Mining Anwendungen gibt der erste International Workshop on Text Mining in [216].

Im folgenden werden einzelne Prozeßschritte und Methoden des Web Content Mining beschrieben. Die Bereitstellung der Daten für Content Mining Analysen wird in Kap. 2.3 beschrieben. Abb. 4.5 ordnet die Datensammlungs- und Datenaufbereitungsschritte in den kompletten Content- bzw. Text Mining-Prozeß ein.

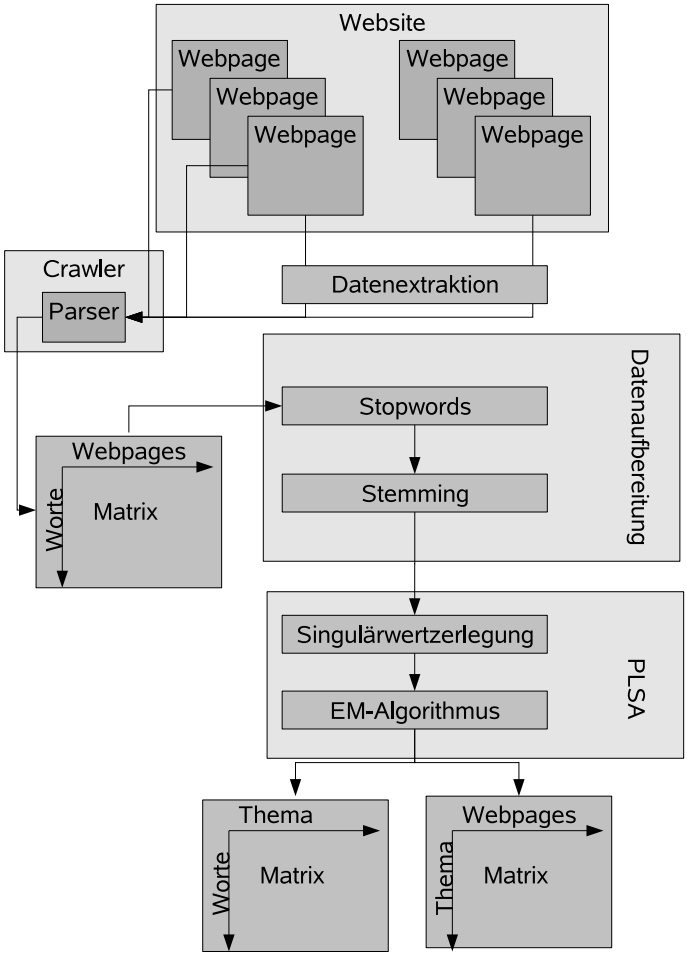


Abbildung 4.5: Text Mining-Prozeß, in Anlehnung an King et al. [147]

4.3.1 Datenaufbereitung

Die Datenaufbereitung hat neben dem Ziel einer sauberen Datenbasis auch den Effekt, die Dimensionalität des Wortraumes zu reduzieren, [162]. Dies ist deshalb notwendig, weil ei-

ne Matrix, die den Wortraum einer Website beschreibt, sehr groß werden kann. Jedes Wort und jedes Dokument erhält einmalig eine Zeile bzw. Spalte in dieser Matrix. Diese Wort-Dokument-Matrix ist nicht nur in ihrer Dimensionalität sehr groß, sondern auch dünn besetzt, [12]. Erst durch eine Reduktion der Dimensionalität und der dichter besetzten Matrix werden die hochdimensionalen Räume für viele Analysen handhabbar. In Kapitel 4.3.3 wird ein Algorithmus beschrieben, der eine weitere erhebliche Dimensionsreduktion und dichtere Besetzung der Matrix erreicht.

Als erste Analyse zu dieser Dissertation wurde von Gedov, Stolz et al. [239], in Anlehnung an Cooper [71], vorgeschlagen, den *Lokalen Kontext* einer Webpage, d.h. die in der Nachbarschaft liegenden Webpages, zur Glättung der Dimensionalität der Worte einer Webpage zu benutzen. Die Inhalte aus dem lokalen Kontext einer Website werden zur Bestimmung des Inhalts der fokusierten Webpage benutzt. Diese Idee wird von Stolz et al. in [240] als Dimensionsreduzierung vor Einsatz des PLSA (siehe unten, Kapitel 4.3.4) eingesetzt. Ziel dieser Arbeit ist eine Identifikation von Themen auf einer Website und stellt einen weiteren Beitrag zu den verwendbaren Methoden für diese Arbeit dar.

Der inhaltliche Zusammenhang und die Verlinkung von Webpages wird von Cohn und Hofmann in [66] untersucht. Das Ziel ihrer Arbeit ist die Vorhersage von Zitaten zwischen inhaltlich verwandten Dokumenten. Die Ergebnisse werden z.B. genutzt, um die Verlinkung von Webpages zu untersuchen. Sie benutzen hierzu den in 4.3.4 beschrieben PLSA-Ansatz, der viele Vorteile bietet und im Folgenden beschrieben wird.

4.3.2 Das Vektorraummodell

Um die Ähnlichkeit von Dokumenten zu bestimmen, können die vorkommenden Worte eines Dokuments (Webpage) zu Vektoren zusammengefaßt werden. Über Distanzmaße lassen sich zwei oder mehrere Webpages miteinander vergleichen und deren direkte inhaltliche Übereinstimmung berechnen. Die inhaltliche Übereinstimmung hängt dabei vom Auftreten der gleichen Worte ab.

Synonyme werden nicht berücksichtigt, sodaß Webpages mit gleicher inhaltlicher Bedeutung nicht als inhaltlich gleich erkannt werden. Umgekehrt führen **Polyseme**, die bei gleichem Wort eine unterschiedliche Bedeutung haben, dazu, daß Webpages mit unterschiedlicher Bedeutung trotzdem als ähnlich erkannt werden. Ein Beispiel für ein Polysem ist das Wort “Bank“, das entweder ein Kreditinstitut oder eine Sitzgelegenheit beschreibt.

Felden beschreibt das Vektorraummodell in [95] in einem Überblick der Modelle und Methoden zur inhaltlichen Analyse von Texten. Auch Baldi et al. [12, S.83f] beschreiben die Ähnlichkeit von Dokumenten im Vektorraum.

Um den Einfluß der Textlänge auf die Häufigkeit einzelner Worte w_j zu eliminieren, wird die Worthäufigkeit n_{ij} zur *Term Frequency* (TF) normalisiert: $TF_{ij} = \frac{n_{ij}}{|p_i|}$. Salton et al. [202, 204, 203] beschreiben das TF-IDF Gewicht, das die TF mit der **Inverse Document Frequency** (IDF) kombiniert $IDF_j = log\frac{n_j}{n}$.

TF-IDF bildet die Wichtigkeit eines Wortes als absolutes Maß ab. IDF verringert sich mit der Zahl an Webpages, auf denen das Wort vorkommt. Das TF-IDF-Gewicht berechnet sich dann wie folgt: $x_{ij} = TF_{ij} \quad IDF_j$. Das TF-IDF-Gewicht stellt lt. Baldi et al.[12, S.84] ein weitverbreitetes Maß dar. Nach Papineni [187] berechnet TF-IDF das theoretisch nachgewiesenermaßen optimale Gewicht eines Wortes. Andere textzentrierte Metriken sind bei Belew et al.[17] zu finden und in Kap. 3.4.2 beschrieben.

4.3.3 Latent Semantic Indexing

Die inhaltliche Beziehung kann laut Baldi et al. [12] durch einfaches Abgleichen gemeinsam vorkommender Worte nicht dargestellt werden. Ein Verfahren, das die inhaltliche Bedeutungszuordnung darstellen kann, ist das Latent Semantic Indexing.

Hofmann [130] beschreibt den von Deerwester et al. in [77] entwickelten Begriff des Latent Semantic Indexing (LSI) so, daß Worte und Dokumente (Webpages) in einem Latent Semantischen Raum dargestellt werden. Dabei soll der hochdimensionale Raum der Wort-Dokument-Matrix auf den niederdimensionalen Latent Semantischen Raum abgebildet werden.

Dies wird durch eine Singulärwertzerlegung (Singular Value Decomposition (SVD)) erreicht. Die Wort-Webpage-Matrix wird durch SVD in drei Matrizen aus Eigenvektoren und Eigenwerten aufgespalten. Die Idee [261] ist diejenige, daß die Wort-Webpage-Matrix aus wichtigen Hauptdimensionen und Nebendimensionen besteht. Die Hauptdimensionen tragen zur Beschreibung des Vektorraumes den Hauptteil bei, d.h. deren Vektoren haben die höchsten Eigenwerte. Sie bestehen aus den für die Bedeutung der Webpages wichtigen Wörtern. Sie können die Wort-Dokument-Kombinationen in hohem Maße alleine beschreiben, ohne auf die unbedeutenderen Nebendimensionen angewiesen zu sein.

Die Nebendimensionen bestehen aus den weniger wichtigen Worten. Sie haben geringe Eigenwerte. Nur die Hauptdimensionen sollen dabei erhalten bleiben. Dabei können Konzepte, das heißt von der Bedeutung her ähnliche Wörter, auch Synonyme, zusammengefaßt werden. LSI generalisiert also die Bedeutung von Wörtern [261].

Die SVD wird beispielsweise bei Handl [114, S.183,423f] oder Hastie et al. [115, S.487f] beschrieben. Ein alternatives, auf Eigenwerten basierendes Verfahren zur Dimensionsreduktion, ist die Multidimensionale Skalierung, wie sie von Hofmann et al. [132] beschrieben wird.

Die Wort-Webpage-Matrix M wird in drei Komponenten zerlegt: $M = U \quad S \quad V$. Die beiden orthogonalen Matrizen U und V enthalten dabei Eigenvektoren. S ist eine Diagonalmatrix

mit den Wurzeln der Eigenwerte, auch Singulärwerte genannt. Die Diagonale der Eigenvek-

	p_1	p_2	p_3	p_4	p_5	p_6
Linux	1	0	0	0	0	1
Unix	0	0	1	0	0	0
Debian	1	0	1	0	0	0
database	0	2	1	0	0	0
relational	0	2	0	0	1	0
data-set	0	0	0	1	2	0
hardware	0	0	0	0	1	0
server	0	0	0	2	0	1

Tabelle 4.1: Transponierte Webpage-Wort-Matrix

	p_1	p_2	p_3	p_4	p_5	p_6
Linux	0,42	-0,01	0,40	0,47	-0,34	0,52
Unix	0,21	0,30	0,31	0,05	-0,20	0,19
Debian	0,41	0,34	0,52	0,19	-0,41	0,41
database	0,29	1,97	0,91	-0,17	0,16	0,09
relational	-0,13	1,85	0,40	0,06	1,05	-0,23
data-set	-0,28	0,25	-0,35	1,24	1,66	0,06
hardware	-0,07	0,07	-0,15	0,96	0,97	0,19
server	0,50	-0,27	0,29	1,68	0,43	0,94

Tabelle 4.2: Transponierte Webpage-Wort-Matrix M_{LSI} nach SVD

tormatrix S besteht aus folgenden Eigenwerten (3,23 2,83 2,03 1,61 1,30 0,43). Über die Eigenwerte der erzeugten Matrix S kann man nun die Dimensionsreduktion steuern, indem nur die k größten Eigenwerte verwendet werden und die anderen auf 0 gesetzt werden. Dann ergibt sich nach Baldi et al.[12, S.89] durch Rekonstruktion $M_{LSI} = U \quad S_{LSI} \quad V$ die neue Webpage-Wort-Matrix in Tab. 4.2. In Tab. 4.2 sieht man das Ergebnis des LSI nach Baldi [12, S.89]. Kommen die Worte *Linux*, und *Unix* nie gemeinsam vor, so wird durch das Bindewort *Debian* ein relativ hoher Wert für *Unix* auf Webpage p_1 zugewiesen, obwohl *Unix* dort gar nicht vorkommt. Dies zeigt, wie der LSI durch andere gemeinsam vorkommende Wörter die Problematik von Synonymen mildern kann.

Wie man sieht, ist M_{LSI} nicht mehr so dünn besetzt wie M, sodaß auch die Problematik der dünn besetzten Matrizen (*Sparse Matrices*) reduziert wird. Der Semantische Raum wird durch die auf die Bedeutungen reduzierte Term-Dokument-Matrix deutlich und spiegelt die den Dokumenten unterliegende Struktur wider. In dieser Struktur wird die Semantik dieser Dokumente sichtbar, [261]. Die Projektion auf die Eigenwerte gibt die Zugehörigkeit zu einem Konzept an.

Ein Problem stellt die Festlegung der Anzahl der zu verwendenden Eigenwerte dar. Hierzu schlägt Handl in [114, S.131f] vor, bei einer Hauptkomponentenanalyse das Kriterium von Kaiser oder Jolliffe zu benutzen. Das Kriterium von Kaiser berücksichtigt nur die Hauptkomponenten, deren Eigenwerte größer sind als der Mittelwert aller Eigenwerte. Es kann aber untersucht werden, welcher Anteil der Gesamtstreuung durch die einzelnen Hauptkomponenten erklärt wird, [114, S.120f].

Auch die Mehrdimensionale Skalierung und Hauptkomponentenanalyse, die bei Handl [114] und Bartell et al. in [14] beschrieben werden, sind Verfahren zur Analyse eines Vektorraumes. Sie können auch dazu genutzt werden, die Zahl der notwendigen Dimensionen auf die wichtigsten Dimensionen zu reduzieren. Letztendlich kann die menschliche Inspektion der Daten nicht ersetzt werden, um die Anzahl der zu verwendenden Eigenvektoren bzw. Dimensionen festzulegen.

4.3.4 PLSA - Probabilistic Latent Semantic Analysis

Hofman [130, S.51] stellt das Latent Semantic Indexing auf eine "statistisch fundiertere Basis" und benennt es *Probabilistic Latent Semantic Analysis* (PLSA). Die Idee des PLSA wurde auch bei Landauer et al. in [156, 259ff] beschrieben.

Das Aspekt Modell

Der Kern der PLSA ist ein statistisches Modell, das sogenannte Aspekt Modell. Es assoziiert eine unbeobachtete Variable z mit jedem Vorkommen eines Wortes w auf einer Webpage p. Baldi et al.[12, S.91] bezeichnen in diesem Modell das Auftreten eines Wortes w in einem Dokument p als Ereignis. Eine nicht direkt ersichtliche, latente Variable z ist nach folgendem Schema mit jedem Ereignis verknüpft:

- *eine ausgewählte Webpage p besitzt die Wahrscheinlichkeit $P(p)$,*
- *zu ihr wählt man eine latente Variable z mit Wahrscheinlichkeit $P(z|p)$.*
- *Ein Wort w beschreibt ein Konzept sprachlich durch die Wahrscheinlichkeit $P(w|z)$.*

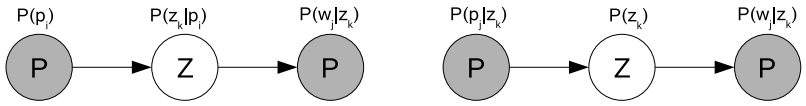


Abbildung 4.6: Aspektmodell als Bayesianisches Netzwerk, nach [12, Fig.4.4]

Abb. 4.6 zeigt, wie durch die nicht beobachtete Konzeptvariable z Wörter von Dokumenten bedingt unabhängig werden. Damit ergibt sich nach [130] für die Wahrscheinlichkeit des Ereignisses $P(w, p)$, daß w in p auftritt, folgendes probabilistisches Modell 4.4:

$$P(w, p) = P(p) \sum_{z} P(w|z)P(z|p) \tag{4.4}$$

Die Dimensionalitätsreduktion wird darin deutlich, daß sich jede Webpage p durch einen Vektor mit k Elementen aus dem Semantischen Raum beschreiben läßt. Eine Webpage wird nicht mehr durch Wörter in einem Vektorraum beschrieben, sondern durch latente Variablen in einem Semantischen Raum. Genauso wird jedes Wort durch einen Vektor der latenten Variablen beschrieben.

PLSA weist im Vergleich zu anderen Clusterverfahren Vorteile auf. Weiss et al. [258, S. 108-116] beschrieben die eindeutige Zuordnung von Webpages zu einem Cluster anhand des k-Means- und Hierarchischen Clusterings. Im Gegensatz dazu und zum Dokumenten-Clustering, wie von Felden [95] beschrieben, ordnet die PLSA Webpages nicht einzelnen

Clustern zu, sondern beschreibt die Themenzugehörigkeit einer Webpage zu allen Themen durch einen Vektor. Dieser Vektor besteht aus den Gewichten $P(z|p)$. Interpretiert man die unbeobachtete Konzeptvariable z als Thema, wird jede Webpage durch einen Themenvektor beschrieben, der die Wahrscheinlichkeiten für eine Webpage angibt, zu den Themen (Konzepten) zu gehören. Diese unscharfe Zuordnung stellt laut Hofmann [130, S.51] eine realitätsnähere Sicht auf den verborgenen, semantischen Raum einer Dokumentensammlung (Website) dar, als eine scharfe Zuordnung zu Clustern.

Die Analogie zu LSI wird von Baldi [12, S.91f] dadurch hergestellt, daß drei Matrizen U, V und Σ eingeführt werden, deren Elemente sich folgendermaßen zusammensetzen: $u_{ik} = P(p_i|z_k), v_{j,k} = P(w_j|z_k)$und $_{kk} = P(z_k)$, mit $i = 1, ..., n$, $j = 1, ..., m$ und $k = 1, ..., K$. Dadurch können alle n m Webpage-Wort-Wahrscheinlichkeiten in einer Matrix zusammengefaßt werden:

$$P = U\Sigma V^T \tag{4.5}$$

Hofmann zeigt in [46, 131], wie die Parameter dieses Modells durch Maximum Likelihood geschätzt werden können, indem der EM-Algorithmus auf die Konzeptvariable z angewandt wird.

Expectation Maximization

Der bei Hastie et al. [115, S.236ff] beschriebene Expectation-Maximization-Algorithmus (EM) ist ein Standard-Werkzeug, um schwierige Maximum-Likelihood-Probleme zu lösen. Im Fall der PLSA werden die beobachteten Webpage-Wort-Häufigkeiten dazu benutzt, die unbekannte Konzept-Variable z zu schätzen.

Die Maximum-Likelihood-Schätzfunktion wird bei Fahrmeier et al. [93, S.376ff] beschrieben. Es wird zunächst eine Likelihood-Funktion L aufgestellt. Die Wahrscheinlichkeiten $P(p), P(z|p), P(w|z)$ werden gemäß [130, 240] durch die Maximierung der folgenden Log-Likelihood-Funktion bestimmt:

$$L = \sum_{p \in P} \sum_{w \in W} n(p, w) log P(p, w) \tag{4.6}$$

wobei $n(p, w)$ beschreibt, wie oft Wort w auf Webpage p vorkommt.

Der EM-Algorithmus, wie er lt. Hofmann von Dempster et al. in [80] beschrieben wird, besteht aus zwei Schritten. Der erste Schritt ist der Expectation-Step (**E-Step**), gefolgt vom Maximization-Step (**M-Step**). Man beginnt mit einer Startverteilung des Modellparameters z. Diese Verteilung kann lt. Baldi et al. [12, S.12] beispielsweise zufällig verteilt sein. Anschließend werden der E-Step und M-Step nacheinander ausgeführt.

Im Expectation-Schritt werden die Erwartungswerte der Konzeptvariable z berechnet, basierend auf den momentanen Schätzungen der Parameter. Nach Hofmann [130] erhält man für den E-Step folgende Gleichung:

$$P(z|p,w) = \frac{P(z)P(p|z)P(w|z)}{\sum_z' P(z')P(p|z')P(w|z')'} \tag{4.7}$$

$P(z|p,w)$ ist die Wahrscheinlichkeit, daß ein Wort w auf einer bestimmten Webpage p durch den entsprechenden z-Faktor erklärt wird: Baldi et al. zeigen in [12, S.91f], daß unter der Annahme der Parameterisierung aus Abb. 4.6 der E-Step sich folgendermaßen vereinfachen läßt:

$$P(z_k|p_i,w_j) \propto P(w_j|z_k)P(p_i|z_k)P(z_k) \tag{4.8}$$

Daraus lassen sich die folgenden Maximization-Schritte ableiten. Im M-Step werden lediglich die Parameter für die im E-Step berechneten posteriori Wahrscheinlichkeiten aktualisiert [12, S.92].

$$P(w|z) = \frac{\sum_p n(p,w)P(z|p,w)}{\sum_{p,w'} n(p,w')P(z|p,w')'} \tag{4.9}$$

$$P(p|z) = \frac{\sum_w n(p,w)P(z|p,w)}{\sum_{p',w} n(p',w)P(z|p',w)'} \tag{4.10}$$

$$P(z) = \frac{1}{R}\sum_{p,w} n(p,w)P(z|p,w), R \quad \sum_{p,w} n(p,w) \tag{4.11}$$

Der Vereinfachung von Baldi et al. [12, S.92] folgend, werden die Parameter durch den M-Step wie folgt aktualisiert:

$$P(w_j|z_k) \propto \sum_{i=1}^{n} n_{ij}P(z_k|p_i,w_j) \tag{4.12}$$

$$P(z_k|p_i) \propto \sum_{j=1}^{|V|} n_{ij}P(z_k|p_i,w_j) \tag{4.13}$$

$$P(z) \propto \sum_{i=1}^{n}\sum_{j=1}^{|V|} n_{ij}P(z_k|p_i,w_j) \tag{4.14}$$

Die iterative Durchführung von E- und M-Step ist ein konvergierender Prozeß, der lt. Hofmann [130] sowie [115, 12, 260] zu einem lokalen Maximum der Log-Likelihood-Funktion 4.6 führt.

Für alle Worte $w \in W$ über alle Konzeptvariablen $z \in Z$ ergibt die Wahrscheinlichkeit für ein Wort $P(w|z)$, bei gegebenem z, eine Matrix . In dieser Matrix wird jedes Wort durch einen Vektor über alle z beschrieben.

In Veröffentlichungen aus Vorarbeiten zu dieser Dissertation [15, 240] wurden die Konzeptvariablen als Themen interpretiert, sodaß sich aus die Themenvektoren der einzelnen Worte ableiten ließen. Der EM-Algorithmus als Teil des PLSA gruppiert semantisch verbundene Worte in Themen und deckt so die semantische Verbindung zwischen den Wörtern einer Dokumentensammlung bzw. einer Website auf. Treten mehrere Wörter oft gemeinsam auf, dann wird deren Wahrscheinlichkeit, dem gleichen Thema z anzugehören, höher.

Genauso werden für alle Webpages $p \in P$ in einer Matrix die Wahrscheinlichkeiten $P(p|z)$ angegeben, für eine Webpage p zu einem Thema z zu gehören. $P(p|z)$ gibt die Wahrscheinlichkeit für ein gegebenes Thema z wieder.

Der Algorithmus transformiert die ursprünglich im hochdimensionalen Wortraum gelegenen Webpages in einen niederdimensionalen Themenraum.

Der PLSA findet auch über Content Mining hinaus Anwendung. Jin et al. verwenden den PLSA in [139] als Web Usage Mining-Methode zur Identifikation ähnlichen User-Verhaltens. Dadurch sollen User-Interessen und Präferenzen besser modelliert werden können.

Beispiel Die Durchführung des PLSA und des EM-Algorithmus soll an einem Beispiel verdeutlicht werden. Zur Veranschaulichung werden in Abb. 4.7, die Initialisierung der Matrizen des EM-Algorithmus gezeigt, und anschließend der E-Step und M-Step durchgeführt.

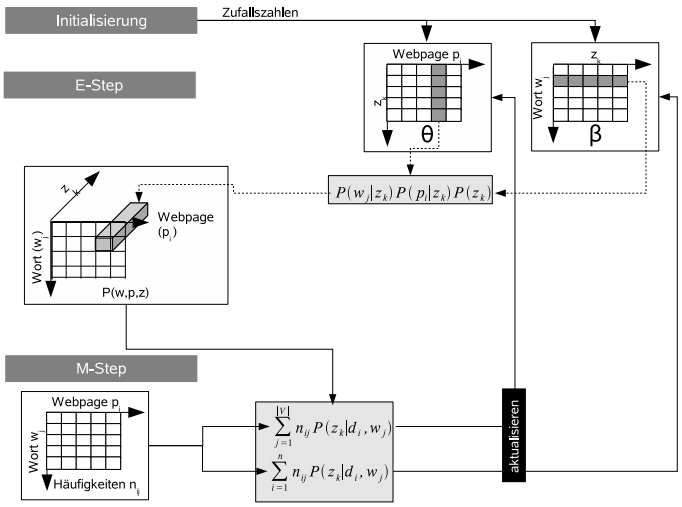


Abbildung 4.7: Darstellung des EM-Algorithmus innerhalb des PLSA

4.3.5 Anwendungsfälle von Content Mining-Analysen

In [15, 238, 240, 241] wurden als Vorarbeiten dieser Dissertation die Anwendbarkeit von PLSA untersucht. Als Ergebnis bleibt festzuhalten, daß PLSA in hohem Maße dafür geeignet ist, den Inhalt einer Website für weitere Analysen verfügbar zu machen. Durch die identifizierten Themen kann auch der inhaltliche Zusammenhang einer Session beschrieben werden. In Verbindung mit einer Usage-Analyse und den vorher beschriebenen Kennzahlen, soll im nächsten Kapitel ein Erfolgsmaß für informationsorientierte Websites erstellt werden.

Xu et al. kombinieren in [266] Usage, Content und Structure durch PLSA direkt, um daraus User Profile zu erstellen.

Eine Anwendung von Text Mining auf Websites im betriebswirtschaftlichen Kontext wird von Lakshminarayan et al. in [155] beschrieben. Es werden Einträge von Usern in Diskussionsforen einer Unternehmenswebsite mit Gaussian Mixture-Modellen und Hierarchischem Clustering analysiert, um eine inhaltliche Synthese zu erhalten. Die Forenbeiträge sollen auf User Feedback hin untersucht und zur Website-Verbesserung eingesetzt werden. Die Analyseziele des Ansatzes von Lakshminarayan [155] überschneiden sich mit den Zielen dieser Arbeit insoweit, daß ein User Feedback erfaßt und analysiert werden soll, um daraus Aussagen über den Erfolg einer Website abzuleiten. Allerdings ist der Ansatz von [155] auf ein direktes, explizites User Feedback angewiesen.

Web Content Mining alleine bietet keine Möglichkeiten, den Erfolg einer Website zu beurteilen. Um ein inhaltliches Verständnis des User-Verhaltens und der User-Absichten zu ermöglichen, ist eine inhaltliche Analyse der besuchten Webpages naheliegend. In diesem Sinne erscheint eine Kombination aus Web Usage- mit Content Mining sinnvoll. Diesem Ansatz kommt Camillo in [48] nach, indem er über eine Singulärwertzerlegung die Distanzen zwischen den Webpages einer Website ermittelt. Diese inhaltlichen Distanzen benutzt Camillo, um die Support- und Confidence Maße einer Clickstream-Analyse (siehe Kap. 4.2.2) anzupassen. Dabei wird die Confidence zwischen inhaltlich sehr nahen Webpages erhöht und damit solche Clicksequenzen bevorzugt.

So wie der PLSA Webpages Konzepten zuweist, so schlagen Zhu et al. [274] vor, das User Verhalten zu benutzen, um semantische Relationen zwischen Webpages zu erkennen. Durch ein Clusterverfahren werden die so erkannten ähnlichen Webpages in einer Link-Hierarchie neu angeordnet. Das Ergebnis ist eine alternative Website-Navigation, die das User Verhalten adaptiert. Ein ähnliches Verfahren wird in Kapitel 4.5 vorgestellt.

Text Kategorisierung Die Kategorisierung von Texten ist für die Beantwortung der Fragestellung dieser Dissertation im absehbaren Rahmen nicht sinnvoll einsetzbar, sodaß auf eingehendere Beispiele verzichtet wird. Die automatische Text Kategorisierung wird von Hoi et

al. in [133] vorgestellt und dahingehend verbessert, daß nicht wie meist üblich nur ein Dokument pro Kategorisierungsschritt bearbeitet wird und das Kategorisierungsmodell nicht jedes Mal neu trainiert werden muß.

Um über eine reine Analyse von losen Webpage-Sammlungen hinauszukommen, analysiert Cohen in [65] neben den Texten auch die Verlinkung der Webpages. Er unterstellt dabei eine sinnvolle Verlinkung, in die der Autor der Website sein Domänenwissen eingebracht hat. Dieser Vorteil für eine erfolgreiche Text Kategorisierung wird bei einer objektiven Evaluation einer Website durch den Nachteil aufgewogen, daß Rückschlüsse auf die Website-Struktur nicht unvoreingenommen sind.

Zusammenfassung von Texten Weiss et al. geben in [258, 157-195] Anwendungsbeispiele für Text Mining: Vorfiltern von E-mails zur gezielten Weiterleitung, Zusammenfassen von Web Dokumenten zur Marktforschung, Themenidentifikation von Nachrichtenartikeln im Sinne des Information Retrieval. Eine weitere Anwendungsmöglichkeit des Text Mining ist die Zusammenfassung von Texten. Das oben in Kap. 4.3.2 beschriebene TF-IDF-Verfahren orientiert sich an den Worthäufigkeiten im Text, um die wichtigsten Worte zu erkennen. Dies stellt weniger eine Zusammenfassung eines Textes dar, als einen Datenbereinigungsschritt, bevor andere Algorithmen, wie hier PLSA, angewandt werden können. Humphreys [136] beschreibt einen Algorithmus, der die wichtigsten Phrasen eines Textes extrahiert. Ähnliche Arbeiten und Weiterentwicklungen sind bei Kerner et al. [111], Mallett et al. [166] und Chen et al. [58] zu finden. Zwar ließe sich das Interesse der User auch auf diesem Wege repräsentieren, jedoch müßte man anschließend wiederum ein Clusterverfahren wie PLSA anwenden, um die unterschiedlichen Interessensbeschreibungen vergleichbar zu machen.

Zamir et al. präsentieren in [270] einen Ansatz, der Dokumente aufgrund gemeinsam vorkommender Phrasen clustert und automatisch eine Zusammenfassung der Webpages liefert. Dieser Algorithmus wird Suffix Tree Clustering genannt und von Crabtree et al. [73] zur Verbesserung von Suchmaschinenergebnissen eingesetzt. Auch dieser Ansatz stellt eine gute Erweiterung zu PLSA dar, die in zukünftigen Arbeiten integriert werden soll.

4.3.6 Semantic Web Mining

Das Semantic Web geht auf eine Initiative von Tim Berners-Lee [26] zurück, die zum Ziel hat, das WWW in ein verteiltes System zur Wissensrepräsentation und Verarbeitung zu entwickeln. Berendt et al. gehen von einer semantischen Anreicherung von Clickstreams in [20, 21] aus. Darauf aufbauend beschreiben sie die Idee des Semantic Web in [22, S.2]. Anstelle bestimmte Dokumente zu finden, soll es möglich werden, Fragen zu beantworten. Dabei treten Fragen zur Infrastruktur, Interoperabilität und einer gemeinsamen Sprache zum Austausch

von Metainformationen über Daten auf. Eine Anwendung von Text Mining für das Semantic Web beschreiben Winkler und Spiliopoulou in [262] mit der automatischen semantischen Annotation von Texten.

Das Semantic Web benötigt eine Sprache [22, S.3], in der die Informationen repräsentiert werden können, die notwendig sind, (a) um Wissen zu repräsentieren und zu verarbeiten, (b) zur Beschreibung des Inhalts (Content) von Dokumenten, (c) zum Austausch der Dokumente und des darin beinhalteten Wissens und (d) zur Standardisierung. In ihrer Arbeit beschreibt Berendt die Bestandteile eines Semantic Web, nach den Empfehlungen des W3C-Konsortiums[1].

Diese sind:

- **XML** (Extensive mark-up language) als Sprache zum mark-up und Annotierung von Dokumenten,
- **RDF** (Resource Description Framework) als Grundlage zur Verarbeitung von Metadaten,
- **OWL** (Ontology Web Language) mit Elementen der Beschreibungslogik und Konstrukten zur Spezifizierung von Semantiken, wie Konjunktionen und Disjunktionen,
- **Ontologien**, die die Relationen zwischen Konzepten beschreiben.

Die Anwendbarkeit von Semantic Web Mining leidet noch unter der geringen Verfügbarkeit semantischer Informationen. Es bleibt abzuwarten, ob sich der Aufwand, die Semantiken zu pflegen, lohnt und sich somit durchsetzt. Neben den unten stehenden Arbeiten sind die meisten Ansätze zum Semantic Web Mining von akademischem Interesse.

Welty geht in [259] auf die Herausforderungen eines Semantic Web ein. Er hält den großen Aufwand bei der Erstellung von Ontologien für ein Haupthindernis und schlägt einfachere Standardsemantiken vor, die einen breiten Bereich abdecken, sogenannte Leight-Weight-Ontologies.

Oberle et al. verbinden Benutzerdaten mit semantischen Informationen in einem *semantic log file* [180]. Wie das Potential des Semantic Web für betriebswirtschaftliche Fragestellungen genutzt werden kann, dafür gibt Lee in [158] ein Beispiel in einer modellbasierten Umsetzung von Unternehmensstrategien über eine Semantische Schicht in die IT-Architektur.

4.4 Web Structure Mining

Wie in Kapitel 2.4 beschrieben, besteht die Struktur einer Website aus der Verlinkung der einzelnen Webpages einer Website. Bei Kosala et al. [151, S.9] wird Web Structure Mining

[1] `http://www.w3.org`

als benachbartes Forschungsgebiet von Sozialen Netzwerken und Zitatanalysen gesehen. Viele Web Structure Analysen lassen sich auf die Graphentheorie zurückführen [28, S.216ff]. Die Idee ist, eine Bewertungsfunktion über die Topologie eines Graphen zu legen, [250].

Die Grundannahme hinter der Web Strukturanalyse ist, lt. Baldi et al. [12, S.125], daß Hyperlinks Informationen über die menschliche Einschätzung einer Webpage enthalten, in dem Falle die Einschätzung des Website Autors. Je mehr eingehende Links eine Webpage aufweist, desto wahrscheinlicher ist es, daß sie eine hohe Relevanz und Qualität besitzt. Diese Sichtweise stammt aus der Literaturwissenschaft, die mittels Zitaten die Wichtigkeit und Relevanz von Artikeln und Büchern beurteilt. Solche Zitat-Netzwerke wurden bereits 1955 von Garfield [102] bewertet.

4.4.1 Ranking-Algorithmen

Die Link-Topologie einer Website bzw. des gesamten Internets wird in Arbeiten von Kleinberg [148] mit Hilfe des HITS Algorithmus untersucht. Der wohl bekannteste Algorithmus zur Analyse der Web Struktur ist der Google Algorithmus, der von Page et al. in [37, 184] als Page-Rank Algorithmus entwickelt wurde. Baldi et al. [12, S.57f] beschreiben den Page-Rank-Algorithmus als die Zeitspanne, die ein User, der sich zufällig durch das Web bewegt, auf dieser Webpage verbringen würde. Es gibt zahlreiche Arbeiten [41, 141, 267, 276], die sich mit dem Page-Rank-Algorithmus beschäftigen und diesen verbessern wollen. Oft liegen diese Verbesserungen im marginalen Bereich und stellen mehr eine akademische Arbeit, als einen Durchbruch in der Anwendbarkeit der Ergebnisse dar. Der ursprüngliche Page-Rank-Algorithmus wurde von den Google-Gründern Page, Brin et al. in [184] vorgestellt.

Bei der Gestaltung und Organisation einer Website läßt ein Website Autor sein Domänenwissen in die Anordnung der Webpages einfließen. Kleinberg analysiert in [148] die Struktur von Websites mit dem Ziel, die Webpages in zwei Kategorien einzuordnen, Hubs und Authorities. Dieser HITS-Algorithmus wurde in Kapitel 3.4.3 beschrieben. Kleinberg wendet seine Analysen auf die Suche im Internet an. Die Bewegung von Usern zwischen diesen Kategorien kann zur Beurteilung der jeweiligen Clicks herangezogen werden. Die Kategorisierung von Webpages anhand ihrer Verlinkung kann bei der Erstellung von Erfolgsmaßen eingesetzt werden. Dies kann dazu genutzt werden, um festzustellen, ob eine Webpage eher der Navigation oder der Bereitstellung von Inhalten dient.

4.4.2 Kombinationen aus Web Content- und Structure Mining

Die enge Verbindung von Suchmaschinen zu Web Content- und Web Structure Mining wird auch in den Arbeiten von Barath und Henzinger [30, 119] und Chakrabati et al. [50, 51, 52, 53]

deutlich, die den von Kleinberg entwickelten HITS-Algorithmus zur Themenidentifikation nutzen, um damit Suchanfragen besser beantworten zu können.

Pant und Menczer beschreiben in [186] eine Arbeit, die Web Structure Mining mit einer inhaltlichen Analyse der Webpages kombiniert. Das Ziel ist ein inhaltlich gesteuerter Crawler, der zu einem Thema interessante Hub-Webpages identifiziert und so gezieltes Crawlen ermöglicht. Die Kombination von Struktur- und Inhaltsanalyse wird auch von Wang et al. in [257] genutzt, allerdings um Suchergebnisse von Suchmaschinen zu clustern.

Eine weitere Arbeit, die Web Structure und Web Content Mining kombiniert, findet sich bei He et al. [118]. Darin wird ein Ähnlichkeitsmaß zur Beschreibung der topologischen Homogenität einer Website beschrieben. Ein solches Maß kann zur Analyse einer Session eingesetzt werden, indem es den Weg des Users durch diese Topologie beschreibt. Aus diesem Bewegungsprofil lassen sich Rückschlüsse auf die Interessen eines User ziehen. Auf dieser Idee baut der in Kapitel 4.5 vorgestellt Ansatz auf.

4.4.3 Verbesserung der Website-Struktur

Martin et al. analysieren in [75] die Struktur einer Website anhand der Zahl der Links und der Entfernung der Webpages untereinander. Die Entfernung, die ein User zurücklegen muß, um von der einen auf die andere Webpage zu gelangen, wird durch eine Kostenfunktion beschrieben. Ziel der Arbeit ist eine Minimierung der Kosten durch Einfügen sogenannter *Hotlinks*. Diese werden identifiziert, indem die Zugriffskosten für tatsächlich stattgefundene User-Zugriffe auf dieser Website minimiert werden, ohne daß die Zahl der Links zu sehr erhöht wird.

Cooper beschreibt in [71] einen Ansatz, um in einem Graphen Webpages zu einem Thema und deren Verlinkung zu identifizieren. Er beschreibt damit das Problem der Identifizierung *geheimer Gesellschaften*, angewandt auf Web Graphen. Der verwendete Algorithmus analysiert die Nachbarschaft einer Webpage und erstellt daraus für jede Webpage iterativ die *Meinung* der Nachbarn. Dehmer untersucht in seiner Dissertation [79] die stukturellen Eigenschaften hypertextbasierter Dokumente und deren strukturelle Vernetzung. Dazu benutzt Dehmer Graphenähnlichkeitsmodelle und beschreibt Verfahren zur Bestimmung der strukturellen Ähnlichkeit von Graphen.

Eine Anwendung dieser Web Structure Mining Ansätze und Graph-Analysen wird im folgenden Kapitel beschrieben. Die Link-Struktur einer Website soll mit Hilfe der von den Usern wahrgenommenen Themen verbessert werden.

4.5 Anwendung von Web Mining auf informationsorientierten Websites

Mit der Publikation [242] wurde als Vorarbeit zu dieser Dissertation ein Ansatz vorgestellt, der Web Usage-, Web Content- und Web Structure Mining kombiniert. Das Ziel der Arbeit bestand in der Verbesserung der Website Struktur. Der Ansatz dazu besteht im Vergleich der Struktur der Website mit der vom User wahrgenommenen Struktur. Abweichungen dienen dann als Indikator von Verbesserungsmöglichkeiten bei der Verlinkung der Webpages. Die vom User wahrgenommene Struktur bezieht sich auf den inhaltlichen Zusammenhang der User Sessions, woraus auf die inhaltlichen Zusammenhänge der Website geschlossen wird.

Abbildung 4.8 zeigt die Vorgehensweise zur Erstellung des Konsistenz-Checks.

User Topic Map

Zunächst werde die Daten gemäß Kap. 2 durch einen ETL-Prozeß aufbereitet. Man erhält zwei Matrizen: eine Content-Matrix mit den Dimensionen über 247 Webpages und 1258 Wörtern, sowie eine Usage-Matrix mit den Dimensionen 5291 Sessions und 247 Webpages. Aus der Multiplikation beider Matrizen erhält man eine Usage-Wort-Matrix, die angibt, welche Wörter innerhalb jeder Session gesehen wurden. Clustert man beispielsweise mit kMeans die Sessions nach Worten, erhält man Gruppen von Sessions, in denen inhaltlich ähnliche Webpages besucht wurden. Jede Session ist durch einen Vektor über alle Worte dargestellt, die die Häufigkeiten der Wörter wie in der Usage-Wort-Matrix beschreibt. Summiert man alle Vektoren innerhalb eines Clusters, erhält man einen Interessensvektor jedes Clusters, den *User Interest Vektor*.

User Interest-Vektor

Zwei dieser User Interest-Vektoren sind in Abb. 4.9 und 4.10 dargestellt. Vorher wurde von jedem Vektor der Durchschnitt über alle User Interest-Vektoren abgezogen. Dadurch wird das relative Gewicht, das ein Wort auf ein Thema hat, deutlich. Dieses relative Gewicht ist in Abb. 4.9 und 4.10 auf der vertikalen Achse abgetragen. Man erhält eine User-Interessen Matrix, indem die einzelnen User Interest-Vektoren, wie in Abb. 4.8 gezeigt, als Spalten dieser neuen Matrix angeordnet werden. In dieser Matrix ist für jedes Wort dessen Gewicht auf ein Thema abgetragen.

Durch Multiplikation mit der Content-Matrix (Webpage-Wort-Matrix) erhält man eine Zuordnung jeder Webpage zu den erkannten Themen, eine *Topic Map*. Diese neue Matrix beschreibt

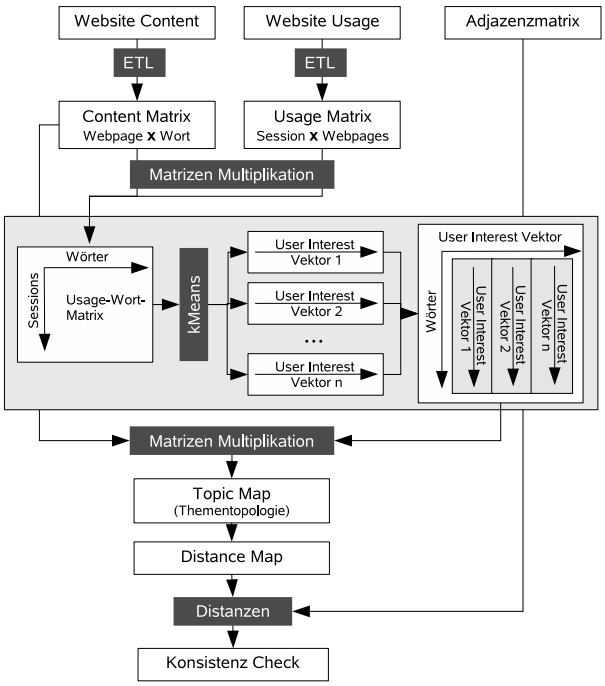


Abbildung 4.8: Berechnungsweg des Konsistenz Checks

die thematische Topologie der gesamten Website. Es lassen sich mit Hilfe dieser Matrix die inhaltlichen Abstände der Webpages zueinander berechnen, (*Distance Map*).

Konsistenz-Check

Die einzelnen Schritte lassen sich so interpretieren, daß die User einer Website dazu benutzt wurden, inhaltliche Zusammenhänge auf einer Website durch ihr Verhalten aufzudecken. Man unterstellt dabei, daß sich in den Aktionen eines Users dessen Interessen widerspiegeln. Man hat mit der Topic Map eine vom User wahrgenommene Thementopologie erstellt.

Dem wird nun die Sichtweise der Website Autoren gegenübergestellt. Wie unter anderem von Kleinberg et al. [148] wird auch hier unterstellt, daß ein Website-Autor bei der Konzeption der Website Struktur sein Domänenwissen einfließen läßt. Die Website-Struktur sollte als Gerüst für die inhaltliche Struktur dienen. Mit der Sicht der Website als Graph beschreibt die Adjazenzmatrix die Linkabstände zwischen allen Webpages auf einer Website. Hier wird von den kürzesten Wegen zwischen zwei Webpages ausgegangen.

Um beide Matrizen vergleichbar zu machen, normiert man die Distance Map und die Adjazenzmatrix auf den gleichen Zahlenbereich, beispielsweise $[0; 1]$. Der Konsistenz-Check be-

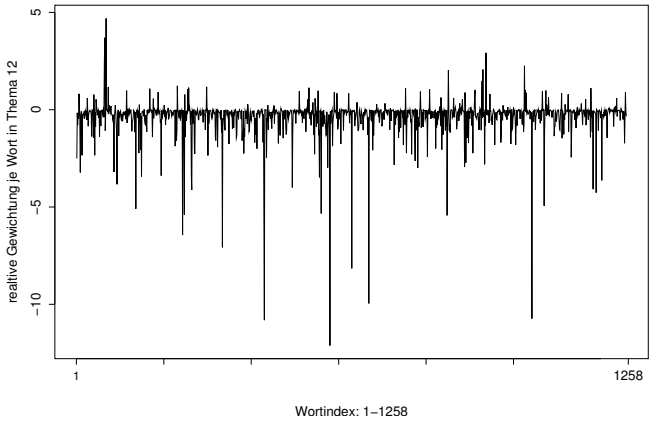


Abbildung 4.9: Themenvektor 1

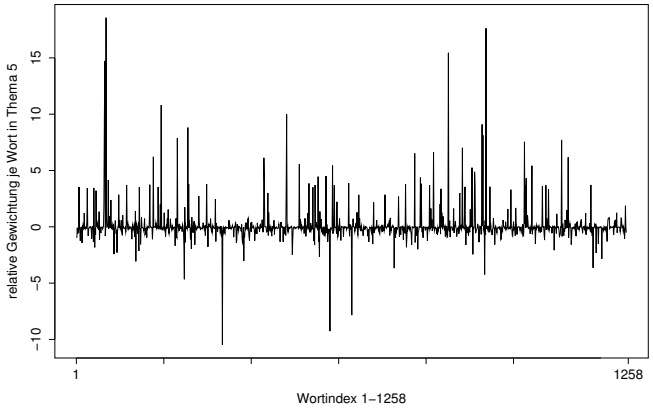


Abbildung 4.10: Themenvektor 2

steht darin, daß man die Differenz zwischen beiden Matrizen für jede Webpage-Kombination bildet. Große Differenzen zeigen ein Abweichen der inhaltlich wahrgenommenen Distanz von der Distanz an, die der Website Autor durch die Link-Struktur geschaffen hat. Die User-Wahrnehmung ist nicht mit der Ansicht des Website-Autors *konsistent*. Diese Abweichung ist ein Indikator für eine Überprüfung der Link-Struktur.

Mit dieser Arbeit [242] wurde gezeigt, wie sich Web Usage, Content und Structure Mining kombinieren lassen, um daraus die User-Wahrnehmung zu approximieren und Verbesserungspotentiale der Usability einer Website aufzudecken. Dieser Ansatz unterscheidet nicht zwischen transaktions- und informationsorientierten Websites. Er stellt einen Schritt auf dem Weg zu einem Erfolgsmaß für Websites im Allgemeinen dar. Vorteile dieser Lösung liegen in ihrem allgemein verwendbaren Ansatz auf alle Websites und in der weitgehend automatisierbaren Durchführung. Ein weiterer Vorteil besteht darin, daß gezielt Schwachstellen einer Website identifiziert werden, sodaß auch große Websites kontinuierlich gepflegt werden können.

4.6 Zusammenfassung

Wie bei der Analyse existierender Metriken und Maße läßt sich dieses Kapitel nur als teilweise erfolgreich zusammenfassen, wenn man die Erfolgsmessung informationsorientierter Websites betrachtet. Einen Ansatz zur Analyse und Beurteilung des Erfolgs von Websites liefern Web Mining-Analysen nicht, weder für transaktionsorientierte noch informationsorientierte Websites. Hier stehen andere Aufgabenstellungen im Vordergrund, wie die Personalisierung von Website-Inhalten und deren Struktur. Mit Beginn der Web Mining-Analysen waren die Methoden nach Usage-, Content- und Structure Mining noch klar getrennt. Da sich das Verhalten der User aber nur aus dem Zusammenhang der Website erklären läßt, bestehen die meisten aktuellen Analysen aus einer Kombination von Web Mining-Methoden.

Ohne valides Wissen über die Motivation der User sehen Schonberg et al. [210, S.56] nur die Möglichkeit, Annahmen über das Verhalten der User zu machen. Allerdings verweisen sie auf die gesammelten Daten, die einen konstanten Strom an verschiedenen Aspekten eines Feedbacks zulassen sollen.

Hahn et al. fordern in [112, 113], daß die Interpretation der gefundenen Web Usage Mining-Ergebnisse erst durch das Wissen um die User-Ziele erfolgreich sein kann. Ihre Arbeit ist auf alle Website-Typen bezogen und nicht nur auf transaktionsorientierte Websites. Sie beziehen sich auf die oben vorgestellte Informationsverbreitungstheorie von Pirolli et al. [192]. Bei einem gegebenen User-Ziel soll man den wahrscheinlichsten Navigationspfad voraussagen können. Umgekehrt soll man bei einem gegebenen Navigationspfad, sowie der Kenntnis der Struktur und des Inhalts einer Website, auf die User-Ziele rückschließen können.

Hahn et al. entwickeln in [112] ein Modell mit zwei Ebenen, der User-Perspektive und der Website-Perspektive. Aus dem Vergleich der User-Ziele mit den Zielen der Website sollen Schwachstellen erkannt und zur Maximierung des wirtschaftlichen Erfolges geändert werden.

In Kap. 4.5 wurde dieses Modell dadurch umgesetzt, indem die Sicht des Website-Autors in Gestalt der Website-Struktur mit der Wahrnehmung der User, repräsentiert durch die wahrgenommene, inhaltliche Topologie einer Website, miteinander verglichen wurde. Das Ziel von Hahn et al. [112], Schwachstellen zu erkennen, konnte damit erreicht werden. Daß sich auf diese Weise der wirtschaftliche Erfolg der Website maximiert, bleibt aber dahingestellt, da er weder hier noch bei Hahn et al. gemessen wird.

4.6.1 Ergebnisse

Somit lassen sich die Kapitel 3 über Web Metriken und Kapitel 4 über Web Mining zusammenfassen, daß es Maße und Metriken gibt, die Aspekte eines gewünschten User-Verhaltens

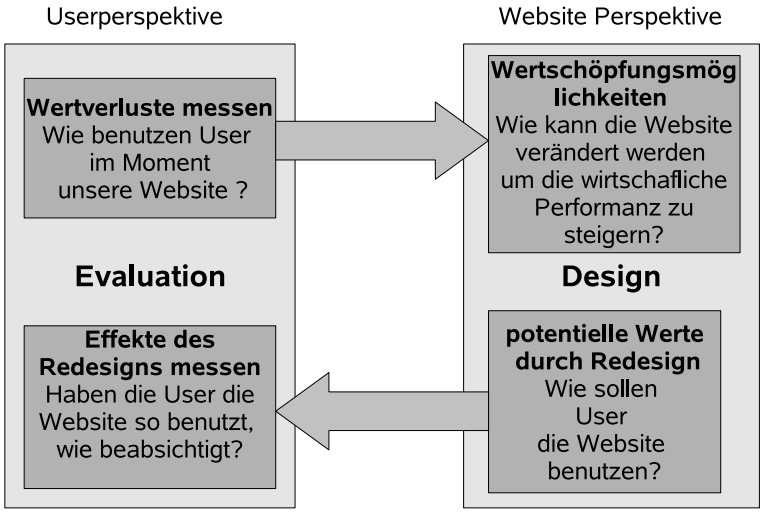


Abbildung 4.11: Website Redesign in Anlehnung an [112, Abb.2]

messen können, z.B. ein bestimmtes (bekanntes) Ziel zu verfolgen (Focus), den Grad der Zielerreichung (Task Completion Rate), ausreichend Zeit auf einer Webpage verbracht zu haben (Duration), in Abhängigkeit von der Aufgabe dieser Webpage (Hub oder Authority).

Über Kennzahlen hinaus gibt es Web Mining-Analysen, die das so bewertete User-Verhalten in den richtigen Zusammenhang einordnen können. So ist eine inhaltliche Analyse einer Website bzw. einer Session wichtig, um die Kennzahl Focus realisieren zu können.

Daraus folgt, daß im nächsten Kapitel die Möglichkeiten untersucht werden sollen, Web Metriken und Web Mining-Analysen zu kombinieren, um gewünschtes User-Verhalten zu erkennen und bewerten zu können.

4.6.2 Schlußfolgerungen

Web Metriken werden meist aus den direkt meßbaren Fakten erstellt, die durch Beobachtung des Users oder durch die inhaltliche und strukturelle Analyse einer Website gesammelt werden. Web Mining Analysen bauen auch auf diesen Daten auf. Sie kombinieren aber häufig Fakten und deren Interpration, um zu ihren Ergebnissen zu gelangen. Wie man in den Qualitätsmodellen von Riemer et al. [201] und Totz et al. [252] in Abb. 1.8 aus Kap. 1 gesehen hat, sind sowohl die Ziele der Website als auch der User unscharf formuliert. Daher soll bei der Erstellung eines eigenen Erfolgsmaßes im folgenden Kapitel 5 ein Modell einer Website beschrieben werden, in welchem klar zwischen Fakten auf der einen Seite und Interpretationen und Hypothesen auf der anderen Seite unterschieden wird. Somit soll klar werden, wann die Aussagen auf Fakten und wann auf Annahmen des Analysten beruhen.

Die Evaluation des Erfolgsmaßes soll durch unabhängige Methoden und Daten erfolgen, die nicht bei der Konstruktion des Maßes eingesetzt wurden. Daher wird vorgeschlagen, die Evaluation mittels explizit erhobener Daten in Form einer User-Studie durchzuführen.

Auch soll die Sichtweise aus Abb. 4.11 in diesem Kapitel, und aus Abb. 1.8 berücksichtigt werden. Beide beschreiben die unterschiedlichen Sichtweisen von Usern und Website-Autoren auf den Erfolg einer Website.

5 Erfolgsmaße für informationsorientierte Websites

In Kapitel 1 wurden die grundlegenden Internet-Geschäftsmodelle beschrieben und diskutiert. Aus deren Analyse ergab sich eine Lücke bei der Erfolgsmessung und Beurteilung von Websites, die keine Transaktionen anbieten, sondern Informationen bereitstellen. Darin liegt der Ausgangspunkt für diese Arbeit.

In diesem Kapitel soll untersucht werden, wie aus den Daten aus Kapitel 2 mit Hilfe von Kennzahlen aus Kapitel 3 und Web Mining Analysen aus Kapitel 4 der Erfolg derartiger Websites gemessen werden kann.

Web Metriken und Maße, die auf Transaktionen basieren, sind bekannt und in ihrer Anwendung etabliert. Vorhandene Maße können auf informationsorientierten Websites nur Teilaspekte erfolgreicher Sessions beschreiben. Aus den Schlußfolgerungen in 4.6.1 sind bei der Erstellung eines Erfolgsmaßes in diesem Kapitel folgende Anforderungen zu berücksichtigen:

- Bewertung der Website und ihrer Benutzung aus Website- und User-Sicht.
- klare Trennung von Fakten, Annahmen und Interpretationen
- Anwendung auf reale Daten
- Evaluation durch eine systematisch unterschiedliche Erhebungsform, beispielsweise eine explizite User-Befragung

An diesen Anforderungen orientiert sich die Vorgehensweise in diesem Kapitel. In Kapitel 5.1 sollen die in den vorigen Kapiteln gewonnenen Einsichten und Erkenntnisse bei der Erstellung eines Erfolgsmaßes einfließen. Zur Verdeutlichung der Zusammenhänge der Website-Elemente und ihrer Benutzung wird ein formales Modell in Kapitel 5.1.6 erstellt.

Anschließend wird das Erfolgsmaß auf die Websites eines großen deutschen Unternehmens angewandt. Durch diese Anwendung werden die Details bei der Berechnung an Beispielen erklärt. Die gewonnenen Ergebnisse werde in Kapitel 5.3 durch eine User-Befragung evaluiert. Da die User-Sicht nicht in allen Fällen mit der Erfolgsbewertung aus Sicht der Website übereinstimmt, wird in 5.4 das Erfolgsmaß unter Berücksichtigung der User-Sicht erweitert und auf eine solide, formale Grundlage gestellt.

5.1 Guidance Performance Indicator GPI

5.1.1 Lösungsansatz

Der User einer informationsorientierten Website ist daran interessiert, sein Informationsbedürfnis zu stillen. Der Anbieter einer Website möchte dem User Informationen vermitteln. Dabei ist es zwar wünschenswert, daß der User die von ihm gewünschten Informationen möglichst schnell und intuitiv findet, andererseits ist auch die Vermittlung anderer Themen im Sinne der Website. Findet der User die gesuchten Informationen nicht, dafür aber andere Inhalte, hat die Website dennoch ein Teilziel erreicht.

In Kapitel 3.1 wurden die Einschränkungen beschrieben, die man bei der Beobachtung des User-Verhaltens auf der Server-Seite hat. Für die Erstellung eines Erfolgsmaßes stehen lediglich diese serverseitigen Daten zur Verfügung. Die Aufmerksamkeit eines Users und dessen Zufriedenheit kann darin nicht direkt gemessen werden. Aus Sicht der Website ist nur erfaßbar, ob ein User durch Erreichen einer bestimmten Webpage die Möglichkeit hatte, den dort präsentierten Inhalt wahrzunehmen.

Da die Ziele und Interessen eines Users ohne zusätzliche Daten auf der Server-Seite nicht zugänglich sind, wird ein Erfolgsmaß vorgeschlagen, das die Zielerreichung einer Website aus Website-Sicht misst.

Jeder Click eines Users soll danach bewertet werden, ob er für die Zielerfüllung der Website hilfreich war. Bewegt sich ein User auf ein Ziel zu, d.h. erreicht er Informationen und hat ausreichend Zeit, diese wahrzunehmen, erhält dieser Click eine positive Bewertung. Die Höhe der Bewertung soll von der Besuchsdauer dieser Webpage abhängen. Auch der Inhalt einer besuchten Webpage soll darauf hin untersucht werden, ob er zum Inhalt der restlichen Session passt. Ein Click, der nicht der Zielerfüllung der Website dient, soll negativ bewertet werden.

Abb. 5.1 zeigt die Vorgehensweise, mit der in diesem Abschnitt ein Erfolgsmaß berechnet wird. Nach der Beschreibung der verwendeten Daten in 5.1.2 werden in 5.1.3 und 5.1.4 zwei Teilmaße erstellt, die zum einen die Effektivität eines Clicks und zum anderen die Effizienz eines Clicks bewerten.

5.1.2 Daten

Für den Erfolg einer Website ist der Bewegungspfad des Users auf einer Website, die Aufenthaltsdauer auf den einzelnen Webpages und der Inhalt dieser Seiten ausschlaggebend. In Kapitel 2 wurden die Datensammlung und -aufbereitung beschrieben, so daß jetzt auf eine bereinigte Datenbasis zugegriffen werden kann.

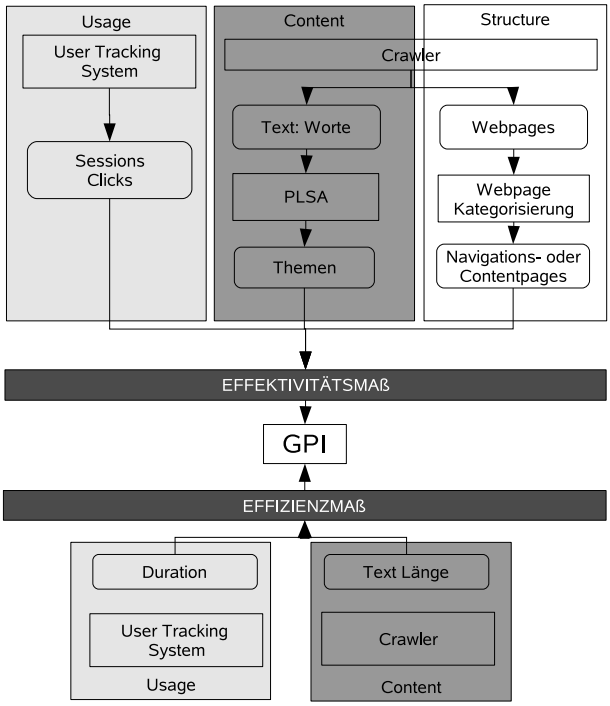


Abbildung 5.1: Berechnung des GPI

Usage Wie man durch die Usage-Analysen in Kapitel 4.2 gesehen hat, ist bei der Betrachtung eines Clicks der Zusammenhang der gesamten Session wichtig. Daher soll ein Erfolgsmaß bei der Bewertung eines Clicks den kompletten Clickstream einer Session berücksichtigen.

Duration In Kapitel 3.3.1 wurde auf die eingeschränkte Aussagefähigkeit der Besuchsdauer (Duration) einer Webpage hingewiesen. Die Duration kann darüber Auskunft geben, inwiefern ein User die Möglichkeit hatte, den Inhalt einer Webpage wahrnehmen zu können. Inwiefern er den Inhalt tatsächlich wahrgenommen hat, bleibt auf der Server-Seite unbekannt.

Die Duration der letzten Webpage einer Session kann gemäß der Erklärungen in Kapitel 3.3.1 nicht erfaßt werden und wird mit einer Duration von 1 Sekunde bemessen.

Content Wie man in der Clickstreamanalyse in Kapitel 4.2.2 gesehen hat, ist ein Click nicht nur als Clickstream einer Session zu betrachten, sondern auch im inhaltlichen Zusammenhang der Website zu analysieren.

Structure Auch die Struktur der Website soll bei der Bewertung eines Clicks mit einfließen, indem die Webpages anhand ihrer Aufgabe innerhalb der Website kategorisiert werden. Dabei orientieren sich die Kategorien an der Arbeit von Spiliopoulou et al. [225], die zwischen Aktions- und Zielseiten unterscheiden. Eine Aktionsseite soll auf informationsorientierten Websites den User zum geeigneten Content leiten und eine intuitive Navigation ermöglichen. Sie wird der Kategorie *Navigationsseite* zugeordnet. Eine Zielseite [225] entspricht einer *Contentseite*, die die Informationen bereithält.

Aus den beschriebenen Daten soll im folgenden der Erfolg einer informationsorientierten Website ermittelt werden. Jeder Click soll danach bewertet werden, ob er zur Zielerreichung der Website beiträgt. Das Problem, die Zielerreichung zu messen, wird in zwei Aspekte aufgeteilt. Zum einen soll die **Effektivität** und zum anderen die **Effizienz** eines Clicks hinsichtlich der Zielerreichung bewertet werden.

Die Effektivität beschreibt die Richtung der Bewegung des Users auf Informationen zu oder von ihnen weg. Die Effizienz beschreibt den Umfang der Zielerreichung, also wie weit sich der User in Richtung einer erfolgreichen Informationsvermittlung bewegt hat.

5.1.3 Effektivität eines Clicks

In diesem Abschnitt wird die Effektivität eines Clicks bewertet. Die Effizienzbewertung wird in 5.1.4 beschrieben.

Nicht jede Webpage dient dazu, Informationen zu vermitteln. Zur Strukturierung einer Website und der intuitiven Zielführung eines Users bedarf es Webpages, die die Navigation des Users unterstützen. In Anlehnung an Spiliopoulou et al. [225] wird zwischen Navigations- und Contentseiten unterschieden. Navigationsseiten sollen den User möglichst intuitiv zu Contentseiten leiten. Contentseiten sollen dem User die Informationen präsentieren, die der Website-Besitzer verbreiten möchte.

Navigations- und Contentseiten müssen, wie bei Spiliopoulou et al. [225] vorgeschlagen, für jede Website individuell identifiziert werden - meist vom Website-Besitzer. Bei der manuellen Klassifikation wurde die Homepage, sowie Sitemap und Suchseiten als Navigationsseiten klassifiziert. Die restlichen Seiten werden der Kategorie Contentseite zugeordnet. Die Bewertung der Transitionen zwischen den verschiedenen Webpage Kategorien wurde gemäß Tab. 5.1 vorgenommen. Unter einer Transition wird in diesem Zusammenhang der Übergang von einer Webpage auf eine andere verstanden. Wird ein User von einer Navigationsseite auf eine weitere Navigationsseite geleitet, hat erstere ihre Aufgabe nicht erfüllt, nämlich den User zu Contentseiten zu leiten. Daher wird diese Transition negativ bewertet. Noch negativer

		Ziel				
		Home	Sitemap	Search	Content	Sess. End
Start	Home	-	-	-	+	-
	Sitemap	-	-	-	+	-
	Search	-	-	-	+	-
	Content	-	-	-	+	0

Tabelle 5.1: Bewertung von Transitionen zwischen Webpage Typen

ist es, wenn eine Navigationsseite am Ende einer Session steht, denn dann hat diese Navigationsseite den User veranlaßt, seine Session abzubrechen. Damit besteht keine Möglichkeit mehr, den User zum Content zu leiten.

Leitet eine Navigationsseite einen User auf eine Contentseite, ist dies ein positiver Beitrag zum Website-Ziel. Endet eine User-Session mit dem Besuch einer Contentseite, sind zwei Szenarien denkbar: Der User hat gefunden, wonach er sucht, bzw. genau das Gegenteil, er hat es nicht gefunden. Da die Website aber ihren Auftrag, Content zu vermitteln, erfüllt hat, wird auch dieser Click aus der Website-Perspektive positiv bewertet.

Der Übergang zwischen Contentseiten kann nicht direkt beurteilt werden. Eine Transition zwischen Contentseiten ist in jedem Fall gewünscht und wird daher positiv bewertet. Die Bewertung kann wesentlich genauer sein, wenn der inhaltliche Zusammenhang der User-Session berücksichtigt wird. Wie dies erreicht werden kann, wird bei der Berechnung des Effektivitätsmaßes gezeigt.

Eine beispielhafte Berechnung des Effektivitätstyps zeigt Abb. 5.2.

Hits-Algorithmus Anstelle einer manuellen Kategorisierung der Webpages zwischen Navigations- und Contentseiten können andere Eigenschaften der Website herangezogen werden. Zum Beispiel kann der Charakter eine Website bezüglich Navigation oder Contentvermittlung durch den Hits-Algorithmus von Kleinberg [148] dargestellt werden. Dieser Algorithmus wurde in Kapitel 3.4.3 beschrieben. Hierbei werden durch Analyse der Verlinkung der Website die Zahl der eingehenden und ausgehenden Links miteinander verglichen. Die so identifizierten *Hubs* dienen hauptsächlich der Navigation, während *Authorities* Inhalte anbieten.

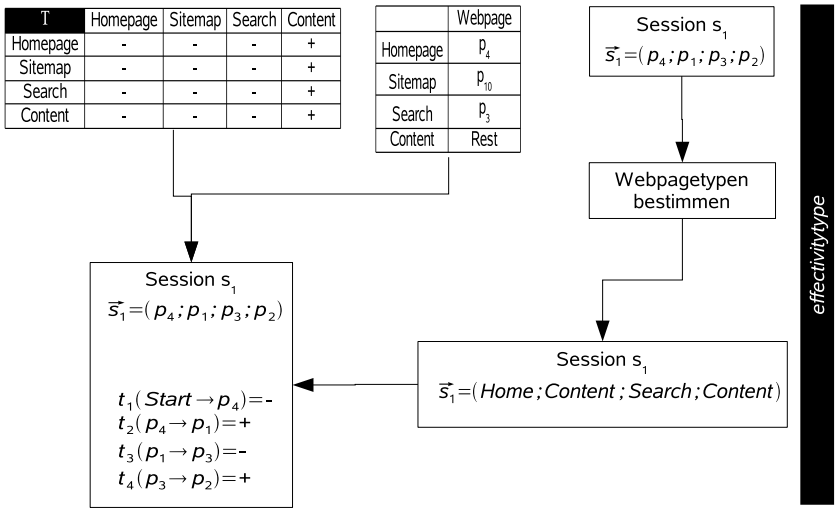


Abbildung 5.2: Beispielhafte Berechnung des Effektivitätstyps

Diese generische Herangehensweise benötigt kein manuelles Eingreifen und kann auf alle Websites angewandt werden. Allerdings kann bei einer manuellen Kategorisierung das Domänenwissen der Website Besitzer einfließen. Nachteile dieses Link-basierten Ansatzes, wie von Ding et al. in [85] beschrieben, liegen vor allem in der Annahme eines gerichteten Webgraphen durch die Verlinkung. Diese Annahme wird durch den clientseitigen Gebrauch des Back-Button, gespeicherte Links oder die History-Funktion eines Browsers, hinfällig, siehe Kap. 2.2.3. Daher wird auf eine Integration des Hits-Algorithmus verzichtet.

Themenübergänge Mit der inhaltlichen Analyse von Übergängen zwischen Contentseiten sollen die User-Sessions dahingehend untersucht werden, ob eine Thementransition durch einen Click innerhalb oder außerhalb des jeweiligen User-Fokuses liegt. Webpages, die innerhalb des User-Fokuses liegen, sollen höher bewertet werden als diejenigen außerhalb.

Für diese Analyse benötigt man die Themen, denen die Webpages zugehören. In 4.3.4 wurde der PLSA-Algorithmus vorgestellt, der zur Identifikation der Themen einer Website genutzt werden kann.

Jeder Webpage wird dabei ein Themenvektor zugeordnet. Aus der Differenz der Themenvektoren zweier Webpages kann der Themenübergang ermittelt werden. Um festzustellen, ob eine Webpage innerhalb oder außerhalb des User-Fokuses liegt, wird der Durchschnitt über die Themenvektoren der Webpages einer Session gebildet. Dieser Themenvektor beschreibt den User-Fokus. Der Inhalt aller in einer Session besuchten Webpages wird als Richtwert für die Themen genommen, die der User auf der Website verfolgt hat. Ausreißer erhalten so ein geringeres Gewicht.

Hat der User nur Webpages zu einem Thema aufgesucht, wird der durchschnittliche Themenvektor nur gering davon abweichen. Die Session ist somit an einem Thema interessiert. Ist in der Session eine Webpage, die aus diesem Themenbereich herausfällt, passt sie nicht in diese Session. Sie soll zwar immer noch positiv bewertet werden, aber nicht mehr so hoch wie Webpages, die im User-Fokus liegen, da sich hierfür der User über mehrere Clicks hinweg interessiert hat.

Hat ein User sehr unterschiedliche Webpages zu unterschiedlichen Themen besucht, wird auch der durchschnittliche Themenvektor ein sehr breites Themengebiet abdecken, so daß es eher wenige Ausreißer geben wird.

Durch die Art, wie der User-Fokus angelegt und berechnet wird, werden bestimmte User-Typen erkannt und danach die Themenübergänge bewertet. Ohne diese Berücksichtigung des Session-Inhalts würde ein User, der an vielen Themen interessiert ist, gegenüber einem thematisch fokusierten User benachteiligt.

Der Website-Besitzer muß einen Schwellenwert festlegen, über dem eine Themenänderung als Ausreißer erkannt wird. An diesem Schwellenwert wird jede Themenänderung einer Session bewertet, um thematische Ausreißer zu erkennen.

Berechnung des Effektivitätsmaßes

Das Effektivitätsmaß wird aus zwei Bestandteilen aufgebaut: dem Transitionstyp und dem Transitionsgewicht. Der Transitionstyp wird durch Tab. 5.1 festgelegt und läßt sich folgendermaßen formal beschreiben:

Definition: Transitionstyp *Sei eine Transition auf der beobachteten Website und* T *die dafür erstellte Transitionstabelle, dann sei der Transitionstyp effectivitytype definiert als*

$$effectivitytype = 1 \quad sign(T_{ij}) \tag{5.1}$$

wobei i der Typ der Startseite der Transition sei und j der Transitionstyp der Zielseite.

Auf diese Weise wird das Vorzeichen des Effektivitätsmaßes festgelegt. Das Transitionsgewicht wird für Transitionen zwischen Contentseiten durch die Themenänderung festgelegt. Bei Transitionen von oder auf eine Navigationsseite wird das Gewicht danach vorgegeben, wie stark eine derartige Transition bewertet bzw. bestraft werden soll.

Definition: Transitionsgewicht *Ist eine Transition zwischen Contentseiten (C → C), dann wird das Transitionsgewicht effectivityweight festgestellt, indem die Themenänderung effectivity$_{GPI}$( ) mit dem Schwellenwert effectivity$_{GPI}(sig)$ in Bezug auf die Themenmischung der Session verglichen wird.*
Ist eine Transition zwischen Navigationsseiten (N → N), wird ein festgelegter Wert zugewiesen, was durch das vorher festgelegte Vorzeichen zu einer negativen Bewertung führt.
Ist eine Transition von einer Navigations- auf eine Contentseite (N → C), wird ein festgelegter positiver Wert zugewiesen.
Ist eine Transition von einer Content- auf eine Navigationsseite (C → N), wird ein festgelegter negativer Wert zugewiesen.
Ist eine Transition, die das Ende einer Session auf einer Navigationsseite beschreibt (N → ⊥), wird ein festgelegter höherer, negativer Wert zugewiesen.
Ist eine Transition, die das Ende einer Session auf einer Contentseite beschreibt (C → ⊥), wird ein positiver Wert zugewiesen.

$$
\textit{effectivityweight} = \begin{cases}
C \rightarrow C : erhebliche\ Themenänderung \\
\quad \textit{effectivity}_{GPI}(\ ) > \textit{effectivity}_{GPI}(sig) \\
\quad \Rightarrow 1 < \textit{effectivityweight} \\
C \rightarrow C : unerhebliche\ Themenänderung \\
\quad \textit{effectivity}_{GPI}(\ ) <= \textit{effectivity}_{GPI}(sig) \\
\quad \Rightarrow 0 < \textit{effectivityweight} < 1 \\
N \rightarrow N :\Rightarrow \textit{effectivityweight} << 0 \\
N \rightarrow C :\Rightarrow \textit{effectivityweight} > 0 \\
C \rightarrow N :\Rightarrow \textit{effectivityweight} < 0 \\
N \rightarrow \perp :\Rightarrow \textit{effectivityweight} << 0 \\
C \rightarrow \perp :\Rightarrow \textit{effectivityweight} > 0
\end{cases} \tag{5.2}
$$

Eine beispielhafte Berechnung des Effektivitätsgewichts zeigt Abb. 5.3. Das Effektivitätsgewicht wird unterschiedlich berechnet, je nachdem zwischen welchen Webpage-Typen man sich bewegt. Besondere Beachtung finden die Transitionen zwischen Contentpages.

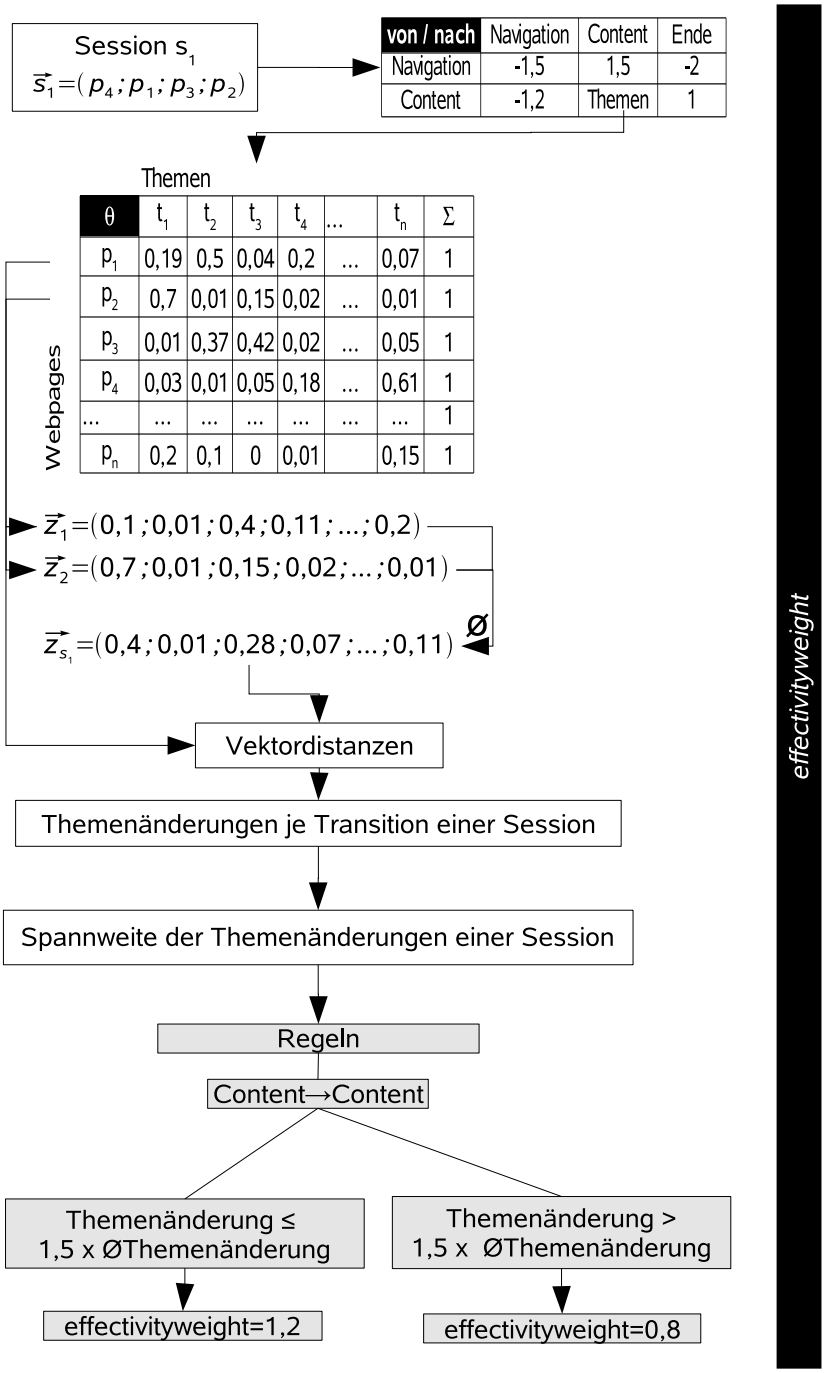


Abbildung 5.3: Beispielhafte Berechnung des Effektivitätsgewichts

5.1.4 Effizienz eines Clicks

Nachdem mit der Effektivität die Zielführung des Users betrachtet wurde, untersucht dieser Abschnitt die Effizienz. Damit wird bewertet, ob der User die Webpage, auf die er geführt wurde, effizient genutzt hat, bzw. die Möglichkeit dazu hatte. Eine Webpage kann vom User effizient wahrgenommen werden, wenn dessen Aufenthalt lang genug oder kurz genug war, je nach Webpage-Typ. Tab. 5.2 zeigt die Bewertung der Aufenthaltsdauer (Duration) in Abhängigkeit des Webpage-Typs.

	kurz	lang
Navigation	+	-
Content	-	+

Tabelle 5.2: Webpage Duration Rating

Ein langer Aufenthalt auf einer Contentseite ist ebenso wünschenswert, wie eine schnelle Weiterleitung von einer Navigationsseite, d.h. eine kurze Duration auf der Navigationsseite. Dies entspricht dem Wunsch nach einer hohen Stickiness auf Contentseiten und einer hohen Slipperiness auf Navigationsseiten (Kapitel 3.3.1).

Das Effizienzmaß soll mit dem Effektivitätsmaß so kombiniert werden, daß eine schnelle Weiterleitung von einer Navigations- auf eine Contentseite sehr positiv bewertet wird. Der Schritt von einer Content- auf eine Navigationsseite wird nach kurzem Aufenthalt auf der Contentseite schlecht bewertet.

Um die Duration zwischen unterschiedlichen Webpages vergleichbar machen zu können, wird die Duration über die Textlänge der jeweiligen Webpage normiert. Um die gesamte Gestaltung einer Webpage miteinzubeziehen, wird vorgeschlagen, die durchschnittliche Duration aller User für jede Webpage zu berechnen und diesen Durchschnitt als Indikator für die zu erwartende Aufenthaltszeit zu benutzen. So kann die Wahrnehmung einer Webpage durch die User miteinbezogen werden.

Berechnung des Effizienzmaßes

Da die Duration der letzten Webpage einer Session nicht gemessen werden kann (siehe Kap. 2), wird dem letzten Click einer Session der Wert 0 zugewiesen. Dadurch würde dieser Click überhaupt nicht gewertet werden. Alternativ kann ein geringes Gewicht z.B. 0.1 zugewiesen werden, wenn man den letzten Click einfließen lassen möchte.

Definition: Effizienzmaß (Durationgewicht) *Sei eine Transition auf der beobachteten Website. Sei d die Duration der betrachteten Transition , d_{avg_p} die durchschnittliche Duration aller User auf Webpage p dann gilt entsprechend Tab. 5.2 für das Durationgewicht und Effizienzmaß efficiency:*

$$\textit{efficiency} = \begin{cases} C: \begin{cases} d >= d_{avg_p} \rightarrow \textit{efficiency} > 1 \\ d < d_{avg_p} \rightarrow 0 < \textit{efficiency} < 1 \end{cases} \\ N: \begin{cases} d >= d_{avg_p} \rightarrow \textit{efficiency} < 0 \\ d < d_{avg_p} \rightarrow 0 < \textit{efficiency} > 0 \end{cases} \\ \perp : d = 0 \end{cases} \tag{5.3}$$

Die genauen Gewichte werden manuell in den oben beschriebenen Bereichen festgelegt. Eine beispielhafte Berechnung des Effizienzmaßes zeigt Abb. 5.4. Für jede Webpage wird die durchschnittliche Aufenthaltsdauer aller Besucher dieser Webpage ermittelt. Die Aufenthaltsdauer wird gemäß der Definition des Effizienzmaßes nach den in 5.1.4 vorgegebenen Regeln bewertet. So werden kurze Durations auf Navigationsseiten und lange Durations auf Contentseiten "belohnt" und im umgekehrten Fall mit niedrigen Gewichtungen "bestraft".

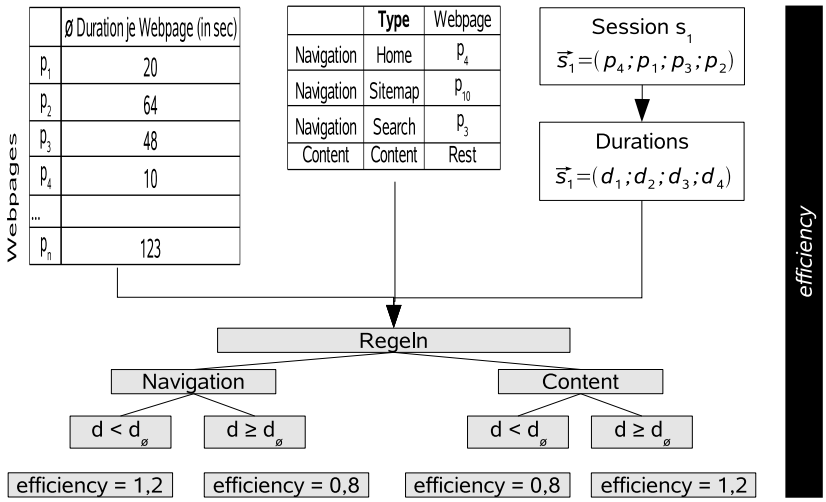


Abbildung 5.4: Beispielhafte Berechnung des Effizienzmaßes

5.1.5 Berechnung des Erfolgsmaßes

Mit dem Effektivitätsmaß und dem Effizienzmaß sind zwei Teilaspekte einer erfolgreichen Informationsvermittlung erfaßt worden: 1. Die gezielte Hinführung zu Informationen und 2. die Möglichkeit, diese Informationen ausreichend wahrnehmen zu können. Da mit den Teilmaßen das Ziel der Website gemessen wird, einen User zu Content zu leiten, wird das daraus ermittelte Gesamtmaß als **G**uidance **P**erformance **I**ndicator bezeichnet.

In Abb. 5.1 wurden die Daten und Analysen in mehreren Prozeßschritten zur Berechnung des GPI zusammengefaßt. Aus Effektivitäts- und Effizienzmaß wird ein Erfolgsmaß für informationsorientierte Websites berechnet.

Definition: Guidance Performance Indicator (GPI) *Kombiniert man die Effektivitätsmaße Transitionstyp (effectivitytype) und Transitionsgewicht (effectivityweight) mit dem Effizienzmaß Durationgewicht (efficiency), erhält man als deren Produkt den GPI-Wert für eine Transition. Die drei Teilmaße wurden so gewählt, daß sie multipliziert den GPI-Wert für eine Transition ergeben.*

Summiert man über die Transitionen einer User Session s, erhält man den GPI dieser Session s

$$GPI_s = \sum_{i=1,\ i \in s}^{|s|} \textit{effectivitytype}_{\ i} \quad \textit{effectivityweight}_i \quad \textit{efficiency}_i \tag{5.4}$$

Summiert man über die Transitionen auf einer Webpage p, erhält man den GPI dieser Webpage

$$GPI_p = \sum_{i=1,\ i \in p}^{|p|} \textit{effectivitytype}_{\ i} \quad \textit{effectivityweight}_i \quad \textit{efficiency}_i \tag{5.5}$$

5.1.6 Modell einer Website und ihrer Benutzung

Die obige Beschreibung des GPI umfaßt zahlreiche Parameter, Annahmen und Interpretationen. Um diese voneinander zu trennen und übersichtlicher darzustellen, sollen sie in einem formalen Modell der Website und ihrer Benutzung abgebildet werden. Abb. 5.5 zeigt dieses Modell in Form eines UML-Diagramms (Unified Modelling Language).

Das Modell besteht aus drei Ebenen (siehe Abb. 5.6,5.7): der syntaktischen und semantischen Ebene sowie der Maßebene.

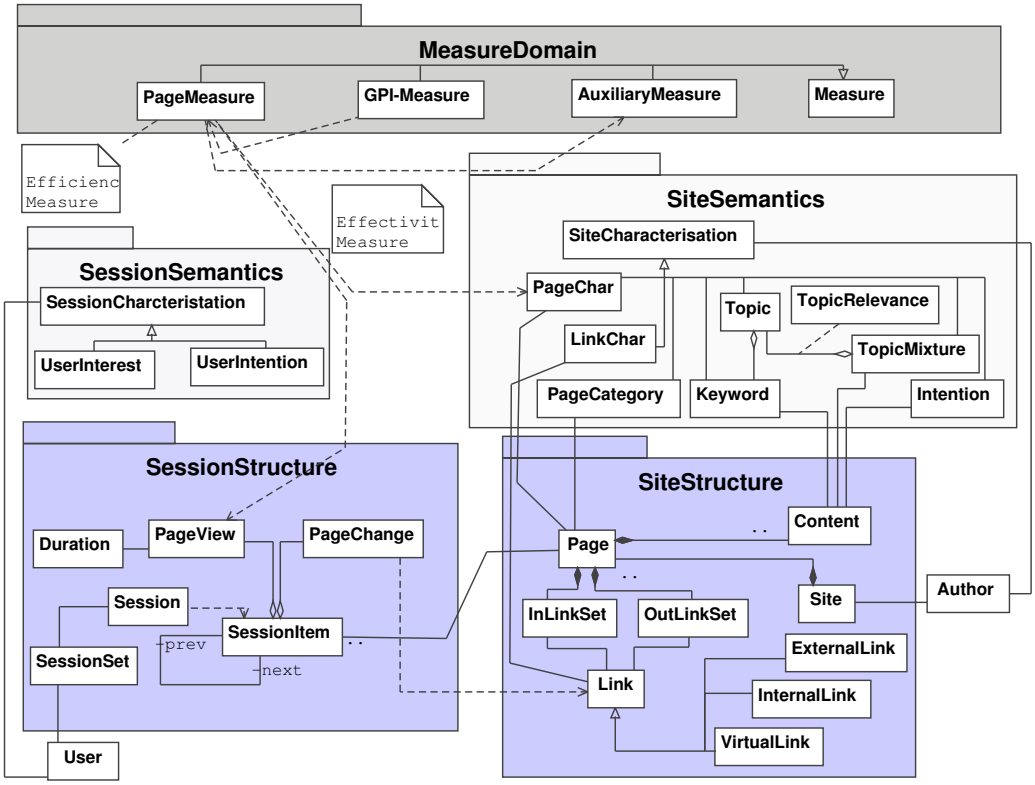


Abbildung 5.5: Formales Modell einer Website zur Berechnung des GPI

Die **syntaktische Ebene** beschreibt die direkt beobachtbaren Fakten einer Website und ihrer Benutzung. Sie bildet die Basis des Modells.

Die **semantische Ebene** ist oberhalb der Syntax angeordnet. Sie besteht aus den Annahmen und Interpretationen, die bei der Erstellung des GPI getroffen wurden.

Die oberste Ebene bildet die **Maßebene**. Der GPI und dessen Teilmaße greifen auf die darunterliegenden Fakten der Syntax und die Annahmen der Semantik zurück.

Das Website-Modell unterscheidet außerdem, auf Ebene der Syntax und Semantik, zwischen **Usern** und **Website**. Die **SessionStructure** beschreibt die Syntax auf der User-Seite. Sie besteht aus den Parametern des Clickstreams (Pageview, Session, Duration), den ein User bei seinem Besuch auf der Website hinterläßt. Diese *Session* kann aus mehreren *SessionItems* bestehen, die einem Click (Pageview) entsprechen. Kann ein wiederkehrender User als solcher identifiziert werden, können ihm mehrere Sessions zugeordnet werden, die ein *SessionSet* bilden. Neben dem einzelnen Click ist die Verweildauer, *Duration*, ein wichtiger Parameter einer User Session. Die Berechnung und Einschränkungen des Maßes Duration wurden in 3.3.1 beschrieben.

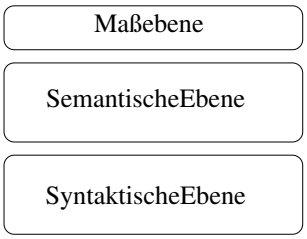


Abbildung 5.6: Ebenen des Website Modells

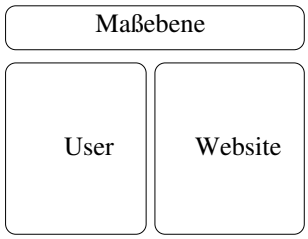


Abbildung 5.7: Seiten des Website Modells

Über der SessionStructure liegt die **SessionSemantics** genannte Semantik der User-Seite. Diesem Teil des Modells sind die Annahmen und Interpretationen über den User zugeordnet. Da auf der Server-Seite, wie in Kapitel 3.1 beschrieben, das User-Verhalten nur indirekt beobachtet werden kann und über die Absichten des Users Unklarheit herrscht, können in der Semantik nur wenige begründbare Annahmen über den User getroffen werden.

Die Website Syntax wird in **SiteStructure** auf der rechten Seite beschrieben. Die Struktur einer Website wird durch die Verlinkung der *Webpages* bestimmt. Es wird die Menge der eingehenden Links (*InLinkSet*) von der Menge der ausgehenden Links (*OutLinkSet*) unterschieden. Liegen Ursprung und Ziel eines Links innerhalb der Website, spricht man von einem *InternalLink*. Ein *ExternalLink* liegt vor, wenn entweder Ursprung oder Ziel außerhalb der Website liegen, das heißt von einer oder auf eine externe Website verlinkt wird. Das Modell ist so angelegt, daß beispielsweise der Hits-Algorithmus durch die In-Out-Links problemlos integriert werden kann.

Ein weiterer Parameter einer Website bzw. Webpage ist deren Inhalt (*Content*). Da der Inhalt von jeder Person subjektiv unterschiedlich wahrgenommen wird, ist die nähere Beschreibung des Website-Inhalts in der semantischen Schicht angeordnet.

Die **SiteSemantics** umfaßt die Annahmen, die über den Inhalt der Website getroffen wurden. Der Inhalt einer Website bzw. Webpage beschränkt sich in dieser Arbeit auf die Informationen, die im Textformat vorliegen. Der Inhalt wird einem oder mehreren Themen (*Topics*) zugewiesen, sodaß der Inhalt einer Webpage durch eine Themenmischung (*TopicMixture*) beschrieben wird. Die Topics werden aus einzelnen Wörtern (*Keywords*) erstellt. Zusätzlich wird angenommen, daß der Autor einer Website mit dem Inhalt eine bestimmte Absicht verfolgt (*Intention*). Die Kategorisierung nach Webpage Typen zur Bestimmung des Effektivitätsmaßes gehört ebenfalls zu SiteSemantics. Zur Berücksichtigung der manuellen Webpage-Kategorisierung wird jede Webpage einer Kategorie (*PageCategory*)zugeordnet und deren Links einer *(LinkChar)*-akterisierung.

In der **Maßebene** liegen die Web Maße, die sowohl aus der Syntax als auch der Semantik einer Website berechnet werden. Man kann durch das Modell genau erkennen, welche Bestandteile

eines Maßes auf Fakten und welche auf Annahmen beruhen.

Das formale Modell einer Website dient in diesem Entwicklungsstadium mehr der Veranschaulichung und Verdeutlichung von Fakten und Interpretationen. In [15] wurde der Nutzen eines Modells durch die formale Herleitung des Effektivitätstyps gezeigt.

Ein weiterer Vorteil wird darin deutlich, daß an das von Schwickert et al. [213, S.3] beschriebene Problem des fehlenden mathematischen Zusammenhangs zwischen Web Kennzahlen durch ein solches Modell herangegangen werden kann. In Kapitel 5.4 wird dieser Vorteil dadurch deutlich, daß durch die Erweiterung des Website Modells ein erweitertes Effektivitätsmaß entwickelt werden konnte.

5.1.7 Ergebnisse

In diesem Abschnitt wird erläutert, wie aus den serverseitig verfügbaren Daten einer Website und ihrer Benutzung ein Erfolgsmaß für informationsorientierte Websites aus Sicht der Website erzeugt werden kann.

Der GPI kombiniert Usage-Daten mit Structure-Daten durch Webpage-Kategorien und Content-Daten mit Hilfe des PLSA. Die genaue Parameterisierung und Bewertungen von gewünschtem und nicht gewünschtem User-Verhalten soll in einer Fallstudie im folgenden Abschnitt 5.2 beschrieben werden. Dabei kommt der GPI auf informationsorientierten Websites eines Unternehmens zum Einsatz.

5.2 Fallstudie

Die Fallstudie wird gemäß des Web Mining-Prozesses durchgeführt, der in 4.1 beschrieben wurde. Dieser besteht aus 1. der Festlegung der Ziele der Data Mining-Analyse, 2. der Datenaufbereitung und -transformation, 3. der Anwendung von Web Mining-Methoden, 4. der abschließenden Interpretation der Ergebnisse und 5. deren Umsetzung.

Die Fallstudie zeigt nur eine mögliche Ausprägung des GPI anhand der Wahl der Effekti-vi täts- und Effizienzgewichte. Ziel dieser Arbeit ist die Entwicklung eines generischen Modells zur Bewertung des Erfolgs von informationsorientierten Websites. Die genaue Kalibrierung der einzelnen Gewichte muß bei Anwendung auf anderen Websites erneut erfolgen. Die verwendeten Gewichte spiegeln hier die Einschätzung des Website-Autors wider und können alternativ durch eine empirische Untersuchung festgelegt werden. Die Grundlage der Festlegung bildet aber weiterhin das in 5.1.6 entwickelte Website-Modell.

5.2.1 Ziel

Mit dieser Fallstudie soll die Berechnung des GPI und dessen Anwendbarkeit auf informationsorientierten Websites gezeigt werden. Dazu werden drei Websites in jeweils zwei Zeiträumen analysiert.

5.2.2 Datensammlung

In Kap. 2 wurde die Bereitstellung und Aufbereitung der Daten beschrieben. Es werden die Usage-, Content- und Structure Daten mehrerer Websites eines Unternehmens verwendet und wie in Kapitel 2 aufbereitet. Tab. 5.3 beschreibt die Rohdaten von zwei Websites aus jeweils zwei Zeiträumen. 45 Tage von November bis Mitte Dezember 2005 sowie 30 Tage im Januar 2006. Die analysierten Websites unterscheiden sich in Größe und Besucherzahl. Website B ist hinsichtlich Zahl der Webpages und Besucherzahl die kleinere der beiden Websites.

5.2.3 Datenaufbereitung und -transformation

Die Rohdaten müssen vor der Analyse aufbereitet und bereinigt werden. Dies wurde in Kap. 2 beschrieben. Hier soll nur auf die Besonderheiten eingegangen werden, die bei diesen Websites aufgetreten sind. Beispielsweise können nur diejenigen Sessions ausgewertet werden, die auch vom Crawler erfaßt wurden. Umgekehrt können auch nur diejenigen Webpages bewertet werden, die in diesem Zeitraum den Usern zugänglich waren. Zuerst werden die Content-

	Website A		Website B	
	11.05-12.05	**01.2006**	**11.05-12.05**	**01.2006**
Sessions	9.877	5.288	1.851	939
Clicks	43.209	18.042	7.218	3.347
Webpages	450	428	86	104
unterschiedliche Wörter	6.063	5.271	1.752	1.982
alle Links	13.778	8.651	2.534	3.143

Tabelle 5.3: Rohdaten der analysierten Websites

und Strukturdaten bereinigt, da sich danach entscheidet, welche Sessions verwendet werden können. Das Ergebnis dieser Datenaufbereitungsschritte wird in Tabelle 5.4 gezeigt.

	Website A		Website B	
	Rohdaten	bereinigt	Rohdaten	bereinigt
Sessions	5.288	4.526	939	853
Clicks	18.042	15.197	3.347	3.050
Webpages	428	335	86	78
unterschiedliche Wörter	5.271	2.383	1.982	830
Links	8.651	8.651	3.143	674

Tabelle 5.4: Bereinigung der Daten: Website A und B, Zeitraum Januar 2006

Transformation der Webpage IDs

Ausgehend von der Homepage, hat der Crawler durch Verfolgung der Linkstruktur auf Website A 428 Webpages gefunden und deren Texte erfaßt. Gleicht man die gesammelten Webpages mit den von den Usern besuchten Webpages ab, sind 335 Webpages sowohl in den Content- wie auch in den Usage-Daten vorhanden. Die übrigen Webpages können nicht berücksichtigt

werden. Da sie überhaupt nicht aufgesucht wurden, tragen sie weder positiv noch negativ zum Erfolg der Website bei.

Um teilweise nicht bewertbare Clickstreams zu vermeiden, wurden nur diejenigen Sessions verwendet, deren besuchte Webpages alle durch den Crawler erfaßt werden konnten. Die Sessionzahl hat sich dadurch um 10-15% reduziert. Da die Sessions in Zeiträumen von mehreren Tagen anfallen, aber nur zu jeweils einem Zeitpunkt in diesem Zeitraum die Website durch den Crawler erfaßt wurde, kann es vorkommen, daß Webpages für den Crawler nicht mehr erreichbar waren. In einem produktiven System müßte man daher die Zeitspanne zwischen Crawlvorgängen möglichst kurz halten, ohne dabei die Datenmengen zu stark zu erhöhen.

Wie von Wilde et al. in [122, S.37] beschrieben, hat auch bei dieser Arbeit die Datensammlung und -aufbereitung den größten Teil der Gesamtzeit in Anspruch genommen. Vor allem die Datenintegration von Usage- und Content-Daten gestaltete sich sehr aufwendig, da die Identifikatoren der Webpages im Content Management und im Web Reporting & Mining System nicht mit den Identifikationsmerkmalen der Webpages durch den Crawler vollständig kompatibel waren. So konnten gleiche Webpages zunächst nicht als solche erkannt werden. Da das Content Management System (CMS) die URL dynamisch zusammensetzt und mehrere Webpage- und Session-Parameter wie beispielsweise die SessionID mitaufnimmt, kann eine Webpage nicht direkt durch deren URL identifiziert werden.

Zwar erhält jede Webpage durch das CMS eine ContentID, doch ist diese nicht immer eindeutig. Artefakte wie Sprachklone einer einzelnen Webpage sollten durch die Datenbank des CMS auf eine konsistente ContentID dieser Webpage aufgelöst werden können. In diesen Fällen mußten die ContentIDs manuell korrigiert werden.

Der Crawler auf der anderen Seite orientiert sich an der URL und den darin enthaltenen Informationen. Zwar ist in nahezu jeder URL der untersuchten Webpages eine ContentID enthalten, allerdings müssen diese IDs ähnlich wie gerade beschrieben auf eine einheitliche, konsistente Basis überführt werden. Mit dieser Basis lassen sich die Usage-Daten aus dem CMS mit den Content-Daten des Crawlers zusammenführen. Ist dies bei Webpages nicht möglich, werden diese von der Analyse ausgeschlossen. Somit ist eine nahezu vollständig automatisierte, konsistente Datentransformation der ContentIDs möglich. Bei der Menge an Webpages wäre ein manuelles Vorgehen nicht durchführbar. Die aus Platzgründen nicht aufgeführte Website C umfaßt 2.209 Webpages. Die Zuweisung der ContentIDs durch das CMS wurde jedoch von manchen Website-Autoren zur Gestaltung der Website in anderem Sinne verwandt. Dadurch wurde die ID-Zuordnung inkonsistent und nicht alle Webpages konnten eindeutig den richtigen IDs zugeordnet werden.

Die Einschätzung der Datensammlung und -aufbereitung als zeitintensivste Phase durch Wilde et al. [122, S.37] wird durch diese Analysen erneut bestätigt.

Bereinigung und Transformation der Keywords

Die gesammelten Texte werden in einzelne Wörter zerlegt. Die Wörter werden ihren Wortstämmen mit Hilfe des Stemming-Algorithmus von Porter [195] zugeordnet. Wird im folgenden Text von Wörtern gesprochen, sind damit die Wortstämme der bereinigten Daten gemeint.

Die Anzahl der unterschiedlichen Wörter (Keywords) führt zu einer Dimensionalität von 335 5271 (Webpages Wörter). Diese $Webpages$ $Wörter$ Matrix ist nur sehr dünn besetzt. Man spricht nach Baldi et al. [12] von einer *Sparse Matrix*. Zusätzlich zu den in 2.3.2 und 4.3.1 beschriebenen Methoden zur Dimensionsreduktion von Worträumen wird hier vorgeschlagen, die am wenigsten besetzten Spalten ($Wörter$) auszuschließen

Je größer die Anzahl gemeinsamer Wörter zwischen den Webpages ist, desto wahrscheinlicher ist es, inhaltliche Zusammenhänge zu entdecken. Daher sollen hier Wörter, die auf weniger als 3 Webpages vorkommen, aus der Webpage-Wort-Matrix entfernt werden. Dies wird damit begründet, daß 2 Webpages bei einer Gesamtzahl von 335 Webpages kein eigenes Thema bilden können.

Zusätzlich werden alle Wörter entfernt, die auf nahezu allen Webpages vorkommen, wie zum Beispiel der Unternehmensname bei Unternehmens-Websites. Kommt ein Wort auf mehr als 75 % der Webpages vor, wird es ausgeschlossen. Auf diese Weise reduziert sich bei den Websites A und B die Dimensionalität der Wörter auf weniger als die Hälfte.

Ergebnis der Datenaufbereitung

Die bereinigten und transformierten Contentdaten liegen als Matrix der Form $Webpage$ $Wort$ vor. Die Usagedaten liegen als Clicks in einer Tabelle sortiert nach Sessions in chronologischer Reihenfolge der Clicks vor. Nach dem Datenaufbereitungsschritt erfolgt wie in Abb. 5.1 gezeigt die Analyse der Daten.

5.2.4 Analyse

Webpage-Kategorisierung

Die Identifikation von Navigationsseiten erfolgt manuell. Dabei werden Homepage, Suchseiten (simple, advanced) und Sitemap als Navigationsseiten kategorisiert, die hauptsächlich der Navigation dienen. Diese Webpages werden von der inhaltlichen Analyse zur Identifikation von Themen ausgeschlossen. Die restlichen Webpages stellen Content zur Verfügung. Auf ihnen werden im folgenden die Themen der Website erkannt.

Themenidentifikation

Bevor mit der Berechnung des GPI begonnen werden kann, müssen die Themen der Website identifiziert werden. Dazu kann beispielsweise der in Abschnitt 4.5 vorgestellte Ansatz benutzt werden, der die von Usern wahrgenommenen Themen approximiert. Allerdings beeinflussen die Struktur und der Inhalt der Website auch die Benutzung der Website. Außerdem ist der in 4.5 vorgestellte Ansatz auf sehr vielen Annahmen begründet, sodaß ein anderer Ansatz vorzuziehen ist.

In 4.3.4 wurde der PLSA vorgestellt, der speziell für die Content-Analyse entwickelt wurde. Er ist in der Lage, die auf einer Website bestehenden Themen für weitere Analysen verfügbar zu machen.

Wie bei den meisten Clusteralgorithmen wird die Clusteranzahl vorgegeben. Die Clusteranzahl kann zum Beispiel durch Analyse der Eigenvektoren der Webpage-Wort-Matrix mittels einer Singulärwertzerlegung oder einer Hauptkomponentenanalyse anhand des Kriteriums von Kaiser bestimmt werden, das bei Handl [114, S.121] beschrieben wird. Ein Beispiel für die Anwendung der Singulärwertzerlegung ist in Kapitel 4.3.3 zu finden.

Man erhält die Zugehörigkeit von Webpages und Worten zu den identifizierten Themen einer Website. Die Themenzugehörigkeit wird durch einen Vektor beschrieben, der die prozentuale Zugehörigkeit eines Wortes bzw. einer Webpage zu allen Themen beschreibt. Das bedeutet, der Themenvektor beschreibt das Mischungsverhältnis der Themen für ein gegebenes Wort oder eine Webpage. Für die Beispielsession in Tab. 5.5 erhält man für Webpage 1171 bei 5 Themen den Themenvektor: {0,02; 0,47; 0,34; 0,12; 0,05}.

Nachdem die Themen der Website durch den PLSA festgelegt wurden, kann mit der Berechnung der Effektivitätsmaße fortgefahren werden. Die GPI-Berechnung wird beispielhaft an den Clicks einer Session demonstriert. Es wird Session 1 von Website A und Session 2 von Website B jeweils im Zeitraum November bis Dezember 2005 ausgewählt. Der GPI besteht, wie in Def. 5.1.5 beschrieben, aus zwei Größen: Effektivitätsmaß, das aus Transitionstyp und Transitionsgewicht gebildet wird, und dem Effizienzmaß.

Effektivitätsmaß: Transitionstyp

Der Transitionstyp richtet sich nach dem Übergang zwischen Webpage-Kategorien. Der Transitionstyp unterscheidet grundlegend zwischen gewünschten Clicks, denen ein positiver Wert zugewiesen wird, und unerwünschten Clicks, die negativ bewertet werden. Die Bewertung dieser Übergänge erfolgt gemäß Tab. 5.1.

In Session 1 bewegt sich der User fast ausschließlich zwischen Contentseiten. Lediglich der Einstieg erfolgt auf einer Navigationsseite, der Homepage. Die Homepage hat den User direkt zu einer Contentseite geleitet und somit ihre Aufgabe erfüllt. Daher wird die Homepage mit

Web-page	Seq Nr.	Dura-tion	avg. Duration	Seitentyp	Duration-gewicht	Transitions-gewicht	Transitionstyp	GPI
Session 1								
Home	1	3	23,23	nav	1,5	1,5	1	2,25
2578	2	96	21,33	cont	0,8	1,137	1	0,91
1067	3	3	21,97	cont	0,4	0,669	1	0,268
1067	4	13	17,48	cont	1,2	1,044	1	1,253
1042	5	6	23,25	cont	0,4	0,6158	1	0,246
1302	6	36	21,11	cont	1,2	1,1279	1	1,535
2579	7	10	27,64	cont	1,2	1,092	1	1,310
1053	8	-	26,31	cont	1	0,6069	0,1	0,061
Session 2								
Home	1	14	22,73	nav	0,8	1,5	-1	-1,2
Search	2	14	22,86	nav	0,8	1,5	-1	-1,2
Search	3	7	22,86	nav	0,8	1,5	-1	-1,2
Search	4	22	14	nav	0,5	1,5	1	0,75
1171	5	-	21.18	cont	1	1	0,1	0,1

Tabelle 5.5: Session mit GPI bewertet

$+1$ positiv bewertet. Die übrigen Clicks erfolgen von Content- zu Contentseiten. Diese Clicks werden als wünschenswert betrachtet und grundsätzlich im positiven Bereich mit $+1$ bewertet. Die Höhe der Bewertung erfolgt anschließend durch eine inhaltliche Analyse, mit der das Transitionsgewicht bestimmt wird. In Session 2 bewegt sich der User zuerst zwischen Navigationsseiten, was negativ zu bewerten ist, da hier keine Informationen vermittelt werden. Damit wird das Ziel von informationsorientierten Websites nicht erreicht. Der Transitionstyp dieser Clicks wird mit -1 bewertet. Erst die Webpage an Position 4 der Session erfüllt ihre Aufgabe, indem sie den User auf eine Contentseite (1171) leitet. Die Webpage hat ihre Aufgabe erfüllt und wird positiv bewertet. Click Nr. 5 liegt am Ende einer Session. Die letzte Webpage einer Session kann, wie in Kap. 3.3.1 beschrieben, nicht eindeutig bewertet werden. Zumindest wird es positiv gesehen, daß die Session auf einer Content- und nicht auf einer Navigationsseite endet. Daher fließt der Transitionstyp dieses Clicks lediglich mit einem geringen aber trotzdem positiven Wert 0,1 ein.

Effektivitätsmaß: Transitionsgewicht

Das Transitionsgewicht bestimmt das Gewicht, mit dem ein positiver oder negativer Transitionstyp in den GPI einfließt. Diese Gewichtung wird im Fall von Contentseiten durch den inhaltlichen Vergleich der besuchten Webpage mit dem Inhalt der gesamten Session vergli-

chen. Siehe hierzu Abb. 5.3. Der Themenvektor der betrachteten Webpage wird mit dem durchschnittlichen Themenvektor der gesamten Session verglichen. Überschreitet die Differenz beider Vektoren einen festgelegten Schwellenwert, wird diese Webpage als inhaltlicher Ausreißer der Session gewertet. Damit sollen Webpages, die innerhalb des inhaltlichen Fokuses des Users liegen, höher gewichtet werden als solche, die einen Inhalt anbieten, der weit vom inhaltlichen User-Fokus abweicht. Der User-Fokus wird als Mittelpunkt aller Themenvektoren der Session durch deren Durchschnitt gebildet.

Es kann entweder ein fester Wert für eine hohe oder niedrige Abweichung festgelegt werden, oder wie im Beispiel in Tab. 5.5 wird die Differenz des Themenvektors einer Webpages zum inhaltlichen Fokus der Session (durchschnittlichen Themenvektor) gebildet. Da eine größere Differenz ein geringeres Gewicht erhalten soll und eine kleine Differenz ein hohes Gewicht, wird das Transitionsgewicht umgekehrt proportional zur Differenz aus User-Fokus und aktuellem Themenvektor festgelegt. Diese umgekehrte Proportionalität wird dadurch erreicht, daß die Differenz aus User-Fokus und aktuellem Themenvektor von 1 abgezogen bzw. dazugezählt wird. Man zieht es ab, wenn die berechnete Themendifferenz größer als der festgelegte Schwellenwert ist, und addiert es zu 1 hinzu, wenn die Themendifferenz kleiner ist. Somit erhält man Transitionsgewichte zwischen 0 und 1 für Themendifferenzen außerhalb des User-Fokuses, und Transitionsgewichte zwischen 1 und 2 für Themendifferenzen innerhalb des User-Fokuses.

Die Festlegung der Gewichte richtet sich zum einen nach der Konstruktion des Gesamtmaßes als Produkt seiner Teilmaße, zum anderen durch die Festlegung der Ziele einer Website und deren Bewertung und Gewichtung durch ein Erfolgsmaß. In 5.4 wird die manuelle Festlegung von Gewichten bei der Berechnung des GPI durch eine generische Lösung ersetzt.

Effizienzmaß: Durationgewicht

Das Effektivitätsmaß hat bereits die Zielorientierung eines Clicks bewertet. Das Effizienzmaß in Form des Durationgewichts soll den Grad der Zielerreichung einbeziehen. Im Fall von informationsorientierten Websites soll bewertet werden, inwieweit der User die Möglichkeit hatte, den Inhalt der Webpage ausreichend lange wahrzunehmen.

Bedingt durch die Anlage des GPI als Produkt aus Transitionstyp, Transitionsgewicht und Durationgewicht, soll das Durationgewicht die zuvor ermittelten Effektivitätsmaße verstärken oder dämpfen.

Die Festlegung des Schwellenwertes richtet sich nach der gewünschten Zeit, die ein User auf einer Webpage verbringen soll. Je nach Webpage Typ soll sich ein User lang oder kurz auf der Webpage aufhalten - auf einer Contentseite lange, auf einer Navigationsseite kurz (siehe Tab. 5.2). Einen objektiven Wert für die optimale Dauer zur Wahrnehmung der Inhalte einer Webpage zu ermitteln, fällt in den Bereich der Psychologie und umfassender User-Studien.

Aufgrund der hohen Anzahl an unterschiedlichen User-Typen auf einer Website wurde die durchschnittliche Duration aller User auf einer Webpage als Schwellenwert ausgewählt.

Es bleibt die in Kap. 3.3.1 beschriebene Einschränkung der Interpretierbarkeit der Maßzahl Duration bestehen, sodaß aus der Duration nicht auf den Zeitraum der Aufmerksamkeit, die der User der Webpage zugewandt hat, geschlossen werden kann. Die Bewertungen der Duration ist aus Sicht des Website-Autors die Zeitspanne, die der User hatte, um den Inhalt wahrnehmen zu können. Auf Contentseiten ist eine zu kurze Duration unerwünscht, da der User so die Inhalte nicht ausreichend wahrnehmen kann. Allerdings können User, die die Website inhaltlich sehr gut kennen, durch einzelne Webpages relativ schnell navigieren. Daher kann hier nur ein durchschnittlich gewünschtes Verhalten bewertet werden.

Ein zu langer Aufenthalt ist wiederum auch nicht wünschenswert, da entweder die Aufmerksamkeit des Users abgelenkt wurde oder die grundlegende Annahme eines sequentiellen, nicht-parallelen Browsingverhaltens nicht mehr gilt, wie in Kapitel 2.2.3 beschrieben. Zu diesem Fallbeispiel wurden zur Konkretisierung von Tabelle 5.2 folgende Gewichte in Tabelle 5.6 festgelegt.

	kurz	mittel	lang
Contentseiten	0,4	1.2	0.8
Navigation → Content	1.5	0.8	0.5
Navigation → Navigation	0.4	0.8	1.5
Webpage → Sessionende	1		

Tabelle 5.6: Durationgewichte in der Fallstudie

Um die durchschnittliche Duration wurden drei Intervalle gebildet: kurz, mittel und lang. Das mittlere Intervall wurde als größtes angelegt, in das die meisten Seitenzugriffe fallen. Es wird das 20% Quantil als untere Grenze für einen kurzen Aufenthalt auf einer Website und das 80% Quantil als obere Grenze vorgeschlagen. Dadurch werden Ausreißer erkannt, und ihnen wird durch ein geringeres Gewicht ein geringerer Einfluß auf die Bewertung der Website gegeben.

Der Aufenthalt auf einer Contentseite wird durch die Effektivitätsmaße grundsätzlich positiv bewertet. Ein kurzer Aufenthalt dämpft diese positive Bewertung mit einem Faktor von *0,4* stärker als ein zu langer Aufenthalt dies mit *0,8* macht. Hält der User sich im mittleren Durationbereich auf, wird die positive Bewertung mit *1,2* noch verstärkt.

Bei Navigationsseiten wird die Durationbewertung weiter differenziert, indem unterschieden wird, wohin der User nach der betrachteten Navigationsseite gegangen ist. Ein kurzer Aufenthalt auf einer Navigationsseite soll positiver bewertet werden als ein langer Aufenthalt, da die Gefahr des Sessionendes, ohne Informationen vermittelt zu haben, besonders groß ist. Man muss hier nun das Vorzeichen aus dem Transitionstyp für den jeweiligen Click berücksichtigen. Soll das Durationgewicht eine positiv zu bewertende Duration belohnen, muß bei einem

positiven Vorzeichen ein Wert > 1 gewählt werden, bei einem negativen Vorzeichen des Transitionstyps muß ein Durationgewicht < 1 und > 0 gewählt werden. Damit wird ein langer Aufenthalt auf einer Navigationsseite, auf die eine weitere Navigationsseite folgt, durch eine *1,5* stärker und bei einem kurzen Aufenthalt mit *0,4* schwächer bestraft. Beim Navigationsmuster $Navigation \rightarrow Content$ ist es genau umgekehrt.

Der letzte Click in einer Session muss auch getrennt bewertet werden. Da die Duration hier nicht gemessen werden kann, wird ein neutrales Durationgewicht von *1* gewählt und die Bewertung den Effektivitätsmaßen überlassen.

Ein anderer Ansatz zur Bestimmung eines Schwellenwertes wurde in [242] gewählt. Hier wurde die Anzahl an Worten und die durchschnittliche Lesedauer pro Wort über alle User Sessions und Webpages als Basis genommen. Für jede Webpage wurde damit eine zu erwartende mittlere Lesezeit in Abhängigkeit von der Länge des Textes berechnet. Beide Berechnungsformen des Schwellenwertes sind nachvollziehbar und liefern sehr ähnliche Ergebnisse. Für jede Website sollte eine an Struktur und Darstellung angepaßte Form der Durationgewichtsberechnung gewählt werden.

GPI-Berechnung

Man erhält den GPI-Wert für einen Click, indem man Effektivitätsmaß, d.h. Transitionstyp *effectivitytype* und Transitionsgewicht *effectivityweight* , mit dem Effizienzmaß, dem Durationgewicht *efficiency* kombiniert. Wie Tab. 5.5 beispielsweise für Session 1 Click 2, Transitionstyp *effectivitytype* , Transitionsgewicht *effectivityweight* , Durationgewicht *efficiency* sodaß sich der $GPI_{1.2} = \textit{effectivitytype}_{1.2} \quad \textit{effectivityweight}_{1.2} \quad \textit{efficiency}_{1.2} = 1 \quad 1,137 \quad 0.8 = 0,91$ ergibt.

5.2.5 Interpretation und Anwendung der Ergebnisse

Der GPI wird für jeden Click einer Session berechnet. Daraus lassen sich GPI-Werte für jede Session und jede Webpage aggregieren und entsprechende Mittelwerte als durchschnittliche Bewertung der Session bzw. Webpage berechnen.

GPI-Aggregation je Webpage

Tab. 5.7 zeigt den aggregierten GPI für drei Webpages aus Website A. Um den Effekt von unterschiedlichen Besucherzahlen je Webpage zu entfernen, bildet man den durchschnittlichen GPI je Click für diese drei Webpages. Man sieht in allen Zahlen sehr deutliche Unterschiede. Sowohl die *Simple Search* als auch die *Advanced Search* sind Navigationsseiten und

unterscheiden sich dadurch, daß die Advanced Search mehr Auswahlmöglichkeiten zur Verfeinerung der Suche anbietet als die Simple Search. Zum Vergleich dient die Bewertung einer Contentseite, die über das Produkt A informiert.

	GPI	Anzahl Clicks	GPI je Click
Simple Search	-350,5	443	-0.79
Advanced Search	14,9	75	0.20
Produkt A	775,5	1193	0,65

Tabelle 5.7: GPI aggregiert je Webpage

Vergleicht man die beiden Suchseiten, erhält die Advanced Search sowohl bei den absoluten als auch den durchschnittlichen GPI-Werten eine höhere Bewertung als die Simple Search. Navigationsseiten werden vom GPI dann positiv bewertet, wenn sie den User schnell zum geeigneten Content führen. Muß der User erst mehrere Navigationsseiten durchqueren, bevor er zu Contentseiten gelangt, wird dies negativ bewertet. Denn hier ist die Gefahr des Sessionabbruchs besonders groß, wenn man bedenkt, daß die durchschnittliche Sessionlänge auf dieser Website gerade 2,4 Clicks beträgt.

Interpretiert man die unterschiedlichen GPI-Werte im Hinblick auf die beschriebene Konstruktion des GPI, gelangt man zu dem Schluß, daß die Advanced Search den User schneller zu Contentseiten leitet als die Simple Search. Die Schlußfolgerung für den Website-Autor ist daher, die Simple Search zu verbessern oder durch die Advanced Search zu ersetzen.

Die Werte des GPI lassen sich lediglich innerhalb einer Website im gleichen Analysezeitraum miteinander vergleichen, es sei denn Inhalt und Struktur einer Website bleiben über einen längeren Zeitraum unverändert. Dann muß allerdings die unterschiedliche Anzahl an Clicks je Webpage im jeweiligen Zeitraum beispielsweise dadurch berücksichtigt werden, daß man den durchschnittlichen GPI je Click für die Webpages berechnet.

Durch den GPI erhält man einen aggregierten Überblick über alle Sessions. Durch die Aggregation verschiedener Werte mit Hinblick auf die Ziele von informationsorientierten Websites erreicht man ein aussagekräftiges Maß für den Erfolg einzelner Webpages und kann problematische Webpages identifizieren. Im Gegensatz zu einer einfachen Clickstream-Analyse, berücksichtigt der GPI neben den Clicksequenzen einer Session auch Duration ebenso wie den Inhalt und Webpage-Typ. Eine mühsame Interpretation mehrerer Clickstreams durch Interessantheitsmaße entfällt.

GPI-Aggregation je Session

Neben der Aggregation je Webpage kann man ebenso auf Basis von Sessions aggregieren. Dadurch lassen sich User-Besuche als Ganzes bewerten. Möchte man das obige Beispiel in Tab. 5.7 detaillierter auf Sessionebene analysieren, wählt man beispielsweise zwei Sessions, die auf die jeweiligen Suchseiten zugegriffen haben. Die Ergebnisse in Tab. 5.8 bestätigen die obige Analyse, daß die Advanced Search den User schneller zum Ziel führt als die Simple Search, da man direkt zum Content geleitet wird.

	GPI_1	GPI_2	GPI_3	GPI_4	$\sum$	je Click
Home→Adv.Search→Content	-1,1	1,2	1,35	...	1,45	0,48
Home → Search → Search → Content	-1,1	-1,2	0,8	1,35	-0,15	-0,04

Tabelle 5.8: GPI aggregiert je Session

Da die erste Session lediglich 3 Clicks benötigt, wäre ein aufsummierter GPI mit dem einer Session von 4 Clicks nicht vergleichbar. Der GPI je Click zeigt, daß die zweite Session schlechter bewertet wird als die erste, da hier die Gefahr eines Abbruchs der Session auf den Simple Search Seiten groß ist. Das Ziel der schnellen und intuitiven Informationsvermittlung wurde bei Session 2 weniger gut erreicht als bei Session 1. Dies spiegelt der GPI korrekt wider.

5.2.6 Schlußfolgerung

Diese Fallstudie dient zur Illustration des Analyse- und Berechnungsprozesses des GPI, sowie zu dessen Anwendungsmöglichkeiten und Vorteilen gegenüber anderen Web Mining-Verfahren bei der Erfolgsmessung. Zur Evaluierung des GPI reicht dies nur zum Teil.

Um dem großen Umfang an zugrundeliegenden Daten noch eine weitere Evaluierungsmöglichkeit hinzuzufügen, wurde eine User-Studie durchgeführt, die die getroffenen Annahmen im Bewertungsmodell und die Konstruktion des GPI validieren soll. Eine User-Studie wurde deshalb gewählt, weil dieses Instrument der expliziten Befragung von Usern bei der Konstruktion des Erfolgsmaßes nicht verwendet wurde. Daher stellt es ein unabhängiges Mittel dar, das Maß unbeeinflußt zu bewerten.

5.3 Evaluierung durch User-Studie

Neben der empirischen Untermauerung der oben beschrieben Ergebnisse soll die Befragung von Usern die bisher unberücksichtigte User-Seite untersuchen und miteinbeziehen. In dem oben entwickelten formalen Modell einer Website in Abb. 5.5 sind auf der User-Seite nur die beobachtbaren Fakten aus der Session berücksichtigt. Details über die Ziele und Absichten eines Users sind auf der Server-Seite unbekannt. Die User-Studie soll diesen Mangel an Daten beheben, um dadurch eine zusätzliche Überprüfung des GPI zu ermöglichen.

Auf der Server Seite sind die verfügbaren Daten auf die beobachtbaren Clickstreams, den Inhalt und die Struktur der Website beschränkt. Auf informationsorientierten Websites bleibt der User meist anonym und gibt seine Ziele nicht preis. Bei transaktionsorientierten Websites gibt er mit einer Transaktion Einblick in seine Interessengebiete, unabhängig davon, ob die Transaktion komplett abgeschlossen wurde. Durch eine direkte Befragung der User kann man dies auch bei informationsorientierten Websites erreichen. Allerdings unter sehr viel aufwendigeren Umständen und nicht kontinuierlich.

5.3.1 Konzeption der User-Studie

Unter Laborbedingungen sollen die befragten User mehrere Aufgaben auf der Website erfüllen bzw. bestimmte Ziele erreichen. Entsprechend dem Modell von Pather et al. in [188], erläutert in Abb. 3.5 in Kap. 3.3.3, werden zunächst die Erwartungen der User und ihre Vorkenntnisse über die Website und das Unternehmen erfragt. Anschließend führt der User die gestellten Aufgaben durch. Jede Aufgabe wird dabei als vorgegebenes Ziel eines Users gesehen, sodaß die Sessions mit der gleichen Aufgabe vergleichbar sind. Die Beobachtung des Users erfolgt durch die SessionID, mit deren Hilfe durch das WRM-System alle Clicks aufgezeichnet werden. Damit erhält man einen Clickstream, bei dem das Ziel des Users bekannt ist und der mit dem GPI ausgewertet werden kann. Eine zukünftige und umfassendere Untersuchung der User-Wahrnehmung durch eine derartige Studie sollte einen größeren Umfang als 30 Beobachtungen aufweisen.

Wie bei dem Fragebogenkonzept von Barnard et al. [13], wurden jedem User drei Ziele und Fragen gestellt, die er auf Website A und B erreichen und beantworten sollte:

- “Welche Produkte für Privatkunden bietet das betrachtete Unternehmen an ?“
- “Wie heißt der Vorstand der Firma B ?“
- “In welcher Stadt betreibt Firma B *Technoparks* ?“

Die Fragen konnten auf jeweils einer oder zwei Webpages einer Website beantwortet werden. Es wurde darauf geachtet, daß die User die Website während der Session nicht verlassen. Diese Laborbedingungen sind für die Vergleichbarkeit der Sessions notwendig, entsprechen

aber nur eingeschränkt der Realität. Da jedoch die beobachteten Clickstreams auch nur auf der Website beobachtet werden können, weichen die Clickstreams aus der User-Befragung davon nicht ab.

Es wurden zehn User befragt und jeweils drei Aufgaben gestellt. Somit stehen 30 auswertbare Sessions zur Verfügung. Der User konnte die Session jederzeit beenden, sobald er den Eindruck hatte, die Aufgabe gelöst zu haben. Er konnte die Session auch abbrechen, ohne sein Ziel erreicht zu haben.

Auswertungskonzept

Der Fragebogen erhebt die User-Bewertung auf zwei Arten: 1. Vergleich von Erwartung und Wahrnehmung und 2. die Benotung von Website-Eigenschaften.

Vergleich von vorheriger Erwartung mit tatsächlicher Wahrnehmung Wie Pather et al. in [188] vorschlagen, werden die Erwartungen mit den Wahrnehmungen der Website durch die User verglichen. Es wurden die Erwartungen der User hinsichtlich der Schnelligkeit und Bequemlichkeit erfragt, in der die gesuchte Information gefunden werden konnte, und nach der Qualität der Informationen. Die gleichen Fragen wurden nach Durchführung der Aufgabe gestellt und damit die User-Wahrnehmung erfragt. Liegt die tatsächliche Clickzahl

Erwartung	**Wahrnehmung gemessen durch**
Information zu finden	Zielseite erreicht
minimale / maximale Anzahl Clicks bis Ziel	Clicks pro Session
Dauer bis Ziel	Duration der Session

Tabelle 5.9: Erwartungswerte vs. Messung der User-Wahrnehmung

einer Session innerhalb der minimal und maximal erwarteten Anzahl an benötigten Clicks, hat die Website die Erwartungen genau erfüllt. Es wird ein positiver Wert zugewiesen (+1). Bei Übertreffen der Erwartungen, d.h. es wurden weniger Clicks benötigt als erwartet, wird ein höherer positiver Wert (+2) zugewiesen. Benötigt ein User mehr Clicks als erwartet, konnte die Website die Erwartungen nicht erfüllen. Dies wird negativ bewertet (-1), sofern eine hohe Abweichung vorliegt mit (-2). Die Bewertung der erwarteten und tatsächlichen Dauer (Duration) der kompletten Session erfolgt parallel, d.h. wenn weniger Zeit benötigt wird als erwartet, wird dies positiv bewertet. Wird mehr Zeit benötigt als erwartet, wird dies negativ bewertet.

Qualitative Einschätzung Die qualitativen Einschätzungen über Erreichbarkeit und Navigierbarkeit wurden vom User erfragt. Es wurde eine Notenskala von 1 bis 5 verwendet. Zu folgenden Punkten wurden von jedem User eine ex ante Bewertung zu seinen Erwartungen sowie eine ex post Bewertung über seine Wahrnehmungen erfragt:

- Gestaltung der Website
- intuitive Zielführung
- Navigierbarkeit
- Strukturiertheit der Website
- Qualität der Informationen

Die Abweichungen zwischen Erwartung und Wahrnehmung wurden wie oben beschrieben auf den Wertebereich von -2 bis +2 normiert.

Gesamtbewertung Es wird der gruppenweise Durchschnitt aus dem *Erwartungs - Wahrnehmungs - Vergleich* und der *Qualitativen Einschätzung* zu einer Gesamtbewertung durch Bildung des Mittelwerts zusammengeführt. Die Bewertung aus der User-Studie wird mit dem GPI verglichen. Dabei wird überprüft, ob die Bewertung einer Session durch den GPI mit der Bewertung durch den User tendenziell übereinstimmt.

Die Übereinstimmung soll durch ein Korrelationsmaß überprüft werden.

Korrelationsmaß

Es soll der Zusammenhang zwischen der GPI-Bewertung und der Bewertung durch die User-Studie gemessen werden. Es wird untersucht, ob die drei Sessions eines befragten Users zueinander jeweils im gleichen Verhältnis stehen. Wird durch User A Session 1 höher eingeschätzt als Session 2, und bewertet der GPI beide Sessions im gleichen Verhältnis, wird die GPI-Bewertung als durch den User bestätigt angesehen.

Hierzu können beispielsweise der empirische Korrelationskoeffizient oder Spearmans Korrelationskoeffizient verwendet werden.

Empirischer Korrelationskoeffizient Der empirische Korrelationskoeffizient r, auch Bravais - Pearson - Korrelationskoeffizient, wird für die Daten $(x_i, y_i), i = 1, ..., n$ wie folgt berechnet [93, S.139ff]:

$$r = \frac{\sum_{i=1}^{n}(x_i \quad x)(y \quad y)}{\sqrt{\sum_{i=1}^{n}(x_i \quad x)^2(y_i \quad y)^2}} \tag{5.6}$$

Spearmans Korrelationskoeffizient Einen alternativen Korrelationskoeffizienten erhält man, wenn man von den ursprünglichen x- und y-Werten zu deren Rängen übergeht [93, S.141f]. Dieser Spearmans Korrelationskoeffizient bewertet dabei die Reihenfolge der Sessions, aber nicht die Gleichheit bezüglich der absoluten Werte, sondern nur die Bewertung der Sessions zueinander. Eine positive Session muß daher nicht durch GPI und User einen positiven Wert erhalten. Es reicht aus, wenn die Reihenfolge der Sessionbewertungen aus User- und GPI-Sicht übereinstimmen.

Der Spearmans Korrelationskoeffizient berechnet sich wie der Bravais - Pearson - Korrelationskoeffizient. Es werden lediglich die Ränge der Werte statt der Werte selbst eingesetzt. Der Vorteil des Spearmans. K. liegt darin, daß die Qualität der Skalen eines oder mehrerer Merkmale nicht abstandstreu sein muß [208, S.96]. Ordinal skalierte Merkmale können somit bei Spearmans auf Korrelation untersucht werden. Da die Bewertung durch die User als auch durch den GPI ordinal-skalierte Werte hervorbringt, soll der Spearmans Korrelationskoeffizient verwendet werden.

5.3.2 Ergebnisse der User-Studie

Gesamturteil der User

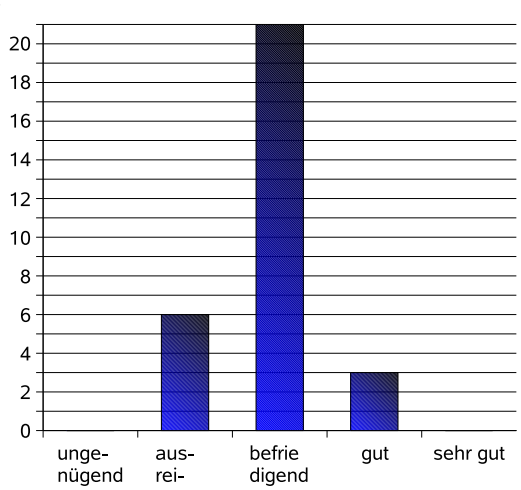


Abbildung 5.8: Gesamtbewertung der Websites

Die befragten User haben beide evaluierten Websites A und B mittelmäßig bewertet. Lediglich 10% haben die Websites nach der Befragung als gut empfunden. Die besten und schlechtesten Noten wurden nicht vergeben und die Websites wurden eher zurückhaltend bewertet.

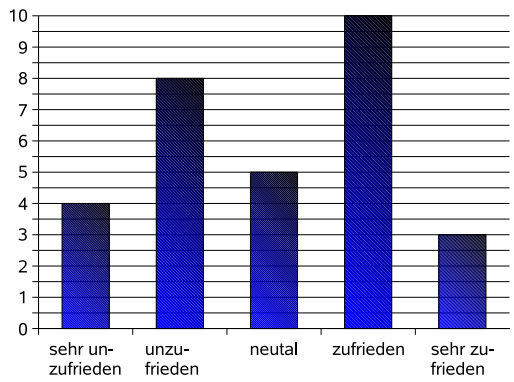


Abbildung 5.9: User-Bewertung der einzelnen Sessions

Differenzierter fällt das Urteil bei einzelnen Sessions aus. In einigen Sessions haben sich die User auf der Website überhaupt nicht zurecht gefunden und ihr Ziel nicht erreicht, was zu einer Bewertung *"sehr unzufrieden"* geführt hat.

Zielerreichung

Neben der qualitativen Bewertung durch die befragten User wird das Erreichen der gestellten Aufgaben kontrolliert. Da zwei Websites in die Studie einbezogen wurden, werden die Ergebnisse in Abb. 5.10 separat dargestellt. Die Ergebnisse entsprechen von ihrer Aussage her der Kennzahl *Task Completion Rate*. Auf Website A wurden zwei Aufgaben gestellt, somit fanden

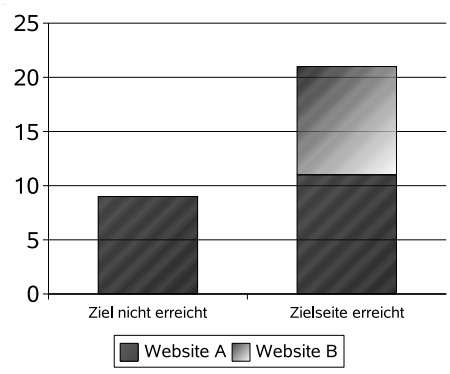


Abbildung 5.10: Zielerreichung nach Website

zwanzig Sessions auf Website A statt, zehn auf Website B. Bei Website B wurde das Ziel in

allen zehn Sessions erreicht. Bei Website A lediglich in elf von zwanzig Fällen, in neun dagegen nicht. Der GPI kann über die Zielerreichung im Sinne einer Task Completion Rate aus 3.3.3 keine Auskunft geben, da auf der Server-Seite die Ziele der User und damit mögliche Zielseiten unbekannt sind.

Anzahl Clicks- Erwartung und Realität

Zur Beurteilung der Servicequalität wird der in Kap. 3.3.3 beschriebene Vergleich zwischen Erwartung und Wahrnehmung angewandt. In dieser Befragung wird der Vergleich für die Schnelligkeit durchgeführt, mit der ein User die Website nach der Zielseite durchnavigiert, gemessen nach Clicks und Duration. Lediglich zwei der Sessions benötigten weniger Clicks

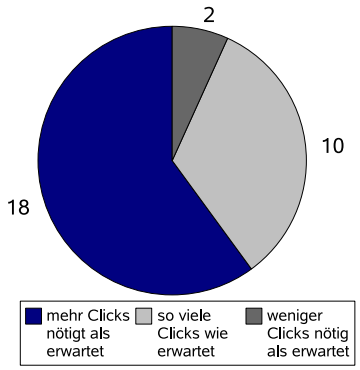


Abbildung 5.11: Erwartete vs. benötigte Clickanzahl

als erwartet. Die überwiegende Mehrzahl von 18 Sessions besuchten mehr Webpages als die User erwartet hatten. Ein Teil davon läßt sich durch die Sessions erklären, die ihr Ziel nicht erreicht haben, denn ein User wird bei einem festen Ziel länger versuchen, dieses zu erreichen. Diese User sind bereit, mehr Webpages aufzusuchen, als sie zunächst beabsichtigt hatten.

Duration - Erwartung und Realität

Neben der Anzahl an Clicks zeigt die Dauer einer Session, die Duration, wie schnell ein User durch die Website navigiert ist. Dieses Geschwindigkeitsmaß liefert ein anderes Bild als die Clicks. In einem wesentlich größeren Anteil an Sessions benötigten die User weniger Zeit als erwartet. Die Sessions im erwarteten Zeitraum und darüber reduzierten sich im gleichen Ausmaß. Ursachen für den Unterschied zwischen Clickanzahl und Duration können in der Gestaltung der Website hinsichtlich Navigation, Darstellung und Informationsdichte liegen, sodaß viele Clicks notwendig sind, die Webpages aber schnell erfaßt werden können und der

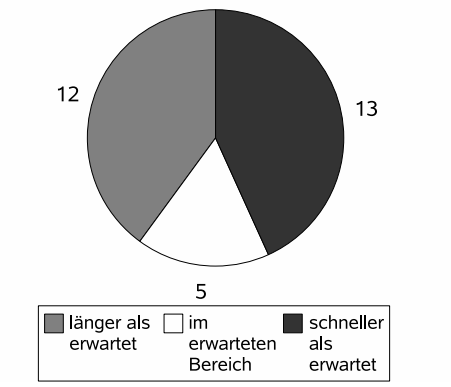


Abbildung 5.12: Erwartete vs. tatsächliche Duration

User schnell zur nächsten Webpage geleitet wird. Die Unterschiede lassen sich auch durch unterschiedliche User-Typen und User-Verhalten erklären, indem die Verweildauer und Suchstrategien sehr individuell ausfallen können. So benutzte ein User stets die Suchfunktion, andere orientierten sich an der Navigationsstruktur.

Die User-Befragung hat zusätzliche Informationen zur Verfügung gestellt:

- die Ziele der User sind bekannt. Sie wurden durch die Aufgabenstellung festgelegt,
- das Erreichen der Ziele wurde erfaßt, sodaß aus User-Sicht der Erfolg einer Session beurteilt werden kann,
- die Erwartungen der User wurden erfragt,
- sowie deren Wahrnehmung der Qualität der Websites

Durch die User-Studie steht das direkte Feedback der User zur Verfügung, das bei der Berechnung des GPI aus serverseitigen Informationen nicht berücksichtigt werden konnte. Im folgenden werden die Sessions anhand der Ergebnisse der User-Studie und des jeweiligen GPI miteinander verglichen.

5.3.3 Vergleich von User-Befragung mit GPI-Bewertung

Die Skalen der User-Befragung können nicht direkt mit denen des GPI verglichen werden. Beispielsweise bewertet der GPI jeden Click innerhalb einer Session und summiert die Werte zu einer Gesamtsessionbewertung auf. Lange Sessions erhalten somit höhere GPI-Bewertungen als kurze Sessions. Dieser Effekt ist gewünscht vor dem Hintergrund, daß lange Sessions vom Website-Autor honoriert werden. Aus User-Sicht kann eine kurze Session schnell zum Ziel führen, eine lange Session dagegen nicht. Um diesen Effekt der Aufsummierung zu eliminieren, werden die GPI-Session-Werte auf GPI-Session-Werte je Click normiert.

Um die Skalen noch besser vergleichbar zu machen, werden für jeden User jeweils die User- wie die GPI-Bewertungen auf einen Mittelwert 0 und eine Standardabweichung 1 skaliert. Dadurch wird der Einfluß von unterschiedlichem Bewertungsverhalten von Usern reduziert. In

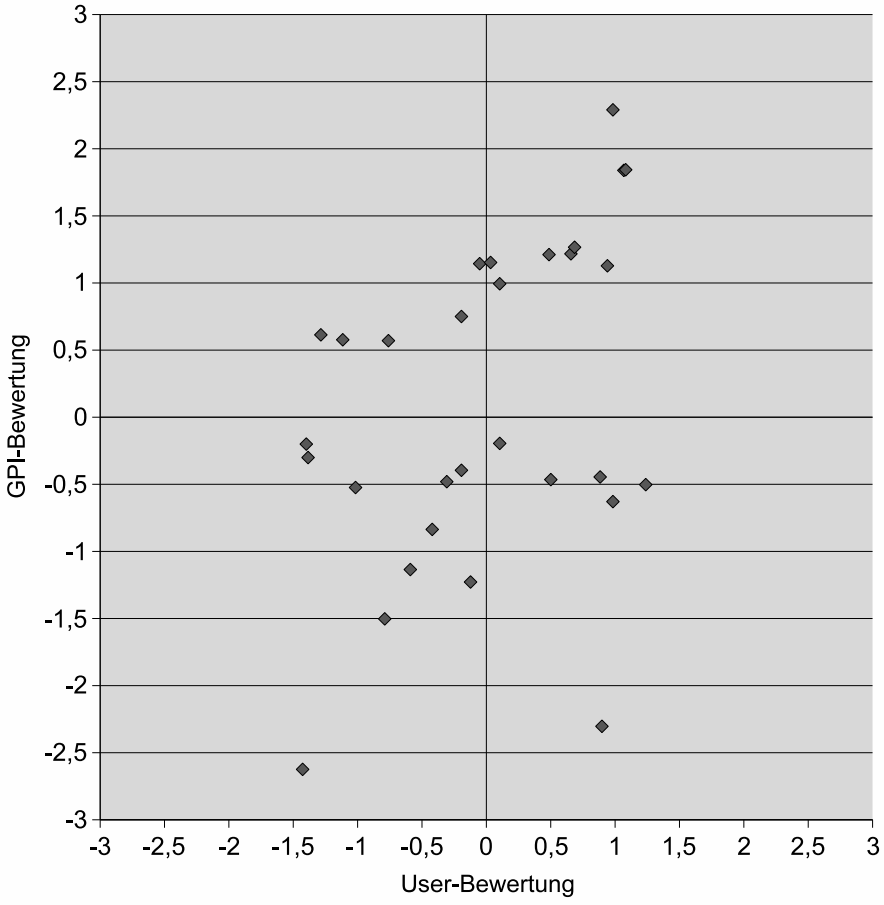


Abbildung 5.13: Vergleich der Bewertungen je Session

Abb. 5.13 sind die standardisierten Bewertungen durch die User den GPI-Scores sessionweise gegenübergestellt. Der obere Teil zeigt die Sessions, die auf Website A durchgeführt wurden, der untere Teil die Sessions auf Website B. Man erkennt die gegensätzlichen Bewertungen. Für eine positive Evaluierung des GPI werden eine grundsätzlich gleichgerichtete Tendenz von GPI und User-Bewertung als ausreichend angesehen. Dazu werden die Korrelationsmaße berechnet.

Der Spearman-Korrelationskoeffizient über alle standardisierten Bewertungen beträgt $0,37$, was auf einen schwach positiv korrelierten Zusammenhang beider Bewertungen hinweist.

Bei der Berechnung des Korrelationskoeffizienten je User erhält man sehr hoch korrelierte

Werte im positiven, wie im negativen Bereich. Hierbei liefern GPI und User-Studie gegensätzliche Beurteilungen. Man kann diese Werte so interpretieren, daß der GPI die Session teilweise genau so sieht wie die User, in einigen Fällen ist die Bewertung allerdings genau konträr. Drei User haben für die Bewertungen ihrer Sessions einen Korrelationskoeffizienten $< \quad 0,79$, drei User $> \ 0,90$, die restlichen User liegen im Bereich von $> \ 0$ und $< \ 0,3$.

Der GPI liefert genau in den Fällen konträre Bewertungen, in denen die User sehr lange benötigten, um das Ziel zu erreichen. Beispielsweise benutzt ein User ausschließlich die Suchfunktion, um zum Ziel zu gelangen. Der GPI bewertet dies sehr negativ, da kein Content vermittelt wird und die Gefahr eines Session-Abbruchs sehr hoch ist. Allerdings findet der User die Zielseite und bewertet seine Session positiv. Hier spielen unterschiedliche User-Typen bzw. User-Verhalten eine Rolle, die vom GPI nicht vorhergesehen werden können und nicht berücksichtigt werden, da keine Informationen über Eigenschaften der User verfügbar sind.

5.3.4 Schlußfolgerung

Die User-Studie zeigt, daß der GPI in der Mehrzahl der Fälle die Sessions so bewertet wie die User. In einigen Fällen jedoch fallen die Beurteilungen genau gegensätzlich aus.

Dies liegt in der Konstruktion des GPI, der lediglich die Informationen verwendet, die auf der Server-Seite verfügbar sind. Dies wird im formalen Website-Modell in Kap. 5.1.6 deutlich, da hier die Semantik auf der User-Seite fast völlig unbekannt bleibt. Dies spiegelt die Problematik des mangelnden direkten User-Feedbacks bei informationsorientierten Websites wider. Folglich bewertet der GPI zwar die Fähigkeit einer Website, einem User Informationen anzubieten, nicht aber dessen Zufriedenheit mit diesem Angebot. Ob der User damit zufrieden ist oder nicht, kann aufgrund mangelnder Daten nicht berücksichtigt werden. Diese Fähigkeit des GPI wurde bei dessen Konstruktion auch nicht beabsichtigt.

Der GPI bewertet User Sessions aus Sicht der Website und ihres Betreibers. Der Erfolg einer Website kann somit teilweise gemessen werden, allerdings nur mit einer eingeschränkten Sichtweise aufgrund eingeschränkter Informationen.

Diesen Mangel an Feedback kennt eine User-Studie nicht. Mit ihr kann man den User direkt nach seinen Zielen, deren Erreichen und die Zufriedenheit des Users befragen. Allerdings ist das Mittel einer direkten Befragung von Usern sehr zeit- und kostenintensiv. Außerdem kann eine Befragungssituation die Antworten beeinflussen. Eine dauerhafte Durchführung im täglichen Betrieb ist nicht möglich.

Der GPI ist in der Lage, das User-Verhalten richtig zu bewerten, wenn die User sich so verhalten, wie es aus Sicht der Website erwartet wird. Unterschiedliche User-Typen und abweichendes Browsing-Verhalten können jedoch zu konträren Bewertungen führen. Somit ist der

GPI ein Maß, welches das User-Verhalten aus Website Sicht richtig wiedergibt, Rückschlüsse auf die User-Seite aber nicht zuläßt.

Es existiert in einigen Sessions eine noch nicht genutzte Informationsquelle, die eine genauere Analyse der User-Seite eröffnet. Im folgenden Abschnitt wird eine Möglichkeit beschrieben, wie basierend auf Suchanfragen an Suchmaschinen, die Absichten der User erfaßt werden können. Kennt man die Absichten der User, kann ein Erfolgsmaß erstellt werden, das dies berücksichtigt.

5.4 Intention Based GPI

Der GPI bewertet aus Mangel an Informationen über die User-Seite eine Website aus Sicht des Website-Autors. Dabei steht die Erfüllung der Ziele einer informationsorientierten Website im Vordergrund - die Verbreitung von Informationen. Es wäre für eine umfassende Bewertung einer Website wünschenswert, die Bewertungsperspektive auf die User-Seite erweitern zu können.

Man erhält in den meisten Fällen, in denen ein User durch Benutzung einer Suchmaschine auf die analysierte Website gelangt, über den Referrer dieses Clicks (Page Views) die Suchanfrage dieses Users. Die Annahme liegt nah, daß der User in einer Suchanfrage an eine Suchmaschine seine Ziele offenbart, die er auf einer Website erreichen möchte.

5.4.1 Suchstrings

Von allen Sessions werden nur diejenigen betrachtet, deren User durch eine externe Suchmaschine, wie Google, Yahoo, Vivisimo, auf die zu analysierende Website geleitet wurden. Diese User haben auf der Suchmaschine aus einer Ergebnisliste gezielt den Link zu der analysierten Website ausgewählt.

Die Session wird mit dem ersten Click auf diese Website vom Web Reporting System erfaßt. Was der User vorher auf der Suchmaschine gemacht hat oder welche anderen Websites er ausgewählt hat, bleibt auf der Server-Seite verborgen. Der erste vom Web Reporting System aufgezeichnete Click trägt in dessen Referrer die Suchanfrage, die durch einen Link auf die Website geführt hat.

Verwandte Arbeiten

Suchmaschinen sind mit deren rasant wachsender Bedeutung im Internet seit längerem Gegenstand wissenschaftlicher Untersuchungen. Hauptsächlich werden Suchstrategien und die Effizienz und Effektivität der Suchmaschinen sowie deren Algorithmen untersucht.

Broder gibt in [39] einen Überblick der Suche im Web. Er unterscheidet bei den Zielen der User zwischen *Navigations-, Informations- und Transaktionsabsichten*. Das Ziel eines Users mit einer navigationsorientierten Suchanfrage ist es, eine bestimmte Website zu erreichen, wohingegen User, die nach bestimmten Informationen suchen, eine informationsorientierte Suchanfrage stellen. Suchanfragen, die auf den Abschluß einer Transaktion abzielen, sind nicht im Fokus dieser Arbeit, da auf den hier betrachteten Websites keine Transaktionen stattfinden. Ebensowenig sind navigationsorientierte Suchanfragen von Interesse, da ein User mit dem ersten Click bereits die Website erreicht hat und damit sein Navigationsziel erfüllt ist.

Diese Arbeit konzentriert sich auf Suchanfragen, die nach einer bestimmten Information suchen, also *informationsorientierte* Suchanfragen.

Lee et al. konzentrieren sich in [160] bei ihren Analysen der User-Ziele auf die Optimierung der Suchmaschine. Wie Teevan et al. in [247], beschreiben sie das User-Ziel als inhärent subjektiv. Teevan sieht Suchanfragen als unpräzise an. Grundsätzlich kann dieser Ansicht zugestimmt werden, allerdings hat ein User bereits einen Link auf die analysierten Websites ausgewählt und somit sein Interessengebiet auf den Kontext dieser Website bewußt eingeschränkt. Dadurch erhöht sich die Präzision der Suchanfragen, wenn man Irrtümer der User ausschließt. Wie Broder et al. in [40], analysieren Lee et al. Suchanfragen aus der Sicht der Suchmaschine. Für diese Dissertation wird eine neue Sichtweise vorgeschlagen. Suchanfragen von Suchmaschinen sollen zur weiteren Analyse des User-Verhaltens auf Websites eingesetzt werden, auf die eine Suchmaschine verlinkt hat.

Shen et al. untersuchen in [215] die aufgerufenen Webpages, die ein User aus den Ergebnissen einer Suchanfrage auswählt. Wie auch in dieser Arbeit unterstellen sie, daß die Ziele eines Users in einer Suchanfrage repräsentiert sind. Unter dieser Annahme unterstellen sie eine implizite Verlinkung, die durch den User erzeugt wurde. Hahn et al. [113, 112] bestätigen diese Ansicht darin, daß bei gegebenem User-Ziel ein wahrscheinlicher Clickstream prognostiziert werden kann. Umgekehrt kann man bei Kenntnis des Navigationspfades sowie der Struktur und des Inhalts einer Website auf die Interessen der User zurückschließen.

Datenaufbereitung der Suchstrings

Eine Suchanfrage besteht aus mindestens einem Suchbegriff, also einem Wort. Nicht alle Suchanfragen helfen dabei weiter, wenn man auf die Absichten der User schließen will. Folgt man der Kategorisierung von Broder in [39], tragen *navigationsorientierte* Suchanfragen keine relevanten Informationen zur Erklärung der User-Absichten. Ein Beispiel für eine navigationsorientierte Suchanfrage ist der Name der Website, bzw. des Unternehmens, der keine weiteren Aufschlüsse über die User-Absichten gibt. Mit Erreichen der Website ist das Ziel des Users bereits mit dem ersten Click erfüllt. Die Suchanfrage ist damit für die restliche Session irrelevant, da das **initiale Interesse** bereits erfüllt ist.

Der Suchstring auf einer referierenden Suchmaschine trägt maximal das initiale Interesse eines Users für eine Session. Mit jedem Click können sich die Absichten, Ziele und Intentionen eines Users ändern. Mit Ausschluß der *navigations- und transaktionsorientierten* Suchanfragen bleiben nur die *informationsorientierten* Suchstrings übrig. Diese Suchstrings werden als das initiale User-Interesse beim Betreten der Website interpretiert. Ein Interessenwechsel des Users während seiner Session bleibt verborgen. Das initiale Interesse bleibt der einzige, direkt greifbare Anhaltspunkt der User-Absichten.

Die Worte aus den Suchstrings wurden in der in Kap.5.2.3 beschriebenen Datenaufbereitung nicht berücksichtigt. In Tabelle 5.10 ist die Bereinigung der Suchstrings zusammengefaßt. Der Anteil an Sessions mit Suchstrings variiert zwischen 8 % und 20 %. Aus den Suchstringsessions lassen sich nach Eliminierung der *navigations- und transaktionsorientierten* Suchanfragen 6-19% der ursprünglichen Sessions verwenden. Die übriggebliebenen informationsorientier-

Website	Zeitraum	Sessions gesamt	Sessions mit Suchstring	bereinigte Suchstringsessions
A	Nov.-Dez.05	9.877	750	564
A	Jan.06	5.288	549	319
B	Nov.-Dez.05	1.851	195	186
B	Jan.06	939	191	180

Tabelle 5.10: Datenaufbereitung der Suchstringsessions

ten Suchstrings werden zunächts wie in Kapitel 4.3 beschrieben aufbereitet und bereinigt. Die einzelnen Worte einer Suchanfrage werden auf ihre Wortstämme reduziert und festgelegte Stopwords entfernt. Damit ist der Datenaufbereitungprozeß für die Suchanfragen beendet. Der folgende Abschnitt beschreibt die Einbindung von Suchstrings in das bestehende Website-Modell. Es wird beschrieben, wie diese zusätzliche Information genutzt werden kann, um eine verbesserte Website-Bewertung zu ermöglichen.

Zusätzlich zu der Entfernung von Stopwords und der Reduzierung auf die Wortstämme werden solche Suchstrings entfernt, die auf der kompletten, analysierten Website nicht vorkommen. Sie werden separat behandelt, um zu erkennen, welche Informationen gesucht wurden, aber nicht auf der Website gefunden werden konnten. Neben Worten, die nicht vorkommen, werden solche Suchbegriffe entfernt, die auf allen Webpages einer Website vorkommen. Sie zählen zu den *navigationsorientierten Suchanfragen* und bringen keine weitere Einsicht in die User-Interessen.

5.4.2 Erweitertes Website-Modell

Im folgenden wird die Referrer Information und die darin enthaltenen Suchanfragen von Suchmaschinen in das bestehende formale Modell einer Website aus Kap. 5.1.6 integriert. Das in Abb. 5.14 gezeigte Modell ist wie das Modell des GPI in Abb. 5.5 in syntaktische, semantische und Maß-Ebene, sowie die User- und Website-Sicht unterteilt. Die syntaktische Ebene beschreibt die direkt beobachtbaren Fakten, während die semantische Ebene die Interpretationen dieser Fakten und die getroffenen Annahmen beschreibt. Die Referrer Information ist Bestandteil eines Clicks (*SessionItems*) und damit Bestandteil der Klasse *SessionStructure*, die die beobachtbaren Fakten auf der User- und Session-Seite beschreibt. Sofern möglich, kann jedem Click ein Referrer zugeordnet werden. Das ergibt im Website-Modell die Assoziation von $0 \ldots 1$ zwischen *SessionItems* und *ReferrerInfo*. Nicht jeder Referrer enthält Suchanfragen (*SearchQuery*) einer Suchmaschine, die ein User benutzt hat, um auf die Website zu gelangen

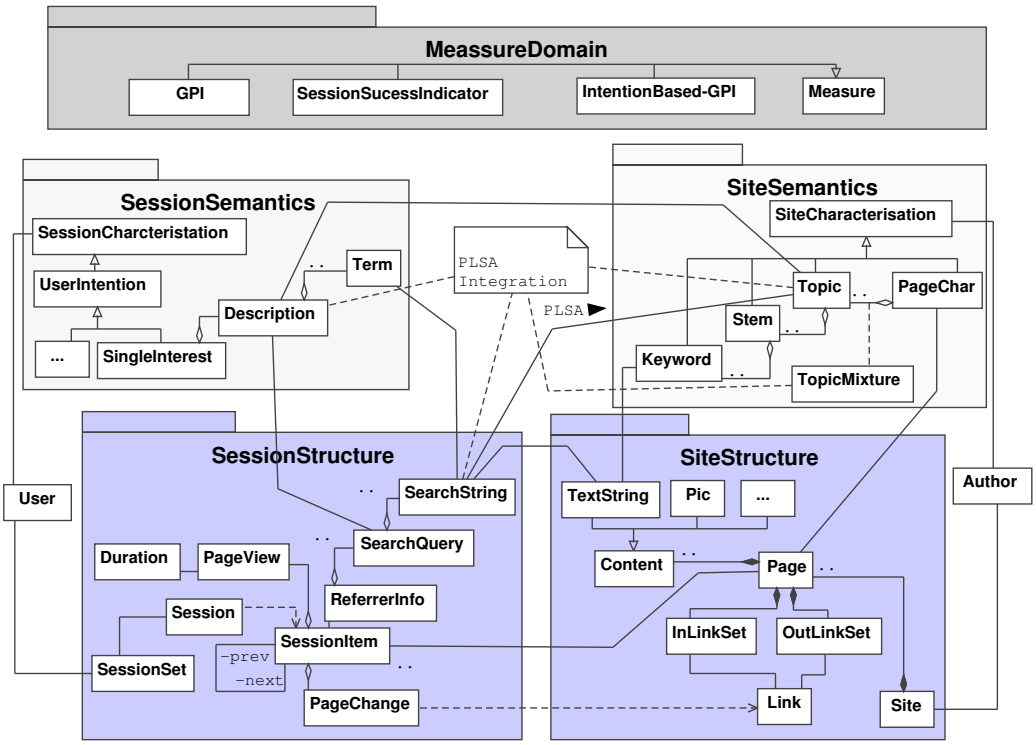


Abbildung 5.14: Erweitertes, formales Modell einer Website zur Berechnung des ibGPI

- daher die Assoziation $0 \ldots 1$. Jede Suchanfrage besteht aus einem oder mehreren Suchstrings (*SearchStrings*).

Eine weitere Änderung des Modells ergibt sich auf semantischer Ebene auf Seiten des Users. Durch die Interpretation der Suchstrings als initiales Interesse der User, kann die *UserIntention* in der semantischen Klasse *SessionSemantics* durch den *SingleInterest* näher beschrieben werden, der sich aus dem *SearchQuery* ableitet.

Um die Suchstrings eines Users mit den besuchten Webpages seiner Session in Verbindung bringen zu können, müssen deren Inhalte miteinander verglichen werden. Der inhaltliche Zusammenhang gibt Aufschluß über die Erreichung der User-Ziele. Kern des neuen Website-Modells stellt der inhaltliche Zusammenhang zwischen dem Inhalt der Webpages und dem Inhalt der Suchstrings auf der User-Seite dar. Der Zusammenhang wird durch den *PLSA*-Algorithmus hergestellt. Als Ergebnis des *PLSA*-Algorithmus erhält man eine Zuordnung von Webpages zu Themen und eine Zuordnung von Wörtern zu Themen. Über die Zuordnung der Wörter zu Themen lassen sich auch die Suchstrings diesen Themen zuordnen.

Beurteilt man die Webpages einer Session unter Berücksichtigung des User-Interesses aus den Suchstrings, kann der Inhalt jeder Webpage, individuell für jede Suchanfrage und damit

individuell für jedes User-Interesse, analysiert und bewertet werden.

In der Klasse *MeasureDomain* befinden sich neben dem bisherigen GPI, der weiterhin Verwendung findet, auch dessen Erweiterungen, die im folgenden beschrieben werden.

5.4.3 Intention Based GPI

Der in Kap. 5.1.5 beschriebene GPI kann die User-Sicht aus Mangel an Informationen nicht berücksichtigen und bewertet eine Website lediglich aus Sicht des Website-Autors. Wie in der User-Studie gezeigt wurde, genügt diese Sichtweise nicht, um eine Website umfassend zu analysieren. Daher wird zusätzlich zum klassischen GPI ein auf Suchstrings basierender GPI vorgeschlagen, der in den Fällen, in denen ein Suchstring als Referrer-Information vorliegt, die User-Absichten berücksichtigt. Die neue Berechnungsweise soll einen GPI ermöglichen, der die manuellen Eingriffe und Einschätzungen durch eine generische Berechnung ersetzt.

Erweiterung des klassischen GPI

Der klassische GPI wurde in Kap. 5.1.5 wie folgt definiert: Kombiniert man die Effektivitätsmaße Transitionstyp (*effectivitytype*) und Transitionsgewicht (*effectivityweight*) mit dem Effizienzmaß Durationgewicht (*efficiency*), erhält man durch deren Multiplikation den GPI-Wert für eine Transition. Die drei Teilmaße wurden so gewählt, daß sie multipliziert den GPI einer Transition ergeben. Summiert man über die Transitionen $_i$ einer User Session s, erhält man den GPI dieser Session s. Dabei umfaßt i alle Elemente einer Session und reicht vom ersten Element $i = 1$ bis zum letzten Element, das der Gesamtzahl aller Elemente $|s|$ einer Session entspricht.

$$GPI_s = \sum_{i=1,\ _i \in s}^{|s|} \textit{effectivitytype}_{\ i} \quad \textit{effectivityweight}_i \quad \textit{efficiency}_i \tag{5.7}$$

Bei der Berechnung des klassischen GPI legt der **Transitionstyp** *effectivityweight* das Vorzeichen des Effektivitätsmaßes fest. Der Transitionstyp *effectivityweight* wurde aus dem Übergang zwischen Navigations- und Contentseiten manuell festgelegt, siehe Tab. 5.1. Das Transitionsgewicht *effectivityweight* $_i$ wurde unter Berücksichtigung des Grades der Themenveränderung abgeleitet bzw. manuell festgelegt. Daraus ergibt sich, wie in Abb. 5.15 gezeigt, die Transitionsbewertung bezüglich deren Effektivität.

Der **Effizienzfaktor** *efficiency* des klassischen GPI in Gleichung 5.7 bleibt unverändert, da die Duration in diesen Analysen immer gleich gemessen wird. Es sind keine zusätzlichen Informationen auf der Server Seite verfügbar, die eine genauere Erfassung der Aufmerksamkeit des Users erlauben. Die Einschränkungen der Interpretierbarkeit der Duration, die in

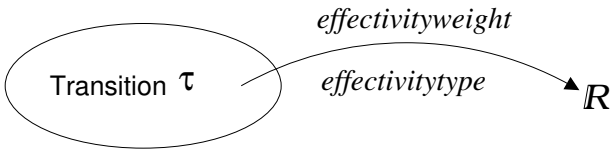


Abbildung 5.15: Schematische Herleitung des klassischen GPI-Effektivitätsmaßes

Kap. 3.3.1 beschrieben wurden, gelten weiter. Es kann somit nicht festgestellt werden, wie lange der User seine Aufmerksamkeit einer Webpage widmet, sondern lediglich, wie lang die Webpage geöffnet war.

Da durch die Referrer Information genauer beschrieben werden kann, inwieweit ein User die richtige Webpage erreicht hat, kann die **Effektivität** der Webpage genauer gemessen werden. Die Effektivitätsmaße *effectivitytype* und *effectivityweight* einer Transition müssen angepasst werden, um die Suchstring-Informationen aus den Referrern integrieren zu können.

Um die Suchstrings und deren Bedeutung nutzen zu können, müssen die Beziehungen zwischen den Elementen einer Website, wie sie im erweiterten formalen Modell einer Website in Abb. 5.14 dargestellt sind, in den Bewertungsalgorithmus eingebaut werden.

Das Ziel ist eine Bewertungsfunktion, deren einzige Eingabe aus einer Transition besteht. Hierzu sind folgende Aufgaben zu erfüllen:

1. Jeder Transition soll die Suchanfrage der jeweiligen Session zugewiesen werden.
2. Diese Suchanfrage soll mit den Inhalten aller Webpages verglichen werden.
3. Aus diesem inhaltlichen Vergleich soll ein Zielmaß entwickelt werden, das den Grad der Übereinstimmung der Suchanfrage mit einer Webpage bestimmt.
4. Aus dem Grad der Zielerreichung wird die Bewertung der Transition bestimmt, die das neue Effektivitätsmaß eines modifizierten GPI bildet.

Abb. 5.16 zeigt die Erweiterungen des Effektivitätsmaßes aus Abb. 5.15. Die Effektivitätsmaße *effectivityweight* und *effectivitytype* werden ersetzt. Im folgenden werden die einzelnen Schritte bis zur Neugestaltung des GPI hergeleitet. Ausgehend von einer Transition wird der dazugehörige Suchstring mit der Funktion $query$ bestimmt. Der Suchanfrage wird durch die Funktion $topic$ ein Themenvektor zugewiesen. Dieser Themenvektor bestimmt durch die Funktion $higherorder$ die Bewertungsfunktion *effectivity*$_{ibGPI}$, die den Effizienzwert der Transition ermittelt.

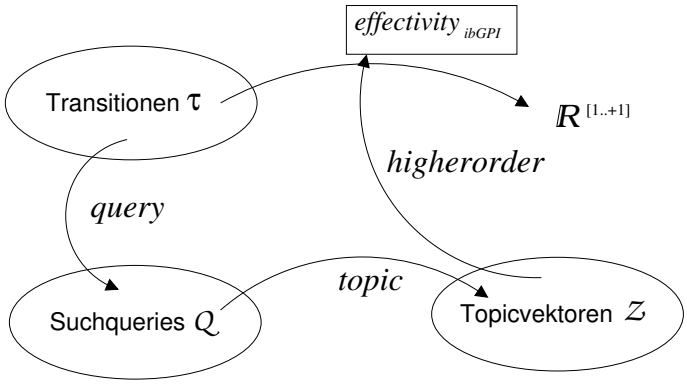


Abbildung 5.16: Schematische Herleitung des neuen GPI-Effektivitätsmaßes

Suchqueryzuordnungsfunktion $query$

Die erste Aufgabe besteht in der Zuordnung einer Transition zu der zugehörigen Suchanfrage $q \in Q$. Eine Transition $\in \mathcal{T}$ sei Element einer Session s und beschreibt den Übergang von einer Start- auf eine Ziel-Webpage. Diese Transition $_s$ korrespondiert zu einer Suchanfrage (SearchQuery) $q \in \mathcal{Q}$. Für jede Session sei eine Suchanfrage verfügbar, die einer Transition durch folgende Funktion zugewiesen wird: $query : \mathcal{T} \rightarrow \mathcal{Q}$ Abb. 5.17 zeigt eine beispielhafte Umsetzung von $query$.

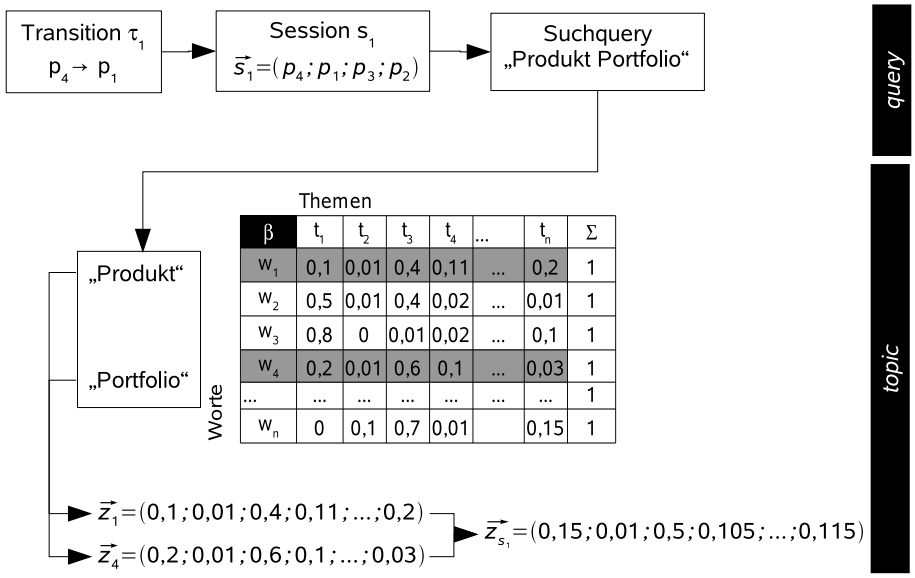
Berechnung der Themenvektoren für Suchstrings mittels der Funktion $topic$

Die Suchanfragen wurden bislang nicht inhaltlich analysiert. Sie müssen in die inhaltliche Analyse der Website durch den PLSA integriert werden. Eine Funktion $topic$ soll einer Suchquery $q \in \mathcal{Q}$ einen Themenvektor $\overrightarrow{z} \in \mathcal{Z}$ zuweisen.

$$topic : \mathcal{Q} \rightarrow \mathcal{Z}$$

In Worten: Eine Funktion *topic* weist jeder Suchanfrage q aus der Menge der Suchanfragen $\mathcal{Q}$ einen Themenvektor $\overrightarrow{z}$ aus der Menge der Themenvektoren $\mathcal{Z}$ zu.

Die Funktion $topic$ wurde in dieser Arbeit wie folgt berechnet: Eine Suchanfrage kann aus mehreren einzelnen Worten (Strings) bestehen. Die aus der Suchanfrage abgeleiteten Suchstrings werden dazu benutzt, um einen Themenvektor $\overrightarrow{z} \in \mathcal{Z}$ zu einem Suchstring zu berechnen. Hierbei korrespondiert $\overrightarrow{z}$ mit der Definition von $\{z_j\}_{j=1}^{L}$ aus Kap. 4.3.4 im Anhang. Der Themenvektor $\overrightarrow{z}$ leitet sich aus der Matrix ab, in der der PLSA (siehe: Kap2.3, Kap. 4.3.4 und Kap. 5.1.2) die Zuordnung von Wörtern zu Themen berechnet hat. Durch die Matrix kann für jede Webpage ein Themenvektor $\overrightarrow{z}$ erstellt werden. $\overrightarrow{z}$ repräsentiert dabei die berechnete, bedingte Wahrscheinlichkeit für ein Wort, zu jedem Thema der Website zu gehören.



β	t_1	t_2	t_3	t_4	...	t_n	Σ
w_1	0,1	0,01	0,4	0,11	...	0,2	1
w_2	0,5	0,01	0,4	0,02	...	0,01	1
w_3	0,8	0	0,01	0,02	...	0,1	1
w_4	0,2	0,01	0,6	0,1	...	0,03	1
...	...	...	...	...	...	...	1
w_n	0	0,1	0,7	0,01		0,15	1

Abbildung 5.17: Funktionen $query$ und $topic$

Der durchschnittliche Themenvektor für eine Suchanfrage wird aus den einzelnen Themenvektoren aus Matrix je Suchstring dieser Suchanfrage berechnet. Man erhält die Themenmischung t für eine Suchanfrage. Jetzt lassen sich die Themenvektoren der Webpages mit dem Themenvektor der Suchanfrage vergleichen.

Eine beispielhafte Berechnung von $query$ und $topics$ wird in Abb. 5.17 beschrieben. Ausgehend von einer Transition $_1$ wird durch die Funktion $query$ über die dazugehörige Session der Suchquery zugewiesen. Die Funktion $topic$ sucht zu jedem Suchstring der Suchanfrage aus der Matrix die Themenvektoren. Als Durchschnitt der einzelnen Themenvektoren berechnet $topic$ den Themenvektor $\overrightarrow{z_{s_1}}$.

Funktion höherer Ordnung $higherorder$ zur Bestimmung von *effectivity*$_{ibGPI}$

Nachdem den Suchstrings einer Suchquery Themenvektoren zugeordnet wurden, ist das initiale User-Interesse darin repräsentiert und verfügbar. Damit können die Transitionen einer Session bewertet werden. Die Bewertung der Effektivität einer Transition erfolgt durch eine Funktion *effectivity*$_{ibGPI}$, die die Transitionen in Abhängigkeit von der jeweiligen Suchanfrage einer Session bewertet. Dadurch erhält man unterschiedliche Funktionen *effectivity*$_{ibGPI}$ in Abhängigkeit vom Topicvektor $\overrightarrow{z}$einer Suchanfrage. Jede Session und ihre Transitionen werden auf diese Weise durch eine individuelle Bewertungsfunktion *effectivity*$_{ibGPI}$ bewertet. Eine Funktion höherer Ordnung soll die Erzeugung der individuellen Bewertungsfunktionen

effectivity$_{ibGPI}$ beschreiben. In Abb. 5.16 sieht man, daß die Funktion $higherorder$ nicht eine Menge (z.B. Suchqueries) verändert, sondern die Funktion *effectivity*$_{ibGPI}$ selbst verändert. Da sie als Funktion auf eine andere Funktion wirkt, spricht man von einer *Funktion höherer Ordnung*.

$$higherorder \ : \ \mathcal{Z} \rightarrow \ \mathcal{T} \rightarrow \mathbb{R} \tag{5.8}$$

In Worten: Die Funktion $higherorder$ nimmt einen Themenvektor $\overrightarrow{z}$aus der Menge der Themenvektoren $\mathcal{Z}$ und erzeugt eine Funktion, die eine Transition aus der Menge der Transitionen $\mathcal{T}$ auf den reellen Zahlenraum abbildet.

Die Funktion $higherorder$ wurde in dieser Arbeit mittels einer **themenbasierten Webpage Rankingfunktion** wie folgt umgesetzt:
Aus den inhaltlichen Übereinstimmungen zwischen den Suchstrings einer Session und den Webpage-Inhalten einer Website, lassen sich die Webpages nach dem Grad der inhaltlichen Übereinstimmung sortieren. Auf diese Weise läßt sich eine Rangfolge bilden, bei der die Webpage mit der größten inhaltlichen Übereinstimmung mit einer Suchquery den höchsten Rang (eins) erhält. Die Funktion $topic$ hat jeder Suchanfrage q den dazugehörigen Themenvektor $\overrightarrow{z}$zugeordnet. Die Themenvektoren der Webpages $\mathcal{P}$ liegen durch die Matrix vor, die durch den PLSA bereits in Kap. 5.1.2 erzeugt wurde und die Form $Webpage \quad Themen$ hat.

Zur Messung der Ähnlichkeit bzw. des Abstands zwischen den Themenvektoren berechnet man beispielsweise den Euklidischen Abstand zwischen den Themenvektoren der Suchanfrage und denen der Webpages. Die Funktion $rank$ benutzt die Euklidische Distanz zwischen dem Themenvektor $\overrightarrow{z}$der Suchanfrage und den Themenvektoren der Webpages$\mathcal{P}$ aus der Matrix . Man erhält für alle $p \in \mathcal{P}$ Distanzen, die die Entfernung zur Suchanfrage q angeben. Die Funktion $rank$ ordnet die Webpages nach steigender inhaltlicher Distanz, sodaß die Webpage mit der geringsten Themendistanz zu der Suchanfrage den höchsten Rang (eins) erhält.

$$rank \ : \ \mathcal{Q} \quad \mathcal{P} \rightarrow \mathbb{N}^{+} \tag{5.9}$$

In Worten: Die Funktion $rank$ weist durch Eingabe einer Suchanfrage q aus der Menge der Suchanfragen $\mathcal{Q}$ jeder Webpage p aus der Menge der Webpages $\mathcal{P}$ eine positive natürliche Zahl zu. Die Idee eines Rankings und deren Anwendungen werden von Pujol et al. [196, S.380ff] zusammengefasst. Zu den bekanntesten Ranking-Algorithmen gehören der *Page-Rank*- [37] und der *HITS*-Algorithmus [148].

Für jede Suchanfrage (Query) q erhält man ein Ranking aller Webpages $p \in \mathcal{P}$ mit $1 \quad rank(q,p) \quad |\mathcal{P}|$, das bedeutet, die Funktion $rank$ liefert Werte im Intervall $[1..Anzahl(Webpages)]$. Durch Berücksichtigung der Gesamtzahl an Webpages auf einer Website kann der Rang auf eine Skala von $[0;1]$ umgerechnet werden, sodaß die Webpage mit Rang 1, also mit der höchsten Übereinstimmung von Suchstring und Webpage Inhalt, auch auf dieser Skala eine 1 erhält. Die Webpage mit dem niedrigsten Rang erhält 0.

$$rank \ : \ \mathcal{Q} \quad \mathcal{P} \rightarrow \mathbb{R}^{[0..1]} \tag{5.10}$$

Dieses Ranking wird für jede Session s bzw. jede Suchanfrage q individuell erstellt. Durch Verwendung der Themen aus der PLSA, wird die inhaltliche Nähe einer Suchanfrage zu den Webpages einer Website nicht durch direkten Wortvergleich hergestellt, sondern über die Themen einer Website. Dadurch können semantische Zusammenhänge berücksichtigt werden, ohne daß ein Wort sowohl im Suchstring als auch auf einer Webpage gemeinsam vorkommen muß.

Die Veränderungen, die die Funktion $higherorder$ in der Funktion *effectivity*$_{ibGPI}$ bewirkt, resultieren aus der individuellen Rankingfunktion $rank$, die für jede Suchquery individuell erzeugt wird. Dadurch ändert sich der Rang der Webpages für jede Suchanfrage. Aus dem unterschiedlichen Ranking kann die endgültige Bewertungsfunktion, wie in Abb. 5.16 dargestellt, die Bewertung einer Transition vornehmen.

Eine beispielhafte Berechnung der $higherorder$-Funktion und deren Rankingfunktion $rank$ zeigt Abb. 5.18.

Ausgehend von dem in Abb. 5.17 berechneten Themenvektor $z_{s_1}^{\rightarrow}$, berechnet die $higherorder$-Funktion individuell für jede unterschiedliche Suchquery eine neue Rankingfunktion $rank$. Hierzu wird die inhaltliche Nähe der Webpages zum Themenvektor $z_{s_1}^{\rightarrow}$berechnet, indem die Euklidische Distanz zwischen den Themenvektoren der Webpages und dem Themenvektor der Suchquery gebildet wird. Die Funktion $rank$ sortiert die Webpages nach steigender Distanz und weist entsprechend deren Position einen Rang zu, sodaß die Webpage mit der geringsten Distanz den Rang *1* und die weitest entfernte Webpage den Rang *Anzahl Webpages* erhält. Der ganzzahlige Rang wird durch Normalisierung in einen Wert im Intervall $[0..1]$ umgewandelt. Dadurch werden die Ränge unabhängig von der Anzahl der Webpages, sodaß ein Vergleich zwischen Websites und unterschiedlichen Zeitpunkten möglich wird. Der Charakter der $higherorder$-Funktion wird deutlich, wenn man statt des Vektors $z_{s_1}^{\rightarrow}$andere Themenvektoren $z_{s_n}^{\rightarrow}$einsetzt. Andere Themenvektoren der Suchquery führen zu anderen Distanzen zwischen Suchanfrage und Webpages, wodurch das komplette Ranking sich individuell für jeden unterschiedlichen Suchquery berechnet.

Effektivitätsbewertungsfunktion *effectivity*$_{ibGPI}$

Die Funktion *effectivity*$_{ibGPI}$ wird durch die Funktion höherer Ordnung $higherorder$ bestimmt. *effectivity*$_{ibGPI}$ bewertet die Effektivität einer Transition auf einer Website.

$$effectivity_{ibGPI} \ : \ \mathcal{T} \rightarrow \mathbb{R}^{[\quad 1..+1]} \tag{5.11}$$

In Worten: Die Funktion *effectivity*$_{ibGPI}$ weist einer Transition aus der Menge aller Transitionen $\mathcal{T}$ einen Wert im reellen Zahlenraum im Intervall $[\quad 1.. + 1]$ zu.

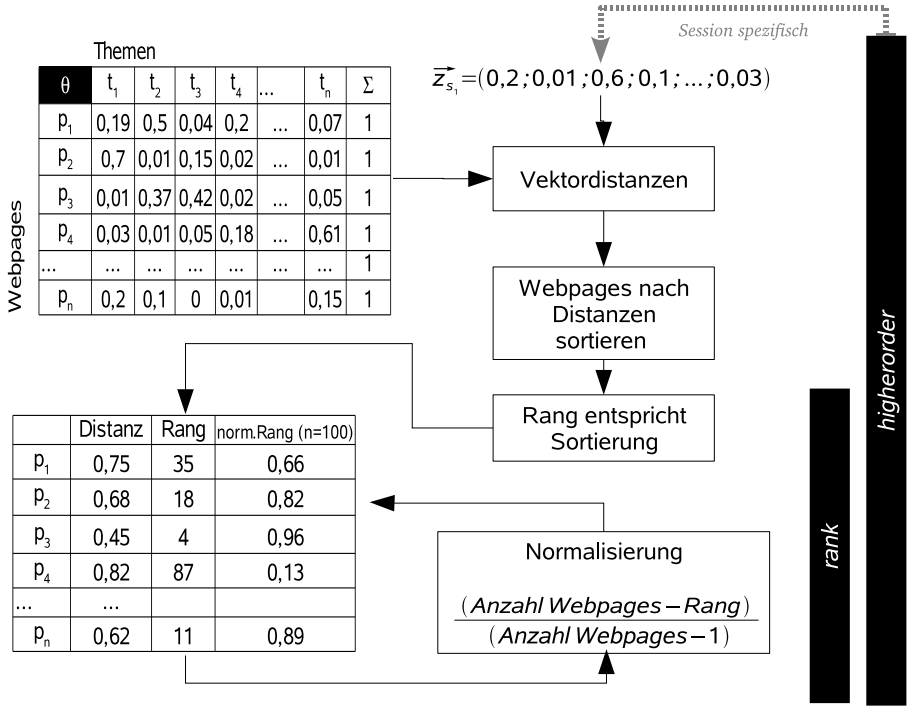


Abbildung 5.18: Funktionen $higherorder$ und $rank$

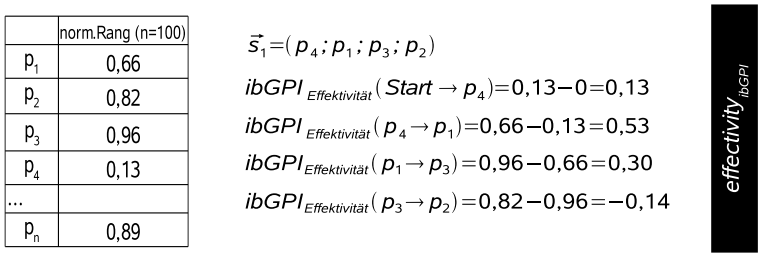


Abbildung 5.19: Funktion *effectivity*$_{ibGPI}$

Die Funktion *effectivity*$_{ibGPI}$ ergibt sich aus der Funktion $higherorder$, die den Webpages p durch die Funktion $rank$ einen Rang zugewiesen hat. In dieser Arbeit und für die analysierten Webpages wird *effectivity*$_{ibGPI}$ auf die nachfolgend beschriebene Weise berechnet. Betrachtet man die Veränderung des Ranges der Webpages durch eine Transition, läßt sich dadurch die Effektivität dieser Transition bestimmen.

Definition: *effectivity*$_{ibGPI}$ *Eine Transition ist dann als erfolgreich anzusehen, wenn der Rang ($rank$) der Zielseite einer Transition p_{target} höher ist als der der Startseite p_{source}. Die Effektivität einer Transition läßt sich durch die Differenz zwischen den Rängen von Ziel- und Startseite ($rank_{p_{source}}$ $rank_{p_{target}}$) berechnen.*

Da die Werte der Funktion $rank$ auf das Intervall $[0..1]$ normiert wurden, liegt das Ergebnis aus der Differenz zwischen p_{source} und p_{target} im Intervall $[\ \ 1.. + 1]$. Eine Transition, die auf eine Webpage mit niedrigerem Rang führt, erhält einen Effektivitätswert im Bereich von $[\ \ 1..0]$. Eine effektive Transition erhält einen Wert in $[0..+1]$. Auf diese Weise ergibt sich das Effektivitätsmaß sowohl mit dem gewünschten Vorzeichen, als auch mit einer auf eine Website standardisierten Gewichtung. Für alle Transitionen liegt das Effektivitätsmaß auf einer Skala von -1 bis 1.

Eine beispielhafte Berechnung der *effectivity*$_{ibGPI}$-Funktion in Abb. 5.19 zeigt den letzten Schritt in der Berechnung des Effektivitätsmaßes für die Transitionen $_1$ $_4$ einer Session s_1. Hierzu wird berechnet, wie sich der Rang der Webpage durch die zu bewertende Transition verändert. Mit der ersten Transition $_1$ im Beispiel startet der User auf Webpage $Start \rightarrow p_4$. Man zieht vom Rang der erreichten Webpage den Rang der vorherigen Seite ab, hier also $0,13$ $0 = 0,13$. Die nächste Transition $_2$ geht von Webpage p_4 auf Webpage p_1, also $p_4 \rightarrow p_1$. Der Rang der Ausgangsseite p_4 wird vom Rang der Zielseite p_1 abgezogen: $0,66$ $0,13 = 0,53$. Dieser Effektivitätswert zeigt eine positive Veränderung, was bedeutet, daß sich der User inhaltlich in Richtung seines Ziels bewegt, das durch den Suchquery repräsentiert wird. Mit der dritten Transition innerhalb seiner Session bewegt sich der User weiter

auf sein Ziel zu, wohingegen sich der User mit seiner letzten Transition vom Ziel entfernt. Dies resultiert in einem negativen Effektivitätsmaß.

Das Effektivitätsmaß bewertet allerdings nicht nur die inhaltliche Richtung, in die sich der User bewegt, sondern gibt auch Auskunft über den Grad der Annäherung an die Webpage mit der größten Übereinstimmung mit dem Suchquery eines Users. Mit der ersten Transition betritt der User die Website auf einer Webpage mit Rang $0,13$. Das bedeutet, daß 13 % der Webpages inhaltlich weniger gut mit dem User-Interesse, bzw. Suchquery, übereinstimmen, aber 87 % der Webpages passendere Informationen bereithalten. Die Bewertung der zweiten Transition von $0,53$ kann so interpretiert werden, daß der User 53 % der Webpages in Richtung der optimalen Webpage übersprungen hat. Mit der dritten Transition verbessert sich der User um weitere 30%, sodaß er fast die optimale Webpage erreicht hat. Diese Transition stellt die maximale Annäherung des Users in dieser Session an sein Ziel dar. Diese maximale Zielannäherung wird weiter unten in Abschnitt 5.4.3 durch die Einführung des SSI als zusätzliches Maß gemessen.

Intention Based GPI - ibGPI

Die Funktion *effectivity*$_{ibGPI}$ ersetzt sowohl den Transitionstyp *effectivityweight* als auch das Transitionsgewicht *effectivitytype* des klassischen GPI.

Um die neue Bewertungsfunktion *effectivity*$_{ibGPI}$ integrieren zu können, ist der in Abbildung 5.16 gezeigte Herleitungsweg zu integrieren. *effectivity*$_{ibGPI}$ wird durch die Funktion höherer Ordnung $higherorder$ näher beschrieben. $higherorder : \mathcal{Z} \rightarrow \mathcal{T} \rightarrow \mathbb{R}$. Die Themenvektoren $\mathcal{Z}$ der Suchanfragen werden durch die Funktion $topic\ \mathcal{Q} \rightarrow \mathcal{Z}$ bestimmt. Die Suchanfragen $\mathcal{Q}$ werden durch die Funktion $query$ einer Transition zugewiesen.

$$effectivity_{ibGPI} : \mathcal{T} \rightarrow \mathbb{R}^{[\ 1..+1]} \tag{5.12}$$

$$higherorder : \mathcal{Z} \rightarrow \mathcal{T} \rightarrow \mathbb{R} \tag{5.13}$$

$$topic : \mathcal{Q} \rightarrow \mathcal{Z} \tag{5.14}$$

$$query : \mathcal{T} \rightarrow \mathcal{Q} \tag{5.15}$$

Setzt man diese ineinander ein, ergibt sich $higherorder\ \ topic(query(\ _i))\quad _i$.

Setzt man dies in die GPI-Gleichung 5.7 anstelle von *effectivityweight* und *effectivitytype*, erhält man einen neuen GPI, der auf Suchstrings und damit auf den User-Absichten basiert.

$$ibGPI_s = \sum_{i=1,\ _i \in s}^{|s|} higherorder\ \ topic(query(\ _i))\quad _i \qquad \textit{efficiency}_i \tag{5.16}$$

Der Intention Based GPI (ibGPI) bewertet, wie der GPI, jede Transition (Click) eines Users einzeln. Er läßt sich, wie in Gleichung 5.16 dargestellt, für eine Session aufsummieren. Die

Bewertung von Sessions und Webpages ist damit von der Anzahl der Clicks abhängig. Positive und negative Clickbewertungen können sich gegenseitig aufheben. Zum einen soll der ibGPI kumulativ Sessions und Webpages bewerten, zum anderen soll eine erfolgreiche Session als solche kenntlich gemacht werden. Daher wird ein separates Erfolgsmaß zur Bewertung des Erfolges einer Session als Ganzes konstruiert.

Session Success Indicator - Zielerreichungsmaß

Als zusätzliches Maß zum GPI und ibGPI wird ein Zielerreichungsmaß berechnet, das die Annäherung des Users an die inhaltlich optimale Webpage misst. So wird der Erfolg einer Session aus Sicht des Users bzw. dessen in den Suchstrings enthaltenen Intentionen gemessen, ohne daß die Zielerreichung durch Kumulierung von negativen und positiven ibGPI Werten verschleiert wird.

Zu diesem Zweck soll der **S**ession **S**uccess **I**ndicator (SSI) berechnet werden. Der SSI ist der höchste Rang einer Webpage, der in einer Session erreicht wurde. Umgeformt auf eine Skala von $[0..1]$, zeigt er die maximale Annäherung an die optimale Webpage an, die dem initialen Ziel des Users am besten entspricht.

Den Rang einer Webpage erhält man durch die Funktion $rank$, siehe 5.10. Die Funktion $rank$ benötigt als Eingabewerte eine Suchanfrage $q \in Q$ und eine Webpage $p \in \mathcal{P}$. Betrachtet man eine Transition $= (p_{source}, p_{target})$ zwischen zwei Webpages $p_{source}, p_{target} \in \mathcal{P}$, dann sei die p_{target} als Zielseite einer Transition durch festgelegt. Der SSI wird durch die Funktion ϱ_{SSI} wie folgt berechnet:

$$\varrho_{SSI} \; = \; max\{rank(query(\quad),\ )\,.\qquad \in s\} \tag{5.17}$$

In Worten: Der SSI berechnet sich aus dem höchsten Rang der Webpages einer Session. Durch die Funktion $query$ erhält man die zu einer Transition gehörige Suchanfrage q. Mit ihr läßt sich der Rang einer Webpage einer Transition berechnen. Aus allen $\in s$ ermittelt ϱ_{SSI} den maximal erreichten Rang in einer Session mit Werten $0 \qquad \varrho_{SSI} < 1$.

Mit dem SSI wird eine Session als Ganzes bewertet. Berechnet man den SSI schrittweise für eine Session für jede Transition dieser Session aufs neue, wird die maximale Annäherung berechnet, die in einer Session bis zu der betrachteten Transition bisher erreicht wurde. Im Gegensatz zum ibGPI verharrt der SSI auf dem höchsten Rang, der erreicht wurde. Der ibGPI würde bei Entfernen von der optimalen Website wieder geringere Werte liefern.

Der SSI mißt wie der ibGPI nur das durch die Suchanfragen bekannte initiale User-Interesse. Die Erfüllung dieses initialen User-Interesses wird nicht durch die Verfolgung von parallelen, unbekannten Interessen behindert. Mit dieser Annahme nimmt man folgende Fehler in Kauf:

- Ein User hat sein initiales Ziel nicht erreicht, aber andere Ziele → die ganze Session wird dennoch negativ bewertet

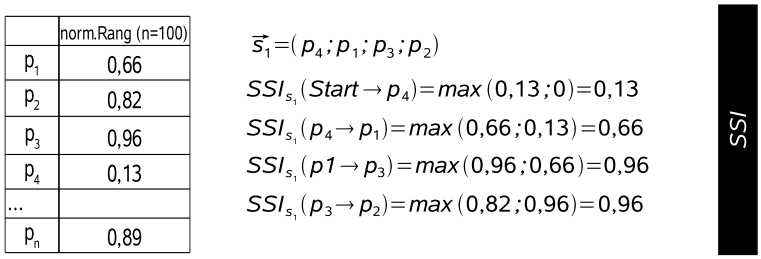


Abbildung 5.20: Funktion SSI

- Ein User hat sein initiales Ziel erreicht, aber ein oder mehrere andere Ziele nicht → die ganze Session wird dennoch positiv bewertet

Eine beispielhafte Berechnung des SSI zeigt Abb. 5.20.

Der SSI mißt die maximale inhaltliche Annäherung eines User an dessen Ziel, das durch den Suchquery repräsentiert wird. Der SSI ist ein Maß für die komplette Session, unabhängig von der Anzahl der Transitionen innerhalb einer Session, wohingegen der ibGPI jede Transition einer Session einzeln bewertet und für die Gesamtbewertung einer Session erst aggregiert werden muß.

5.4.4 Fallstudie

Der ibGPI und der SSI wurden erstellt, um die Defizite zu beseitigen, die in der User-Studie festgestellt wurden. Daher werden beide Maße auf die gleichen Daten wie der GPI angewandt. Die in den Tabellen 5.11 und 5.12 dargestellten Sessions, stammen aus der User-Studie aus Kap. 5.3 und wurden zum einen auf Website A zum anderen auf B ausgeführt.

Die Sessions in der User-Studie verfolgen die vorgegebenen Aufgaben, die einem initialen User-Interesse entsprechen. In der User-Studie liegen jedoch keine Suchstrings vor. Um diese Sessions dennoch für die ibGPI-Auswertung nutzbar zu machen, werden die Sessions mit künstlich hinzugefügten Suchstrings ausgestattet, die der Aufgabenstellung aus der User-Studie entsprechen. Die Sessions der User-Studie werden so erweitert, als ob die befragten User zuvor eine Suchmaschine benutzt hätten, in der sie nach der Aufgabenstellung der User-Studie gesucht hätten. Dies entspricht dem Vorgehen bei der Kennzahl *Task Completion Rate* bzw. *User Focus*, die in Kapitel 3.3.1 erläutert wurden.

Es wurden zwei Sessions ausgewählt, deren Bewertung durch den klassischen GPI von der User-Bewertung in beide Richtungen abweicht. Der User aus Session 1 in Tab. 5.11 hat eine positive Bewertung seiner Session abgegeben, wohingegen der klassische GPI ein sehr

Seq.No.	Webpage	GPI	*efficiency*	Rang	$effectivity_{ibGPI}$	ibGPI	SSI
1	Home	-1,2	0,8	23	0,63	0,50	0,63
2	Search	-1,2	0,8	64	-0,63	-0,50	0,63
3	Search	-1,2	0,8	64	0	0	0,63
4	adv. Search	0,75	0,5	56	0,12	0,01	0,63
5	1171	0,1	1	4	0,81	0,41	0,93
Total		-2,75				0.42	0.93

Tabelle 5.11: Fallstudie ibGPI: Session 2 auf Website B

negatives Bild dieser Session liefert. Die Webpage 1171 war die Zielseite zur Frage in der User-Studie. Die Session startet auf der Homepage, die für die Suchanfrage dieser Session mit Rang 23 versehen wurde. Insgesamt umfaßt die Website 64 Webpages, sodaß die erste Transition in der Funktion $effectivity_{ibGPI}$ einen Wert von 0,63 erzielt. Mit dem Effizienzmaß *efficiency* von 0,8 ergibt sich ein ibGPI-Wert von 0,50. Der SSI startet mit dem gleichen $effectivity_{ibGPI}$-Wert von 0,63. Die Gesamtbewertung der Session durch den GPI ergibt einen Wert von -2,75, beim ibGPI von 0,42. Die durchschnittlichen Bewertungen je Transition beim GPI -0,55, beim ibGPI von 0,08. Der User hat die Session positiv bewertet. Der ibGPI-Wert liegt näher am User-Urteil. Außerdem zeigt der SSI-Wert von 0,93 an, daß der User auf Webpages war, die seiner Suchanfrage sehr gut entsprochen haben. Betrachtet man den Weg des User durch die Seite, beschreibt der ibGPI und der SSI den Weg des Users sehr genau: Session Start auf einer gut passenden Webpage, die Entfernung vom Ziel über die Suchseiten und die Wiederannäherung an die Zielseite. Sowohl der ibGPI als auch der SSI spiegeln das User-Verhalten und dessen Einschätzung der Session wesentlich besser wider als der klassische GPI.

Tabelle 5.11 zeigt ebenfalls eine Session aus der User-Studie, diesmal auf Website A. Auch hier weicht das User-Urteil von der Bewertung durch den klassischen GPI ab. Der User war sehr unzufrieden mit seiner Session auf dieser Website. Website A umfaßt 357 Webpages.

Der User startet die Session auf der Homepage, die für die Suchanfrage dieser Session den nahezu höchsten Rang erhalten hat. Daher bleibt der SSI mit 0,99 über die gesamte Session konstant. Der klassische GPI bewertet die Session sowohl in der Summe mit 7,83 als auch je Transition mit 0,98 sehr positiv. Der ibGPI bewertet die Session in der Summe mit 1,23 und je Transition mit 0,15 nur schwach positiv. Der SSI ist allerdings sehr hoch. Der niedrige ibGPI und der dagegen sehr positive SSI beschreiben die ambivalente Sichtweise des Users auf

Seq.No	Web-page	GPI	*efficiency*	Rang	$effectivity_{ibGPI}$	ibGPI	SSI
1	Home	2,25	1,5	3	0,99	1,49	0,99
2	2578	0,91	0,8	62	-0,16	-0,13	0,99
3	1067	0,27	0,4	58	0,01	0,002	0,99
4	1067	1,25	1,2	4	0,15	0,18	0,99
5	1042	0,25	0,4	55	-0,14	-0,06	0,99
6	1302	1,54	1,2	39	0,04	0,05	0,99
7	2579	1,31	1,2	65	-0,03	-0,04	0,99
8	1053	0,06	0,4	303	-0,67	-0,26	0,99
Total		7,83				1,23	0,99

Tabelle 5.12: Fallstudie ibGPI: Session 1 auf Website B

seine Session sehr gut. Der User hat bereits mit dem ersten Click eine gute Webpage erreicht und versucht durch weiteres Navigieren auf der Website vergeblich, genauere Informationen zu finden. Der User hat durch die vierte Transition nochmals eine Zielseite mit der gesuchten Information erreicht, allerdings ist er nur kurz dort geblieben, was in einem geringen *efficiency* Wert resultiert.

Auch in diesem Fall spiegeln der ibGPI und der SSI das User-Verhalten realistischer wider als der klassische GPI. Dadurch kann der Website-Autor die richtigen Schlußfolgerungen ziehen, zum Beispiel im letzten Fall die Navigation verbessern.

Da die User-Studie unter Laborbedingungen stattgefunden hat, kann von einem einzigen User-Ziel ausgegangen werden. Würde dieser Clickstream unter normalen Bedingungen beobachtet werden, könnte man nicht unterscheiden, ob der User **1.** sein Ziel mit dem ersten Click erreicht hat und sich für weitere Themen hat interessieren lassen und zu einer positiven Beurteilung der Session gelangt, oder **2.** das initiale Ziel wie in der User-Studie noch weiterverfolgt und eher zu einer schlechten Beurteilung der Sessions gelangt. Auf diese eingeschränkte Interpretierbarkeit wurde in Kapitel 5.4.3 hingewiesen.

5.5 Zusammenfassung

5.5.1 Guidance Performance Indicator

Der GPI stellt den Ausgangspunkt für die Bewertung von informationsorientierten Websites dar. Es wurde gezeigt, welche Informationen auf dieser Art von Websites zur Verfügung stehen. In einem formalen Modell wurden die Zusammenhänge zwischen diesen Informationen dargestellt, sodaß daraus ein Bewertungsmodell abgeleitet werden konnte. Dieses Bewertungsmodell ist der GPI, der aus Sicht der Website die User-Aktionen nach deren Beitrag zum Erfolg der Website bewertet.

Aus den Übergängen zwischen Webpage-Typen wird bewertet, inwieweit ein bestimmter Typ zum Ziel einer Website beitragen kann. Aus dieser Überlegung heraus wird das Erreichen von Contentseiten als wünschenswert angesehen, da hier die Inhalte von informationsorientierten Websites vermittelt werden. Navigationspages dienen dagegen zur Strukturierung der Website und zur gezielten Hinleitung auf Contentseiten.

Neben der Unterscheidung von Webpage-Typen wird jeder Click in den inhaltlichen Zusammenhang der zugehörigen Session gestellt. Auf diese Weise lassen sich in Bezug auf den individuellen Sessioninhalt inhaltliche Ausreißer erkennen und bewerten.

Die Effizienz eines Clicks wird durch die Aufenthaltsdauer auf einer Webpage erfaßt.

Aus Effektivitätsmaßen und Effizienzmaß wird der GPI für jeden Click berechnet und kann durch entsprechende Aggregation der GPI-Werte Webpages oder User Sessions bewerten.

Der GPI bewertet eine Website und deren Benutzung rein aus Sicht der Website bzw. deren Autoren. Die Sessions und Clicks werden dahingehend bewertet, ob sie zu den Zielen der informationsorientierten Website beitragen, indem Informationen ausreichend lange wahrgenommen werden konnten.

Wegen mangelnder Informationen auf der User-Seite, wie im Website Modell dargestellt, können die Ziele der User nicht berücksichtigt werden. Dieser Mangel wurde in der User-Studie deutlich, die zur Evaluation des GPI durchgeführt wurde.

Durch die Studie wurde bestätigt, daß der GPI in der Lage ist, User-Verhalten richtig zu bewerten, sofern es den Erwartungen der Website entspricht. Weichen User von diesen erwarteten Verhaltensmustern ab, kann der GPI dies nicht immer richtig darstellen.

Diesen Mangel behebt der Intention Based GPI.

5.5.2 Intention Based Guidance Performance Indicator

Durch Einbeziehung der Referrer Information und der darin enthaltenen Suchstrings als zusätzliche Informationsquelle, kann auf das User-Interesse geschlossen werden. Das formale Website-Modell wurde um die Suchstrings erweitert. Aus den damit ersichtlichen Zusammenhängen des Inhalts von Suchanfrage und Inhalt der Webpages wurde der ibGPI abgeleitet.

Der ibGPI erstellt für jede Suchanfrage eine individuelle Bewertungsfunktion, die die inhaltliche Übereinstimmung der besuchten Webpages mit den in der Suchanfrage repräsentierten User-Zielen vergleicht und bewertet. Hierzu wurden aus dem Website-Modell mehrere Funktionen abgeleitet, die eine lückenlose Beschreibung des neuen Effektivitätsmaßes erlauben.

Der ibGPI ist im Vergleich zum GPI wesentlich vorteilhafter, indem er manuelle Eingriffe und Einschätzungen bei der Festlegung von Gewichten des Effektivitätsmaßes überflüssig macht.

Der ibGPI verzichtet bei der Berechnung des Effektivitätsmaßes nicht nur gänzlich auf manuelle Eingriffe, sondern erstellt die Bewertungsfunktion durch eine Funktion höherer Ordnung $higherorder$ individuell für jede Suchanfrage. Der ibGPI ist dadurch wesentlich robuster und kann genauer berechnet werden. Die Höhe unterschiedlicher ibGPI-Werte ist im Gegensatz zum GPI direkt interpretierbar und kann besser nachvollzogen werden. Auch über mehrere Bewertungszeiträume hinweg sind die ibGPI-Werte vergleichbarer als die des GPI.

Durch den GPI und den ibGPI kann eine Website und ihre Benutzung aus zwei Perspektiven beurteilt werden, wie dies Hahn et al. in [113, 112]in Abb. 4.11 und Riemer et al. in [201] in Abb. 1.8 vorgeschlagen haben. Die Anwendbarkeit beider Maße wurde in Fallstudien und einer User-Befragung überprüft und validiert.

5.5.3 Anwendungsmöglichkeiten

Fortlaufende Website-Optimierung

Die direkte Anwendung einer Erfolgsmessung liegt in der Ableitung eines Verbesserungspotentials und dem Erkennen von Schwachstellen auf der Website. Dazu dienen der GPI bzw. ibGPI, die nicht nur Sessions auf ihren Erfolg hin bewerten, sondern auch einzelne Webpages. Wie in der Fallstudie in 5.2 gezeigt, können Webpages miteinander verglichen werden und problembehaftete Webpages erkannt werden. Beide Erfolgsmaße werden aus Daten erstellt, die durch ein User Tracking-System gesammelt werden können. Lediglich der Crawling-Vorgang zur Sammlung der Content- und Structure-Daten und die Datenintegration bedarf eines größeren zeitlichen Aufwands. In welchen zeitlichen Abständen der Website Inhalt erfaßt werden soll und kann, muß durch Erfahrungswerte festgelegt werden.

Die Suchanfragen können auch zu einer Anpassung der Website-Inhalte verwendet werden. Extrahiert man die Suchstrings, die nicht auf den Webpages vorkommen, könnte man daraus auf das Fehlen von Inhalten schließen, die vom User auf der Website erwartet werden.

Website-Modell

Die Darstellung der Zusammenhänge einer Website und deren Benutzung hat zur Berechnung des ibGPI sehr stark beigetragen. Dadurch konnte beispielsweise die higherorder Funktion abgeleitet werden. In Barth und Stolz et al. [15] wurde ebenfalls der Nutzen eines Website-Modells durch die formale Herleitung des Effektivitätstyps gezeigt.Dieses Modell ist noch keineswegs vollständig. Es kann aber durch konsequente Erweiterung den von Schwickert et al. [213, S.3] beschriebenen Mangel an mathematischem Verständnis über den Zusammenhang zwischen Web Kennzahlen beseitigen helfen. Auf diesem Wege können nicht nur bestehende Kennzahlen validiert, sondern auch neue Kennzahlen abgeleitet werden.

Bewertung und Analyse von Suchmaschinenresultaten

Der oben vorgestellte Ansatz zur Analyse von Suchmaschinenanfragen aus Referrer-Informationen eignet sich auch zur Analyse der Qualität der Suchmaschinenergebnisse. Dabei können sowohl Website externe als auch interne Suchmaschinen analysiert werden.

Mit den Berechnungen des ibGPI wurde für jede Suchanfrage von externen Suchmaschinen ein Ranking erstellt, das den Webpages der analysierten Website nach deren inhaltlicher Übereinstimmung mit den Suchstrings einen Rang zugewiesen hat. So wird eine alternative Ergebnisliste für eine Suchanfrage erstellt. Vergleicht man die gefundenen Webpages durch eine Suchmaschine mit dem Ranking durch den ibGPI, können Webpages identifiziert werden, die inhaltlich besser zu den Suchanfragen der User passen. Da die Ranking-Algorithmen von Suchmaschinen meist nicht frei zugänglich sind, bietet das ibGPI-Ranking einen Vergleichsmaßstab.

Die Resultate einer Suchmaschinenanalyse können bei internen Suchmaschinen direkt einfließen, da hier die Einwirkungsmöglichkeiten größer sind als auf externe Suchmaschinen wie Google. Zur Qualitätssteigerung stehen nicht nur die Suchanfrage und die Suchergebnisse sondern auch die komplette Session eines Users zur weiteren Analyse zur Verfügung. Die Zufriedenheit eines Users mit den Suchmaschinenergebnissen kann durch eine Analyse von dessen kompletter Session auf der analysierten Website wesentlich genauer ermittelt werden. So kann bei der Bewertung der Suchmaschinenergebnisse berücksichtigt werden, welche weiteren Webpages der User aufgesucht und wie lange er dort verweilt hat.

Abb. 5.21 zeigt das identifizierte Verbesserungspotential auf Webpage A. Für alle Suchanfragen an eine externe Suchmaschine wurde das Ranking für den ibGPI durchgeführt. Dadurch

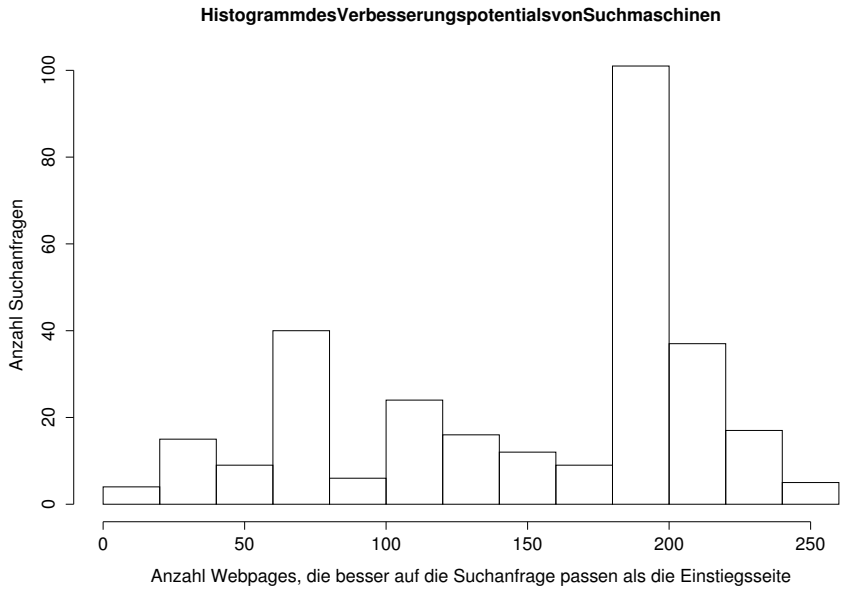


Abbildung 5.21: Suchmaschinen-Verbesserungspotential (Webpage A, Jan. 06)

wurden die Webpages identifiziert, die am geeignetsten für die jeweilige Suchanfrage sind. Die erste Webpage einer Session, also die Einstiegsseite, wurde von der externen Suchmaschine empfohlen. Betrachtet man den Rang dieser Einstiegsseite, ergibt sich daraus das Verbesserungspotential, das aus allen höherrangigen Webpages besteht. Die Anzahl an höherrangigen Webpages für jede Suchanfrage ist in Abb. 5.21 als Histogramm gezeigt.

Gegen die Anwendung auf externe Suchmaschinen spricht zum einen die geringe Einflußmöglichkeit. Zum anderen kann die Website-Politik verlangen, daß jeder User die Website vorzugsweise durch eine bestimmte Seite, wie die Homepage, betritt. Diese Einwände gelten nicht für die Anwendung auf Website interne Suchmaschinen.

Personalisierung

Neben der Integration in ein Bewertungsmodell zur Verbesserung statischer Websites, kann das vorgestellte Modell zur Auswertung von Suchanfragen in Referrern dazu genutzt werden, einem User gezielt interessante Inhalte anzubieten. Eine derartige Personalisierung benötigt keinerlei Kundenhistorie, um direkt auf die Kundenwünsche reagieren zu können.

Betritt ein User die analysierte Website ausgehend von einer Suchmaschine, kann in den meisten Fällen der Suchstring aus den Referrer-Informationen extrahiert werden. Mit der oben gezeigten Analyse kann darauf direkt reagiert werden. Da das Webpage Ranking individuell für

jede Suchanfrage ist, können auch individuelle Webpages empfohlen oder als Link einem User angeboten werden.

Der Vorteil dieser Personalisierungstechnik liegt darin, daß keine vorherige Analyse anderer User oder eine Kundenhistorie notwendig sind, um Empfehlungen für einen User ausgeben zu können. Sie ist allerdings auf User beschränkt, die interne oder externe Suchmaschinen benutzt haben.

User Profiling

Wie durch die User-Studie deutlich wurde, unterscheiden sich User sehr stark voneinander. Insbesondere deren Browsingverhalten unterscheidet sich stark. So ist eine "Generation Google" daran gewöhnt, eine Suchmaschine zu benutzen, um gezielt an Informationen zu gelangen. Wenn man diese User-Strategien kennt und bei der Gestaltung einer Website berücksichtigen kann, erhöht sich deren Usability erheblich. Solche Arbeiten wurden bereits unter dem Stichwort User Profiling durchgeführt. Zum einen können mit Hilfe der Suchstrings diese User-Profile wesentlich detaillierter werden. Zum anderen werden die Sessions durch GPI und ibGPI bewertet, wodurch ein weiteres Profilierungskriterium verfügbar wird.

Um eine weitere Verbesserung der Erfolgsmessung einer Website zu erreichen, wäre eine unterschiedliche Betrachtung von User-Typen sinnvoll. Betrachtet man die Session aus der User-Studie in Tab. 5.11, kann daraus ein User-Profil erstellt werden. Durch eine weitere Studie müßten zunächst User-Profile erkannt und deren Charakteristika identifiziert werden. So können Sessions diesen User-Profilen zugeordnet und individuell bewertet werden.

Beispielsweise könnte man für ein User-Profil einer Session wie in Tab. 5.11 mehrere Ziele unterstellen. Dadurch käme man zu einer differenzierteren Analyse der Website und ihrer User.[1]

[1] Um die in dieser Arbeit durchgeführten Analysen nachvollziehen zu können, sind User-Daten einer Website in hoher Qualiät notwendig. Die Analysen wurden mit der frei zugänglichen Statistiksoftware *R (http://www.r-project.org/)* [198] durchgeführt. Der PLSA-Algorithmus wurde von dem Autor dieser Arbeit in *R* selbst umgesetzt, ist aber noch nicht im Internet verfügbar. Die Sammlung der inhaltlichen Daten erfolgte mit einem Crawler, der sich im Eigentum der Siemens AG befindet.

6 Zusammenfassung und Ausblick

Dieses Kapitel gibt einen abschließenden Überblick der Herangehensweise an die Aufgabenstellung dieser Arbeit und eine Zusammenfassung der erzielten Ergebnisse. Die sich daraus ergebenden Erkenntnisse werden vor dem Hintergrund der Aufgabenstellung und ihre Anwendbarkeit diskutiert. Über diese Arbeit hinaus wird ein Ausblick auf Ansätze gegeben, die im Anschluß an diese Arbeit weiterverfolgt werden sollen.

6.1 Überblick und Diskussion der Ergebnisse

Websites haben sich für Unternehmen als fester Bestandteil ihrer Geschäftstätigkeit etabliert. Der Umfang und die Art der Internetaktivitäten wurde durch die dahinterliegenden Geschäfts- und Erlösmodelle beschrieben. Die damit verbundenen Unternehmensziele und deren Umsetzung wurden in Kapitel 1.1.4 vorgestellt. Die zunehmende Bedeutung dieses Kommunikations- und Vertriebskanals verlangt eine kontinuierliche Analyse und Kontrolle der Website. Mit steigenden Investitionen erlangt die Frage nach deren Rentabilität einen hohen Stellenwert.

Aus der Beschreibung der Erlösmodelle in 1.1.5 ließ sich diese Frage teilweise beantworten. Bei einer Unterscheidung zwischen transaktions- und informationsorientierten Websites wurde deutlich, daß lediglich die Erfolgsmessung bei transaktionsorientierten E-Commerce Websites, die Produkte oder Dienstleistungen anbieten, ausreichend erforscht ist und angewandt wird.

Der Erfolg von Websites, deren Geschäftsmodell im Angebot von Unternehmensinformationen liegt, ist dagegen kaum untersucht. Der in Kapitel 1.2 beschriebene Mangel an Kennzahlen und Methoden zur Untersuchung des Erfolgs von informationsorientierten Websites stellt den Ausgangspunkt dieser Dissertation dar.

Um die Möglichkeiten und Methoden abschätzen zu können, mit denen an diese Aufgabenstellung herangegangen werden kann, wurden in Kapitel 2 zunächst die verfügbaren Daten beschrieben. Dabei wurde ein besonderes Augenmerk auf die Datenqualität gelegt, die durch die beschriebenen Datensammlungs-, Datenaufbereitungs- und Datenbereinigungstechniken erreicht werden konnte.

Die Zahl der verfügbaren Daten hängt von deren Erhebung ab. Durch explizite Befragung der User einer Website können fast alle Aspekte zur Beurteilung der Website aus Usersicht erfaßt werden. Diese kosten- und zeitintensive Erhebungsform eignet sich jedoch nicht zur dauerhaften Analyse von Websites und ist zudem nur auf einen Teil der User beschränkt. Durch Beobachtung der User aus Sicht der Website können alle User erfaßt werden, aber der Umfang der verfügbaren Informationen ist wesentlich eingeschränkter. Trotzdem beschränkt sich diese Arbeit auf die serverseitig verfügbaren Daten.

Ausgehend von dieser Basis wurden in Kapitel 3 und 4 vorhandene Kennzahlen und Methoden untersucht, die geeignet erscheinen, den Erfolg auf informationsorientierten Websites meßbar machen zu können. Kennzahlen und Web Metriken wurden in Kapitel 3 erläutert und zu einer Kennzahlen Systematik zusammengefaßt. Aus dieser Übersicht wurde das Fehlen geeigneter Maße erneut deutlich. Einzelne Kennzahlen sind im Gegensatz zu transaktionsorientierten Websites nicht in der Lage, den komplexen Prozeß der Informationsvermittlung wiederzugeben. Lediglich Ansätze zur Beurteilung von Teilaspekten des Erfolgs informationsorientierter Websites sind vorhanden und konnten als Grundlage des hier vorgestellten Ansatzes dienen.

Neben der direkten Messung durch Kennzahlen versprechen Web Mining-Analysen genaue Einblicke in die Wirkungszusammenhänge einer Website und die Beweggründe der User für ihr Verhalten. Kapitel 4 erläuterte Web Mining-Methoden, mit deren Hilfe die Benutzung, der Inhalt und die Struktur einer Website untersucht werden können. Um den komplexen Prozeß der erfolgreichen Informationsvermittlung richtig abbilden zu können, müssen die Aktionen der User im inhaltlichen und strukturellen Zusammenhang der Website analysiert werden. Insbesondere die in Kapitel 4.3 vorgestellten Textanalysen waren für die Erstellung eines Erfolgsmaßes hilfreich.

Aus den Erkenntnissen der Kapitel 2, 3 und 4 wurde in Kapitel 5 ein neuer Ansatz zur Erstellung eines Erfolgsmaßes für informationsorientierte Websites vorgestellt.

In Kapitel 5.1 wurden einzelne Useraktionen im inhaltlichen Zusammenhang der jeweiligen User Session dahingehend untersucht und bewertet, ob sie dem Ziel einer Website entsprechen. Das Ziel einer informationsorientierten Website wird in der Vermittlung von Informationen gesehen. Da die User-Interessen unbekannt bleiben, kann nur aus dem inhaltlichen Zusammenhang einer Session zwischen Webpages unterschieden werden, die innerhalb oder außerhalb des inhaltlichen User-Fokuses liegen. Zusätzlich werden die Webpages Kategorien zugeordnet, um aus der Transition zwischen diesen Kategorien ein weiteres Bewertungskriterium zu erhalten. Besuche auf Webpages, die Informationen vermitteln, werden anders bewertet als Besuche auf solchen Webpages, die zur Strukturierung der Website dienen. Zur Messung der Effizienz einer User-Aktion wird die Duration herangezogen. Aus diesen Teilbewertungen wird als Gesamtmaß der Guidance Performance Indicator berechnet.

Zur Veranschaulichung der Zusammenhänge innerhalb einer Website und deren Benutzung wurde in Kapitel 5.1.6 ein Website-Modell erstellt. Dieses Modell dient neben der Veranschau-

lichung außerdem zur klaren Trennung von Fakten und Annahmen, die bei der Erstellung des GPI getroffen werden.

Der GPI ist als größtenteils regelbasierter Ansatz auf viele manuelle Eingriffe und die Festlegung von Parametern angewiesen. Der GPI wurde in Kapitel 5.2 auf zwei Websites angewandt. Darin zeigt sich dessen Fähigkeit, Schwachstellen auf einer Website zu entdecken und die User-Aktionen aus Sicht der Website zu beurteilen.

Als zusätzliche Evaluationsmöglichkeit wurde in einer User-Studie in Kapitel 5.3 überprüft, ob der GPI auch fähig wäre, die User-Sicht abzubilden. Dies war in einigen Fällen nicht möglich, in denen die User nicht den von der Website erwarteten User-Typen entsprachen.

Um diesen Mangel zu beheben, wurden in Kapitel 5.4 Referrer-Informationen als zusätzliche Datenquelle eingeführt. Die darin teilweise enthaltenen Suchanfragen referenzierender Suchmaschinen erlauben Rückschlüsse auf die Interessen der User. Somit wurde es möglich, neben der Website-Sicht des GPI die User-Sicht einfließen zu lassen. Durch die Erweiterung des Website-Modells in 5.4.2 konnte eine neue Erfolgsbewertungsfunktion in Kapitel 5.4.3 formal hergeleitet werden.

Die Bewertung der Effektivität einer User-Aktion wird nun nicht mehr wie beim GPI durch Regeln und festgelegte Gewichtungen erreicht, sondern durch eine Funktion berechnet. Diese Funktion ihrerseits wird durch eine Funktion höherer Ordnung festgelegt, die für jede Session eine individuelle Bewertungsfunktion erstellt. Der ibGPI kann nicht nur die vom GPI korrekt beschriebenen Sessions ebenso genau wiedergeben, sondern auch die Fälle aus der User-Studie. Weitere Vorteile des ibGPI liegen in dessen Berechnung als Funktion begründet. Die Höhe der ibGPI-Werte sind direkt interpretierbar und zwischen verschiedenen Sessions und Websites vergleichbar. Die GPI-Werte lassen Website übergreifende Vergleiche kaum zu.

Der GPI und ibGPI ergänzen sich zu einer Gesamtbetrachtung des Erfolgs einer Website, da der GPI den Erfolg rein aus der Website-Sicht betrachtet und der ibGPI die User-Sicht miteinschließt. Der GPI kann alle User Sessions bewerten, wohingegen der ibGPI auf die Verfügbarkeit von Referrer-Informationen angewiesen ist. Durch die parallele Anwendung beider Maße kann ein komplettes Bild des Erfolgs einer informationsorientierten Website geschaffen werden, wie dies in Abbildung 1.8 gefordert wurde.

6.2 Ausblick

Der so berechnete Intention Based GPI (ibGPI) ist ein generisches Erfolgsmaß, das session-individuell berechnet wird. Die sich daraus ergebenden Anwendungsmöglichkeiten werden in Kapitel 5.5.3 beschrieben. Beispielsweise durch die Identifikation von User-Profilen können der GPI und ibGPI weiter verfeinert werden. Dabei ist auf das formale Website-Modell

hinzuweisen. Durch konsequente Erweiterung dieses Modells und die Formulierung von mathematischen Zusammenhängen können die bestehenden Maße und Kennzahlen integriert und verbessert werden. In mehreren Veröffentlichungen [15, 238, 239, 242, 254] wurden bereits Teilaspekte und Anwendungsfälle des GPI, ibGPI und des formalen Modells untersucht und vorgestellt.

In Kapitel 2.2.3 wurde bereits auf die Einschränkung hingewiesen, die sich durch neue Browsertechnologien auf Annahmen über das User-Verhalten ergeben. In [254] wird von Stolz et al. erstmals auf die Problematik hingewiesen, die sich daraus für Web Analysen ergeben. Darin wurde auch ein Lösungsansatz vorgestellt, der aber noch weiterer Forschung bedarf. Will man dieses User-Verhalten korrekt darstellen, muß nicht nur der GPI und ibGPI angepaßt werden, sondern auch die anderen Kennzahlen erneut betrachtet werden.

Ehrenwörtliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Dissertation selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.
Die Arbeit wurde noch keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

München, den 15. April 2007

Lebenslauf

Carsten Dirk Stolz carsten.stolz@ku-eichstaett.de
geboren am 15.06.1975 in Ludwigsburg, Deutschland

Akademische Ausbildung

09.2002 - 10.2007	Dissertation bei Prof. Dr. K.D. Wilde als Doktorand der Siemens AG, München
10.1997 - 03.2002	Studium der Betriebswirtschaft an der Wirtschaftswissenschaftlichen Fakultät Ingolstadt der Kath. Universität Eichstätt-Ingolstadt, Abschluß: Diplom-Kaufmann

Berufliche Ausbildung

08.1996 - 06.1997	Techn. Dienst Ev. Diakonissenkrankenhaus Stuttgart
08.1994 - 07.1996	Ausbildung zum Bankkaufmann (Finanzassistent), Dresdner Bank AG in Stuttgart

Schulische Ausbildung

09.1985 - 06.1994	Leibniz - Gymnasium in Stuttgart, Abitur
09.1981 - 07.1985	Grundschule Hattenbühl in Stuttgart

Berufliche Tätigkeiten

seit 12.2006	b.telligent, München, Consultant für Business Intelligence, Data Warehouse und Data Mining
03.2006 - 10.2006	Universität Eichstätt-Ingolstadt, Lehrstuhl für Wirtschaftsinformatik, Wissensch. Mitarbeiter
09.2002 - 12.2005	Siemens AG, Corporate Technology, Neural Computation, seit 10.2003 CT IC 1 Knowledge Management
08.2000 - 09.2000	Investmentbank Dresdner Kleinwort Benson Singapur, Praktikum bei Global Finance
03.2000 - 04.2000	DaimlerChrysler AG in Stuttgart, Werkstudent im Bereich Finance Administration
08.1999 - 09.1999	DaimlerChrysler AG in Stuttgart, Praktikum im Bereich Finance Administration
03.1999 - 04.1999	Dresdner Bank AG in Stuttgart, Außenhandelsfinanzierung
11.1998 - 02.1999	Lehrstuhl für Wirtschaftsinformatik, Studentische Hilfskraft
08.1998 - 09.1998	Dresdner Bank AG in Stuttgart, Firmenkunden-Kreditabteilung
06.1997 - 08.1997	Dresdner Bank AG in Böblingen, Finanzberater

Literaturverzeichnis

[1] ABBOTT, M. ; CHIANG, K.P. ; HWANG, Y.S. ; PAQUIN, J. ; ZWICK, D.: The Process of Online Store Loyality Formation. In: *Advance in Consumer Research* 27 (2000), S. 145–150

[2] AGARWAL, R. ; VENKATESH, V.: Assessing a firm's Web presence: A heuristic evaluation procedure for the measurment of usability. In: *Information Systems Research* 13 (2002), S. 168–186

[3] AGGARWAL, C.C. ; AL-GARAWI, F. ; YU, P.S.: Intelligent crawling on the World Wide Web with arbitrary predicates. In: *WWW '01: Proc. 10th Int. Conf. on World Wide Web*. New York, NY, USA : ACM Press, 2001, S. 96–105

[4] AGRAWAL, R. ; IMIELINSKI, T. ; SWAMI, A.N.: Mining Association Rules between Sets of Items in Large Databases. In: BUNEMAN, P. (Hrsg.) ; JAJODIA, S. (Hrsg.): *Proc. ACM SIGMOD Int. Conf. on Management of Data 1993*, 207–216

[5] AGRAWAL, R. ; MANNILA, H. ; SRIKANT, R. ; TOIVONEN, H. ; VERKAMO, A.I.: Fast Discovery of Association Rules. In: *Advances in Knowledge Discovery and Data Mining*, AAAI Press/The MIT Press, 1996, S. 307–328

[6] AGRAWAL, R. ; SRIKANT, R.: Fast Algorithms for Mining Association Rules. In: *Proc. 20th VLDB Conference Santiago, Chile*, 1994

[7] AGRAWAL, R. ; SRIKANT, R.: Mining sequential patterns. In: YU, P. (Hrsg.) ; CHEN, A. S. P. (Hrsg.): *11th Int. Conf. on Data Engineering*. IEEE Computer Society Press, 3–14

[8] ALADWANI, A.M. ; PALVIA, P.C.: Developing and validating an instrument for measuring user-perceived web quality. In: *Information Management* 39 (2002), Nr. 6, S. 467–476

[9] AMITAY, E. ; CARMEL, D. ; DALROW, A. ; LEMPEL, R. ; SOFFER, A.: Topic Distillation with Knowledge Agents. In: *Proc. of the 11th Text Retrieval Conference TREC 2002 NIST Special Publication:SP 500-251*

[10] ANAND, S.S ; MULVENNA, K.: On the Deployment of Web Usage Mining. In: BERENDT, B. (Hrsg.) ; HOTHO, A. (Hrsg.) ; MLADENIC, D. (Hrsg.) ; V.SOMOREN, M. (Hrsg.) ; SPILIOPOULOU, M. (Hrsg.) ; STUMME, G. (Hrsg.): *Web Mining: From Web to Semantic Web*, Springer, 2004 (Lecture Notes in Artificial Intelligence 3209), S. 23–42

[11] ARNDT, D. ; KOCH, D.: Datenschutz im Web Mining - Rechtliche Aspekte des Umgangs mit Nutzerdaten. In: WILDE, K.D. (Hrsg.): *Handbuch Web Mining im Marketing*, Vieweg, September 2002, S. 77–103

[12] BALDI, P. (Hrsg.) ; FRASCONI, P. (Hrsg.) ; SMYTH, P. (Hrsg.): *Modeling the Internet and the Web - Probabillistic Methods and Algorithms*. Wiley, 2003

[13] BARNARD, L. ; WESSON, J. L.: Usability issues for E-commerce in South Africa: an empirical investigation. In: *SAICSIT '03* (2003), S. 258–267

[14] BARTELL, B. T. ; COTTRELL, G. W. ; BELEW, R. K.: Latent semantic indexing is an optimal special case of multidimensional scaling. In: *Proc. of SIGIR '92*. New York, NY, USA : ACM Press, 1992, S. 161–167

[15] BARTH, M. ; SKUBACZ, M. ; STOLZ, C.: Web Performance Indicator by Implicit User Feedback - Application and Formal Approach. In: *LNCS: Intl. Conf. WISE 2005 New York*, Springer, 2005

[16] BAUMGARTEN, M. ; BÜCHNER, A. G. ; ANAND, S. S. ; MULVENNA, M. D. ; HUGHES, J. G.: User-Driven Navigation Pattern Discovery from Internet Data. In: *WEBKDD '99: Revised Papers from the International Workshop on Web Usage Analysis and User Profiling*. London, UK : Springer-Verlag, 2000, S. 74–91

[17] BELEW, R.K. ; VAN RIJSBERGEN, C. J.: *Finding out about: a cognitive perspective on search engine technology and the WWW*. New York, NY, USA : Cambridge University Press, 2000

[18] BENSBERG, F.: Segmentierung im Online-Marketing. In: HIPPNER, H. (Hrsg.) ; MERZENICH, M. (Hrsg.) ; WILDE, K.D. (Hrsg.): *Handbuch Web Mining im Marketing*, vieweg, 2002, S. 163–192

[19] BENSBERG, F.: Website Optimierung - Aufgabenstellung und Vorgehensweise. In: HIPPNER, H. (Hrsg.) ; MERZENICH, M. (Hrsg.) ; WILDE, K.D. (Hrsg.): *Handbuch Web Mining im Marketing*, Vieweg, 2002, S. 248–265

[20] BERENDT, B.: Web usage mining, site semantics, and the support of navigation. In: *WEBKDD 2000 - Worskhop on Web Mining for E-Commerce – Challenges and Opportunities*, 2000

[21] BERENDT, B.: Using Site Semantics to Analyze, Visualize, and Support Navigation. In: *Data Minining and Knowledge Discovery* 6 (2002), Nr. 1, S. 37–59

[22] BERENDT, B. ; HOTHO, A. ; MLADENIC, D. ; V.SOMOREN, M. ; SPILIOPOULOU, M. ; STUMME, G.: A Roadmap for Web Mining: From Web to Semantic Web. In: BERENDT, B. (Hrsg.) ; HOTHO, A. (Hrsg.) ; MLADENIC, D. (Hrsg.) ; V.SOMOREN, M. (Hrsg.) ; SPILIOPOULOU, M. (Hrsg.) ; STUMME, G. (Hrsg.): *Web Mining: From Web to Semantic Web*, Springer, 2004 (Lecture Notes in Artificial Intelligence 3209), S. 1–22

[23] BERENDT, B. ; SPILIOPOULOU, M.: 2.3.1 Assoziations- und Pfadanalyse - Entdeckung von Abhängigkeiten. In: HIPPNER, H. (Hrsg.) ; MERZENICH, M. (Hrsg.) ; WILDE,

K.D. (Hrsg.): *Handbuch Web Mining im Marketing*, Vieweg, September 2002, S. 143–161

[24] BERGMARK, D.: Collection synthesis. In: *Proc. of JCDL '02 ACM/IEEE Conf. on Digital Libraries*. New York, NY, USA : ACM Press, 2002, S. 253–262

[25] BERNERS-LEE, T. ; CAILLIAU, R. ; LUOTONEN, A. ; NIELSEN, H. F. ; SECRET, A.: The World-Wide Web. In: *Commun. ACM* 37 (1994), Nr. 8, S. 76–82

[26] BERNERS-LEE, T. ; FISCHETTI, M. ; BERNERS-LEE, T. (Hrsg.) ; FISCHETTI, M. (Hrsg.): *Weaving the Web*. Harper, 1999

[27] BERRY, L.L.: Relationship Marketing. In: BERRY, L.L. (Hrsg.) ; SHOSTACK, G.L. (Hrsg.) ; UPAH, G.D. (Hrsg.): *Emerging Perspectives on Services Marketing*, 1983, S. 25–28

[28] BERRY, M.J.A. ; LINOFF, G.S. ; ELLIOTT, R.M. (Hrsg.): *Data Mining Techniques*. Wiley Computer Publishing, 1997

[29] BERTHON, P.: The World Wide Web as an advertising medium: toward an understanding of conversion efficiency. In: *Journal of Advertising Reseaerch* 36 (1996), S. 43ff

[30] BHARAT, K. ; HENZINGER, M.R.: Improved algorithms for topic distillation in a hyperlinked environment. In: *Proc. of SIGIR '98*. New York, NY, USA : ACM Press, 1998, S. 104–111

[31] BIRKHOFER, B.: Ertragsmodelle - Einnahme- und Erlösquellen im innovativen Absatzkanal des Electronic Commerce. In: SCHLÖGEL, M. (Hrsg.) ; TOMCZAK, T. (Hrsg.) ; BELZ, C. (Hrsg.): *Roadmap to e-Business - Wie Unternehmen das Internet erfolgreich nutzen*, Thexis Verlag, 2002, S. 430–452

[32] BOLLINGER, T.: Assoziationsregeln - Analyse eines Data Mining Verfahrens. In: *Zeitschrift Informatik Spektrum* 19 (1996), S. 257–261

[33] BORGELT, C. ; KRUSE, R.: Induction of Association Rules: Apriori Information. In: *15th Conference on Computational Statistics 2002*, 2002

[34] BORGES, J. ; LEVENE, M.: Data Mining of User Navigation Patterns. In: *WEBKDD*, 92-111

[35] BOTAFOGO, R. A. ; SHNEIDERMAN, B.: Identifying aggregates in hypertext structures. In: *HYPERTEXT '91: Proceedings of the third annual ACM conference on Hypertext*. New York, NY, USA : ACM Press, 1991, S. 63–74

[36] BOULICAUT, J-F. ; JEUDY, B.: Using Condensed Representations for Interactive Association Rule Mining. In: *Proc. of 6th PKDD Principles of Data Mining and Knowledge Discovery, 2002, Helsinki, Finland* Bd. 2431, Springer, 2002

[37] BRIN, S. ; PAGE, L.: The anatomy of a large-scale hypertextual Web search engine. In: *Computer Networks and ISDN Systems* 30 (1998), Nr. 1–7, S. 107–117

[38] BRINCK, T. ; HA, S.S. ; PRITULA, N. ; LOCK, K. ; SPEREDOLOZZI, A. ; MONAN, M.: Making an impact: Redesigning a Business School Web Site around Performance Metrics. In: *ACM* 6 (2003)

[39] BRODER, A.: A taxonomy of web search. In: *SIGIR Forum* 36 (2002), Nr. 2

[40] BRODER, A. ; GLASSMAN, S. ; MANASSE, M. ; ZWEIG, G.: Syntactic clustering of the web. In: *Proc. of 6th WWW'97*, 1997, S. 391–404

[41] BRODER, A.Z ; LEMPEL, R. ; MAGHOUL, F. ; PEDERSEN, J.: Efficient PageRank approximation via graph aggregation. In: *Information Retrieval* 9 (2006), Nr. 2, S. 123–138

[42] BROMBERGER, J. (Hrsg.): *Internetgestütztes Customer Relationship Management*. Deutscher Universitäts-Verlag, 2004

[43] BROWNE, J. ; HIGGINS, P. ; HUNT, I.: E-Business Principles, Trends and Visions. In: GASOS, J. (Hrsg.): *E-Business Applications*, Springer, 2002, S. 3–16

[44] BÜCHNER, A. ; BAUMGARTEN, M. ; MULVENNA, M. ; ANAND, S. ; HUGHES, J.: *Navigation Pattern Discovery from Internet Data*. submitted to ACM Workshop on Web Usage Analysis and User Profiling (WebKDD'99), 1999

[45] BUYS, M. ; BROWN, I.: Customer satisfaction with internet banking web sites: an empirical test and validation of a measuring instrument. In: *SAICSIT '04: Proc. Conf. South Africa IT research in developing countries* (2004), S. 44–52

[46] CAI, L. ; HOFMANN, T.: Text categorization by boosting automatically extracted concepts. In: *Proc. of SIGIR '03*. New York, NY, USA : ACM Press, 2003, S. 182–189

[47] CALERO, C. ; RUIZ, J. ; PIATTINI, M.: A Web Metrics Survey Using WQM. In: *Web Engineering, ICWE 2004, Munich, Proc.*, Springer, 2004, S. 147–160

[48] CAMILLO, F.: Clickstream Analysis, Semiotic Interpretation and Semantic Text Mining for a Distance Measurement on the Hypertextual Map of an Internet-portal. In: SIRMAKESSIS, S. (Hrsg.): *Text Mining and its Applications - Results of the NEMIS Launch Conference*, Springer, 2004

[49] CATLEDGE, L. ; PITKOW, J.: Characterizing browsing behaviors on the world wide web. In: *Computer Network ISDN Systems* 27 (1995), S. 1068–1073

[50] CHAKRABARTI, S: Data mining for hypertext: a tutorial survey. In: *SIGKDD Explor. Newsl.* 1 (2000), Nr. 2, S. 1–11

[51] CHAKRABARTI, S. ; BERG, M. van d. ; DOM, B.: Focused crawling: a new approach to topic-specific Web resource discovery. In: *Computer Networks (Amsterdam, Netherlands: 1999)* 31 (1999), Nr. 11–16, S. 1623–1640

[52] CHAKRABARTI, S. ; DOM, B.E. ; INDYK, P.: Enhanced hypertext categorization using hyperlinks. In: HAAS, L.M. (Hrsg.) ; TIWARY, A. (Hrsg.): *Proc. of SIGMOD-98, ACM*

International Conference on Management of Data. Seattle, US : ACM Press, New York, US, 1998, S. 307–318

[53] CHAKRABARTI, S. ; DOM, B.E. ; KUMAR, S.R. ; RAGHAVAN, P. ; RAJAGOPALAN, S. ; TOMKINS, A. ; GIBSON, D. ; KLEINBERG, J.: Mining the Web's Link Structure. In: *Computer* 32 (1999), Nr. 8, S. 60–67

[54] CHAMONI, P. ; GLUCHOWSKI, P. ; CHAMONI, P. (Hrsg.) ; GLUCHOWSKI, P. (Hrsg.): *Analytische Informationssysteme - Business Intelligence Technologien und Anwendungen*. Bd. 3. Springer, 2006

[55] CHAN, P.K.: Constructing Web User Profiles: A non-invasive Learning Approach. In: *WEBKDD*, 1999, S. 39–55

[56] CHAPMAN, P. ; CLINTON, J. ; KERBER, R. ; KHABAZA, T. ; REINARTZ, T. ; SHEARER, C. ; WIRTH, R.: CRISP-DM 1.0 Step-by-step data mining guide / CRISP-DM consortium. 2000. – Forschungsbericht

[57] CHAPMAN, P. ; CLINTON, J. ; KHABAZA, T. ; REINARTZ, T. ; WIRTH, R.: The CRISP-DM Process Model / CRISP-DM consortium. 1999. – Forschungsbericht

[58] CHEN, M. ; SUN, J.-T. ; ZENG, H.-J. ; LAM, K.-Y.: A practical system of keyphrase extraction for web pages. In: *CIKM '05: Proc. of Conf. on Information and Knowledge Management*. New York, NY, USA : ACM Press, 2005, S. 277–278

[59] CHEUNG, C.M.K. ; LEE, M.K.O.: Consumer satisfaction with internet shopping: a research framework and propositions for future research. In: *ICEC '05: Proc. of Conf. on Electronic Commerce*. New York, NY, USA : ACM Press, 2005, S. 327–334

[60] CHI, E. H. ; PIROLLI, P. ; CHEN, K. ; PITKOW, J.: Using information scent to model user information needs and actions and the Web. In: *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*. New York, NY, USA : ACM Press, 2001, S. 490–497

[61] CHI, E. H. ; PIROLLI, P. ; PITKOW, J.: The scent of a site: a system for analyzing and predicting information scent, usage, and usability of a Web site. In: *CHI '00: Proc. SIGCHI* (2000), S. 161–168

[62] CHI, E. H. ; ROSIEN, A. ; SUPATTANASIRI, G.: The bloodhound project: automating discovery of web usability issues using the InfoScent simulator. In: *Proc. of CHI* (2003), S. 505–512

[63] CHO, N. ; PARK, S.: Development of Electronic Commerce User-Consumer Satisfaction Index (ECUSI) for Internet Shopping. In: *Industrial Management & Data Systems* 101 (2001), S. 400–405

[64] CLAYPOOL, M. ; LE, P. ; WASED, M. ; BROWN, D.: Implicit interest indicators. In: *Intelligent User Interfaces*, 2001, S. 33–40

[65] COHEN, W.W.: Improving a Page Classifier with Anchor Extraction and Link Analysis. In: *Advanced Neural Information Processing Systems (NIPS)*, 2002, S. 1481–1488

[66] COHN, D. ; HOFMANN, T.: The Missing Link - A Probabilistic Model of Document Content and Hypertext Connectivity. In: *Advanced Neural Information Processing Systems (NIPS)* Bd. 13, NIPS, 2000

[67] COOLEY, R.: The Use of Web Structure and Content to Identify Subjectively Interesting Web Usage Patterns. In: *ACM Transaction on Internet Technology* 3 (2003), May, Nr. 2, S. 93–116

[68] COOLEY, R. ; MOBASHER, B. ; SRIVASTAVA, J.: Web Mining. Information and Pattern Discovery on the World Wide Web. In: *9th International Conference on Tools with Artificial Intelligence (ICTAI '97)* (1997), S. 558. ff

[69] COOLEY, R. ; MOBASHER, B. ; SRIVASTAVA, J.: Data Preparation for Mining World Wide Web Browsing Patterns. In: *Knowledge and Information Systems* 1 (1999), S. 5–32

[70] COOLEY, R. ; TAN, P.-N. ; SRIVASTAVA, J.: Websift: The web site information filter system. In: *WEBKDD'99 (San Diego)*, 1999

[71] COOPER, C.: Classifying Special Interest Groups in Web Graphs. In: J.D.P. ROLIM, S. V. (Hrsg.): *LNCS:Randomization and Approximation Techniques : 6th International Workshop, RANDOM 2002, Cambridge, MA, USA, September 13-15, 2002. Proceedings* Bd. 2483, Springer Berlin / Heidelberg, 2002, S. 263ff

[72] CORNELSEN, J.: Kundenbewertung mit Referenzwerten. In: GÜNTER, B. (Hrsg.) ; HELM, S. (Hrsg.): *Kundenwert*, 2001, S. 155–187

[73] CRABTREE, D. ; GAO, X. ; ANDREAE, P.: Improving Web Clustering by Cluster Selection. In: *WI '05: Proc. of Conf. on Web Intelligence (WI'05)*. Washington, DC, USA : IEEE Computer Society, 2005, S. 172–178

[74] CUTLER, M. ; STERNE, J.: *E-Metrics Business Metrics For The New Economy*. `http://www.netgencom/emetrics` (Zugriff: 27.05.2004), 2000

[75] CZYZOWICZ, J. ; KRANAKIS, E. ; KRIZANC, D. ; PELC, A. ; MARTIN, M.V.: Enhancing Hyperlink Structure for Improving Web Performance. In: *Journal of Web Engineering* 1 (2003), Nr. 2, S. 93–127

[76] DAI, H. ; MOBASHER, B.: Using ontologies to discover domain-level web usage profiles. In: *Second Semantic Web Mining Workshop at PKDD 2001, Helsinki, Finland* PKDD, 2002

[77] DEERWESTER, S C. ; DUMAIS, S. T. ; LANDAUER, T. K. ; FURNAS, G. W. ; HARSHMAN, R. A.: Indexing by Latent Semantic Analysis. In: *Journal of the American Society of Information Science* 41 (1990), Nr. 6, S. 391–407

[78] DEHASPE, L. ; TOIVONEN, H.: Frequent query discovery: a unifying ILP approach to association rule mining / K.U.Leuven. 1998 (CW-258). – Forschungsbericht

[79] DEHMER, M.: *Strukturelle Analyse Web-basierter Dokumente*, Universität Darmstadt, Diss., 2005

[80] DEMPSTER, A.P. ; LAIRD, N.M. ; RUBIN, D.B.: Maximum likelihood from incomplete data via the EM algorithm. In: *Journal of the Royal Statistical Society* 39 (1977), S. 1–38

[81] DESHPANDE, M. ; KARYPIS, G.: Selective Markov Models for Predicting Web Page Accesses. In: *ACM Transactions on Internet Technology* 4 (2004), May, Nr. 2, S. 163–184

[82] DEUTSCHER MULTIMEDIA VERBAND: *Viele Premium-Marken brauchen Relaunch ihres Webangebots*. `http://www.dmmv.de/shared/data/pressclipping`. Version: 2004. – Zugriff: 08.08.2006

[83] DEVARAJ, S. ; FAN, M. ; KOHLI, R.: Antecedents of B2C Channel Satisfaction and Preference: Validating e-Commerce Metrics. In: *Information Systems Research* 13 (2002), Nr. 3, S. 316–333

[84] DHYANI, D. ; KEONG NG, W. ; BHOWMICK, S.S.: A Survey of Web Metrics. In: *ACM Computing Surveys* 34 (2002), December, Nr. 4, S. 469–503

[85] DING, C. ; ZHA, H. ; HE, X. ; HUSBANDS, P. ; SIMON, H.: Link Analysis: Hubs and Authorities on the World Wide Web / LBNL. 2002 (47847). – Forschungsbericht

[86] DONG, G. ; LI, J.: Interestingness of Discovered Association Rules in Terms of Neighborhood-Based Unexpectedness. In: *PAKDD '98: Proc. of Pacific Conf. on Research and Development in Knowledge Discovery and Data Mining*. London, UK : Springer-Verlag, 1998, S. 72–86

[87] DONG, G. ; LI, J.: Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In: *SIGKDD'99, Fifth International Conference on Knowledge Discovery and Data Mining*, 43-52

[88] EIGHMEY, J.: Profiling User Responses to Commercial Web Sites. In: *Journal of Advertising Research* 37 (1997)

[89] EIRINAKI, M. ; VAZIRGIANNIS, M.: Web Mining for Web Personalization. In: *ACM Transactions on Internet Technology* 3 (2003), S. 1–27

[90] EIRINAKI, M. ; VAZIRGIANNIS, M. ; KAPOGIANNIS, D.: Web Path Recommendations based on Page Ranking and Markov Models. In: *WIDM '05: Proc 7th ACM Int. workshop on Web Information and Data Management*. New York, NY, USA : ACM Press, 2005, S. 2–9

[91] EROGLU, S.A. ; MACHLEIT, K.A. ; DAVIS, L.M.: Empirical Testing of a Model of Online Store Atmospherics and Shopper Responses. In: *Psychology and Marketing* 20 (2003), S. 139–150

[92] ETZIONI, O.: The world wide web: Quagmire or gold mine. In: *Communications of the ACM* 39 (1996), S. 65–68

[93] FAHRMEIER, L. ; KUENSTLER, R. ; PIGEOT, I. ; TUTZ, G. ; FAHRMEIER, L. (Hrsg.) ; PIGEOT, I. (Hrsg.) ; TUTZ, G. (Hrsg.): *Statistik - Der Weg zur Datenanalyse*. 5. Springer, 2004

[94] FAYYAD, U. M. ; PIATETSKY-SHAPIRO, G. ; SMYTH, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: *Commun. ACM* 39 (1996), Nr. 11, S. 27–34

[95] FELDEN, C.: Data Mining. In: CHAMONI, P. (Hrsg.) ; GLUCHOWSKI, P. (Hrsg.): *Text Mining als Anwendungsbereich von Business Intelligence*, Springer, 2006, S. 283–304

[96] FELDEN, C.: Extraktion, Qualitätssicherung und Klassifikation unstrukturierter Daten. In: *HMD-Praxis der Wirtschaftsinformatik* 247 (2006), S. 54–62

[97] FELLBAUM, C. (Hrsg.): *WordNet: An Electronic Lexical Database*. MIT Press, 1998

[98] FLAKE, G.W. ; FRASCONI, P. ; GILES, C.L. ; MAGGINI, M.: Machine Learning for the Internet. In: *ACM Transaction on Internet Technology* 4 (2004), May, Nr. 2, S. 125–128

[99] FLEMING, J. ; FLEMING, J. (Hrsg.): *Web Navigation: Designing the User Experience*. O'Reilly & Associates, 1999

[100] FOX, C.: Lexical analysis and stoplists. In: *Information retrieval: data structures and algorithms*. Upper Saddle River, NJ, USA : Prentice-Hall, Inc., 1992, S. 102–130

[101] FRANKE, J. ; NAKAHAEIZADEH, G. ; RENZ, I. ; FRANKE, J. (Hrsg.) ; NAKAHAEIZADEH, G. (Hrsg.) ; RENZ, I. (Hrsg.): *Text Mining: Theoretical Aspects and Applications*. Physica, 2003

[102] GARFIELD, E.: Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. In: *Science* 122 (1955), July, Nr. 3159, S. 108–111

[103] GASOS, J. ; THOBEN, K.-D. ; GASOS, J. (Hrsg.) ; THOBEN, K.-D. (Hrsg.): *E-Business Applications*. Springer, 2003

[104] GAUL, W. ; SCHMIDT-THIEME, L.: Mining web navigation path fragments. In: *Proc. of Workshop on Web Mining for E-Commerce - Challenges and Opportunities, Boston, MA*, 2000

[105] GAUL, W. ; SCHMIDT-THIEME, L.: Recommender systems based on navigation path features. In: *KDD2001 Workshop WEBKDD2001 - Mining Log Data Acreoss All Customer Touchpoints*, 2001

[106] GAUL, W. ; SCHMIDT-THIEME, L.: Aufzeichnung des Nutzerverhaltens - Erhebungstechniken und Datenformate. In: HIPPNER, H. (Hrsg.) ; MERZENICH, M. (Hrsg.) ; WILDE, K.D. (Hrsg.): *Handbuch Web Mining im Marketing*, vieweg, 2002, S. 35–52

[107] GAUL, W. ; SCHMIDT-THIEME, L.: Web Controlling und Recommendersysteme. In: HIPPNER, H. (Hrsg.) ; MERZENICH, M. (Hrsg.) ; WILDE, K.D. (Hrsg.): *Handbuch Web Mining im Marketing*, vieweg, 2002, S. 234–247

[108] GROB, H.-L. ; BROCKE, J. vom: Internetökonomie. In: GROB, H.-L. (Hrsg.) ; BROCKE, J. vom (Hrsg.): *Internetökonomie: Ein interdisziplinärer Beitrag zur Erklärung und Gestaltung hybrider Systeme*, Vahlen, 2006, S. 3–20

[109] GRÖNROOS, C.: Techn. Report: Strategic Management and Marketing in the Service Sector / Swedish School of Economics and Business Administration. 1982. – Forschungsbericht

[110] GUDIVADA, V.N. ; RAGHAVAN, V.V. ; GROSKY, W.I. ; KASANAGOTTU, R.: Information retrieval on the World Wide Web. In: *Internet Computing, IEEE* Bd. 1, 1997, S. 58–68

[111] HACOHEN-KERNER, Y. ; GROSS, Z. ; MASA, A.: Automatic Extraction and Learning of Keyphrases from Scientific Articles. In: GELBUKH, A. (Hrsg.): *LNCS: Proc. of CICLing, Computational Linguistics and Intelligent Text Processing: 6th International Conference* Bd. 3406, 2005, S. 657 – 669

[112] HAHN, J. ; KAUFFMAN, R. ; PARK, J.: Designing for ROI: Toward a Value-Driven Discipline for E-Commerce Systems Design. In: *Proc. of the 35th Hawaii International Conference on System Sciences, Hawaii*

[113] HAHN, J. ; KAUFFMAN, R.J.: *Evaluating Web Site Performance In Internet-Based Selling From a Business Value Perspective*. 2001. – submitted to: Int. Conf. on Electronic Commerce, Austria

[114] HANDL, A. ; HANDL, A. (Hrsg.): *Multivariate Analysemethoden - Theorie und Praxis multivariater Verfahren unter besonderer Beruecksichtigung von S-Plus*. Springer, 2002

[115] HASTIE, T. ; TIBSHIRANI, R. ; FRIEDMAN, J. ; HASTIE, T. (Hrsg.) ; TIBSHIRANI, R. (Hrsg.) ; FRIEDMAN, J. (Hrsg.): *The Elements of Statistical Learning - Data Mining, Inference and Prediction*. Springer, 2001

[116] HAUPTMANN, A.: Integrating and using large databases of text, image, video and audio. In: *IEEE Intelligent Systems* 14 (1999), S. 34–35

[117] HAWKEY, K. ; INKPEN, K.: Web browsing today: the impact of changing contexts on user activity. In: *CHI 05 extended abstracts on Human factors in computing systems*, ACM Press, 2005, S. 1443–1446

[118] HE, X. ; ZHA, H. ; DING, C.H.Q. ; SIMON, H.D.: Web document clustering using hyperlink structures / Pennsylvania State University, Dept. of Computer Science and Engineering, College of Engineering. 2001 (e&s-tech rept 38482012626400). – Forschungsbericht

[119] HENZINGER, M.: Web Information Retrieval - an Algorithmic Perspective. In: *European Symposium on Algorithms*, 2000, S. 1–8

[120] HETTICH, S. ; HIPPNER, H.: 11 Assoziationsanalyse. In: HIPPNER, H. (Hrsg.) ; KÜSTERS, U. (Hrsg.) ; MEYER, M. (Hrsg.) ; WILDE, K.D. (Hrsg.): *Handbuch Data Mining im Marketing*, Vieweg, 2001, S. 427–463

[121] HIPPNER, H.: CRM-Grundlagen, Ziele und Konzepte. In: HIPPNER, H. (Hrsg.) ; WILDE, K.D. (Hrsg.): *Grundlagen des CRM - Konzepte und Gestaltung*, Gabler, 2006, S. 17–44

[122] HIPPNER, H. (Hrsg.) ; KÜSTERS, U. (Hrsg.) ; MEYER, M. (Hrsg.) ; WILDE, K.D. (Hrsg.): *Handbuch Data Mining im Marketing*. Vieweg, 2001

[123] HIPPNER, H. ; MERZENICH, M. ; WILDE, K.D.: Grundlagen des Web Mining. In: HIPPNER, H. (Hrsg.) ; MERZENICH, M. (Hrsg.) ; WILDE, K.D. (Hrsg.): *Handbuch Web Mining im Marketing*, Vieweg, September 2002, S. 2–31

[124] HIPPNER, H. (Hrsg.) ; MERZENICH, M. (Hrsg.) ; WILDE, K.D. (Hrsg.): *Handbuch Web Mining im Marketing*. 1. Vieweg, 2002

[125] HIPPNER, H. ; MERZENICH, M. ; WILDE, K.D.: Web Mining im E-CRM. In: SCHÖGEL, M. (Hrsg.) ; SCHMIDT, I. (Hrsg.): *eCRM- mit Informationstechnologien Kundenpotenziale nutzen* Bd. 1, symposion, 2002, S. 87–102

[126] HIPPNER, H. ; RENTZMANN, R. ; WILDE, K.D.: Aufbau und Funktionalität von CRM-Systemen. In: HIPPNER, H. (Hrsg.) ; WILDE, K.D. (Hrsg.): *Grundlagen des CRM - Konzepte und Gestaltung*, Gabler, 2006, S. 45–74

[127] *Kapitel* 2. In: HIPPNER, H. ; WILDE, K.D.: *Der Prozess des Data Mining im Marketing*. Vieweg, 2001, S. 21–91

[128] HIPPNER, H. (Hrsg.) ; WILDE, K.D. (Hrsg.): *Grundlagen des CRM - Konzepte und Gestaltung*. Gabler, 2006

[129] HO, C.F. ; WU, W.H.: Antecedents of customer satisfaction on the Internet: an empirical study of online shopping. In: *Proc. of the 32nd Int. Conf. on Systems Sciences*, 1999

[130] HOFMANN, T.: Probabilistic latent semantic indexing. In: *Proc. of SIGIR '99*. ACM Press, 50–57

[131] HOFMANN, T.: Unsupervised Learning by Probabilistic Latent Semantic Analysis. In: *Machine Learning* 42 (2001), S. 177 – 196

[132] HOFMANN, T. ; BUHMANN, J.: Multidimensional Scaling and Data Clustering. In: TESAURO, G. (Hrsg.) ; TOURETZKY, D. (Hrsg.) ; LEEN, T. (Hrsg.): *Advances in Neural Information Processing Systems* Bd. 7, The MIT Press, 459–466

[133] HOI, S. C. ; JIN, R. ; LYU, M. R.: arge-scale text categorization by batch mode active learning. In: *Proc. of 15th Conference WWW '06.*, ACM Press, New York, NY, 2006, S. 633–642

[134] HOWARD, J.A. ; SHETH, J.N. ; HOWARD, J.A. (Hrsg.) ; SHETH, J.N. (Hrsg.): *The Theory of Buying Behaviour*. John Wiley & Sons Inc, 1969

[135] HUANG, L.: A survey on web information retrieval technologies / ECSL. 2000. – Forschungsbericht

[136] HUMPHREYS, J.B.K.: PhraseRate: An HTML Keyphrase Extractor / National Science Foundation. 2002 (CCR-9988360). – Forschungsbericht

[137] IBM: *IBM Intelligent Miner User's Guide*. 1996. – Version 1 Release 1 aus Handbuch Data Mining im Marketing

[138] IVORY, M.Y. ; SINHA, R.R. ; HEARST, M.A.: Empirically validated web page design metrics. In: *CHI '01: Proc. of SIGCHI Conf. on Human factors in computing systems* (2001), S. 53–60

[139] JIN, X. ; ZHOU, Y. ; MOBASHER, B.: Web usage mining based on probabilistic latent semantic analysis. In: *KDD '04: Proc. of ACM SIGKDD* (2004), S. 197–205

[140] JONEN, A. ; LINGNAU, V. ; MÜLLER, J. ; MÜLLER, P.: Balanced IT-Decision-Card - Ein Instrument für das Investitionscontrolling von IT-Projekten. In: *Wirtschaftsinformatik* 46 (2004), S. 196–203

[141] KAMVAR, S. ; HAVELIWALA, T. ; MANNING, C. ; GOLUB, G.: Exploiting the block structure of the web for computing PageRank / Stanford University, Technical Report. 2003. – Forschungsbericht

[142] KAUFMAN, R.J. ; WALDEN, E.A.: Economics and electronic commerce: Survey and directions for research. In: *Int. Journal of Electronic Commerce* 5 (2001), S. 5–116

[143] KHALIFA, M. ; SHEN, N.: Effects of Electronic Customer Relationship Management on Customer Satisfaction: A Temporal Model. In: *Proc. of Int. Conf. on System Sciences*, 2005

[144] KHOLI, R. ; DEVARAJ, S. ; MAHMOOD, M.A.: Understanding Determinants of Online Consumer Satisfaction: A Decision Process Perspective. In: *Journal of Management Information Systems* 21 (2004), S. 115–135

[145] KIM, J. ; LEE, J. ; HAN, K. ; LEE, M.: Business as buildings: Metrics for architectural quality of Internet business. In: *Information Systems Research* 13 (2002), S. 205–223

[146] KIM, S.Y. ; LIM, Y.J.: Consumer's Perceived Importance and Satisfaction with Internet Shopping. In: *Electronic Markets* 11 (2001), S. 148–154

[147] KING, O. ; KOBAYASHI, M.: Information Retrieval and Ranking on the Web: Benchmarking Studies II / Solution Research Center of IBM Tokyo Reseaerch Laboratory. 1999. – Forschungsbericht

[148] KLEINBERG, J. M.: Authoritative sources in a hyperlinked environment. In: *Journal of the ACM* 46 (1999), Nr. 5, S. 604–632

[149] KLEIST, V.F.: An approach to evaluating e-business information systems projects. In: *Information Systems Frontiers* 5 (2003), S. 249–263

[150] KOBAYASHI, M. ; TAKEDA, K.: Information retrieval on the web. In: *ACM Computing Surveys* 32 (2000), Nr. 2, S. 144–173

[151] KOSALA ; BLOCKEEL: Web Mining Research: A Survey. In: *SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM* 2 (2000)

[152] KRÜGER, W. ; BACH, N.: Geschäftsmodelle und Wettbewerb im e-Business. In: BUCHHOLZ, W. (Hrsg.) ; WERNER, H. (Hrsg.): *Supply Chain Solutions - Best Practices im E-Business*, Schaeffer-Poeschel, 2001, S. 29–51

[153] KÜPPERS, B.: *Data Mining in der Praxis - Ein Ansatz zur Nutzung der Potentiale von Data Mining im betrieblichen Umfeld*, TU Graz, Diss., 1998

[154] KÜSTERS, U.: Data Mining Methoden: Einordnung und Überblick. In: HIPPNER, H. (Hrsg.) ; KÜSTERS, U. (Hrsg.) ; MEYER, M. (Hrsg.) ; WILDE, K.D. (Hrsg.): *Handbuch Data Mining im Marketing*, Vieweg, 2001, S. 95–130

[155] LAKSHMINARAYAN, C. ; YU, Q. ; BENSON, A.: Improving Customer Experience via Text Mining. In: *Databases in Networked Information Systems* Bd. 3433, 2005, S. 288 – 299

[156] LANDAUER, T.K. ; FOLTZ, P.W. ; LAHAM, D.: Introduction to Latent Semantic Analysis. In: *Discourse Processes* 25 (1998), S. 259–284

[157] LEDERER, A.L. ; MIRCHANDANI, D.A. ; SIMS, K.: The seaerch for strategic advantage from the World Wide Web. In: *Int. Journal of Electronic Commerce* 5 (2001), S. 117–133

[158] LEE, J.: Model-driven business transformation and the semantic web. In: *Commun. ACM* 48 (2005), Nr. 12, S. 75–77

[159] LEE, J. ; PODLASECK, M. ; SCHONBERG, E. ; HOCH, R.: Visualitation and Analysis of Clickstream Data of Online Stores for Understanding Web Merchandising. In: *Data Mining and Knowledge Discovery* 5 (2001), S. 59–84

[160] LEE, U. ; LIU, Z. ; CHO, J.: Automatic Identification of User Goals in Web Search. In: *WWW Conference* IW3C2, ACM, May 2005

[161] LI, J. ; ZHANG, X. ; DONG, G. ; RAMAMOHANARAO, K. ; SUN, Q.: Efficient Mining of High Confidence Association Rules without Support Thresholds. In: *PKDD99* Bd. 1704, Springer, 406–411

[162] LIN, J. ; FERNANDES, A. ; KATZ, B. ; MARTON, G. ; TELLEX, S.: *Extracting Answers from the Web Using Knowledge Annotation and Knowledge Mining Techniques*

[163] LIU.J.; YAO, Y. ; ZHONG, N. (Hrsg.): *Web Intelligence*. Springer, 2003

[164] LU, Z. ; YAO, Y. ; ZHONG, N.: Web Log Mining. In: ZHONG, N. (Hrsg.) ; LIU, J. (Hrsg.) ; YAO, Y. (Hrsg.): *Web Intelligence*, Springer, 2003, S. 173–194

[165] MADEJA, N.: *Corporate Success in Electronic Business - Results from an Empirical Investigation*, WHU Vallendar, Diss., 2005

[166] MALLETT, D. ; ELDING, J. ; NASCIMENTO, M.A.: Information-Content Based Sentence Extraction for Text Summarization. In: *ITCC* 02 (2004), S. 214

[167] MANNILA, H. ; TOIVONEN, H. ; VERKAMO, A.I.: Discovery of Frequent Episodes in Event Sequences. In: *Data Mining and Knowledge Discovery* 1 (1997), Nr. 3, S. 259–289

[168] MEFFERT, H. ; MEFFERT, H. (Hrsg.): *Marketing-Management: Analyse - Strategie - Implementierung*. Gabler, 1994

[169] MEIER, A. ; STROMER, H. ; MEIER, A. (Hrsg.) ; STROMER, H. (Hrsg.): *e-Business & eCommerce*. Springer, 2005

[170] MENA, J. ; MENA, J. (Hrsg.): *Data Mining Your Website*. Digital Press, 1999

[171] MENCZER, F. ; PANT, G. ; SRINIVASAN, P.: *Topic-driven crawlers: Machine learning issues*. 2002. – `http://dollar.biz.uiowa.edu/fil/Papers/TOIT.pdf` (Zugriff: 24.07.2006)

[172] MILLER, G.A. ; FELLBAUM, C. ; TENGI, R. ; WAKEFIELD, P. ; PODDAR, R. ; LANGONE, H. ; HASKELL, B.: *http://wordnet.princeton.edu/*. Zugriff: 24.07.2006. – Lexikalische Datenbank - Synonyme

[173] MOBASHER, B.: In: SINGH, M.P. (Hrsg.): *Web Usage Mining and Personalization*, Chapman & Hall/ CRC Press, 2004, S. 1–35

[174] MOBASHER, B. ; COOLEY, R. ; SRIVASTAVA, J.: Automatic Personalization Based on Web Usage Mining. In: *Communications of the ACM* 43 (2000), S. 142–151

[175] MOBASHER, B. ; DAI, H. ; LUO, T. ; NAKAGAWA, M.: Effective Personalization Based on Association Rule Discovery from Web Usage Data. In: *Workshop On Web Information And Data Management* (2001), S. 9 – 15

[176] MOBASHER, B. ; DAI, H. ; LUO, T. ; SUN, Y. ; ZHU, J.: Integrating Web Usage and Content Mining for More Effective Personalization. In: *Proc. of the Int'l Conf. on E-Commerce and Web Technologies (ECWeb2000)* (2000)

[177] MÜLLER, A. ; THIENNENM, L. von: *Das Controlling Cockpit für IT-Projekte - Das Controlling von E-Business-Vorhaben mit der Balanced Scorecard*. `www.business-process-solutions.de`, 2002. – Zugriff: 13.01.2005

[178] NASRAOUI, O. ; CARDONA, C. ; ROJAS, C.: Using retrieval measures to assess similarity in mining dynamic web clickstreams. In: *KDD '05: Proc. of SIGKDD Conf. on Knowledge Discovery in Data Mining*. New York, NY, USA : ACM Press, 2005, S. 439–448

[179] NIELSON, J. ; NIELSON, J. (Hrsg.): *Usability Engineering*. Academic Press, 1993

[180] OBERLE, D. ; BERENDT, B. ; HOTHO, A. ; GONZALEZ, J.: Conceptual User Tracking. In: *Proceedings of the Atlantic Web Intelligence Conference* (2002), S. 155 – 164

[181] OLSINA, L. ; LAFUENTE, G. ; ROSSI, G.: Specifying Quality Characteristics and Attributes for Websites,. In: MURUGESAN, S. (Hrsg.) ; Y. DESPHANDE, Y. (Hrsg.): *Web Engineering : Software Engineering and Web Application Development* Bd. 2016/2001, Springer, 2001, S. 266ff

[182] OLSINA, L. ; ROSSI, G.: Measuring Web application quality with WebQEM. In: *Multimedia, IEEE* 9 (2002), S. 20–29

[183] OLSTON, C. ; CHI, E.H.: ScentTrails: Integrating browsing and searching on the Web. In: *ACM Trans. Comput.-Hum. Interact.* 10 (2003), Nr. 3, S. 177–197

[184] PAGE, L. ; BRIN, S. ; MOTWANI, R. ; WINOGRAD, T.: *The PageRank Citation Ranking: Bringing Order to the Web*. Stanford Digital Library Technologies Project, 1998

[185] PALMER, J.W.: Web Site Usability, Design and Performance Metrics. In: *Information Systems Research* 13 (2002), S. 151–167

[186] PANT, G. ; MENCZER, F.: Topical crawling for business intelligence. In: KOCH, T (Hrsg.) ; SOLVBERG, I. (Hrsg.): *Proc. of 7th European Conf. on Research and Adv. Technology for Digital Libraries (ECDL)*, Springer, 2003

[187] PAPINENI, K.: Why inverse document frequency ? In: *Proc. North American Association for Computational Lunguistics*, 2001, S. 25–32

[188] PATHER, S. ; ERWIN, G. ; REMENYI, D.: Measuring e-Commerce effectiveness: a conceptual model. In: *SAICSIT '03* (2003), S. 143–152

[189] PIATSKY-SHAPIRO, G. ; BRAACHMAN, R. ; KHABAZA, T. ; KLOESGEN, W. ; SIMOUDIS, E.: An overview of issues in developing industrial data mining and knowledge discovery applications. In: *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, 1996, S. 89–95

[190] PIROLLI, P.: Exploring browser design trade-offs using a dynamical model of optimal information foraging. In: *CHI '98: Proceedings of the SIGCHI conference on Human factors in computing systems*. New York, NY, USA : ACM Press/Addison-Wesley Publishing Co., 1998, S. 33–40

[191] PIROLLI, P. ; PITKOW, J. ; RAO, R.: Silk from a Sow's Ear: Extracting Usable Structures from the Web. In: *Proc. ACM Conf. Human Factors in Computing Systems, CHI*, ACM Press

[192] PIROLLI, P.L.T. ; CARD, S.K.: Information Foraging. In: *Psychological Review* 106 (1999), S. 643–675

[193] PIROLLI, P.L.T. ; PITKOW, J.E.: Distributions of surfers paths through the World Wide Web: Empirical characterizations. In: *World Wide Web* 2 (1999), Nr. 1-2, S. 29–45

[194] PITKOW, J.: Web Characterization Activity Characterization Metrics / W3C. Version: 1998. `http://www.w3c.org/WCA/Metrics.html`. – Forschungsbericht. – Elektronische Ressource. Zugriff: 04.05.2004

[195] PORTER, M. F.: An algorithm for suffix stripping. In: *Program* 14 (1980), S. 130–137

[196] PUJOL, J.M. ; SANGÜESA, R. ; DELGADO, J.: A Ranking Algorithm Based on Graph Topology to Generate Reputation or Relevance. In: ZHONG, N. (Hrsg.) ; LIU, J. (Hrsg.) ; YAO, Y. (Hrsg.): *Web Intelligence*, Springer, 2003, S. 380–393

[197] PYLE, D. ; CERRA, D. (Hrsg.): *Data Preparation For Data Mining*. Morgan Kaufmann Publishers, Inc, 1999

[198] R DEVELOPMENT CORE TEAM: *R: A Language and Environment for Statistical Computing*. `http://www.R-project.org`. Version: 2007

[199] RAJOPAL, S. ; VENKATACHALAM, M. ; KOTHA, S.: Does the Quality of Online Customer Experience Create a Sustainable Competitive Advantage for E-Commerce Firms ? / School of Business Administration, University of Washington, Seattle and Center of Electronic Business and Commerce, Stanford. 2001. – Forschungsbericht

[200] RESNICK, P. ; VARIAN, H. R.: Recommender systems. In: *Commun. ACM* 40 (1997), Nr. 3, S. 56–58

[201] RIEMER, K. ; MÜLLER-LANKENAU, C. ; KLEIN, S.: Internet-Qualitätsmanagement - Klassifikation und Anwendung von Methoden der Web-Evaluation. In: GROB, H.-L. (Hrsg.) ; BROCKE, J. vom (Hrsg.): *Internetökonomie: Ein interdisziplinärer Beitrag zur Erklärung und Gestaltung hybrider Systeme*, Vahlen, 2006, S. 249–277

[202] SALTON, G. ; SALTON, G. (Hrsg.): *Automatic Information Organization and Retrieval.* McGraw Hill, 1968

[203] SALTON, G. ; SALTON, G. (Hrsg.): *Automatic text processing*. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc., 1988

[204] SALTON.G ; MCGILL, M.J. ; SALTON.G (Hrsg.) ; MCGILL, M.J. (Hrsg.): *Introduction to modern information retrieval*. McGraw-Hill, 1983

[205] SANE SOLUTIONS: *Analyzing Web Site Traffic - A Sane Solutions White Paper*. www.sane.com, 1999-2003. – Zugriff: 07.11.2005

[206] SANE SOLUTIONS: *NetTracker Log File Analysis vs. Page Tagging - A Guide for Comparing Web Analytics Methodologies*. www.sane.com, 2003. – Zugriff: 07.11.2005

[207] SÄUBERLICH, F.: Vorverarbeitung von Web-Daten - Pre-Processing. In: WILDE, K.D. (Hrsg.): *Handbuch Web Mining im Marketing*, Vieweg, September 2002, S. 107–123

[208] SCHIRA, J. ; SCHIRA, J. (Hrsg.): *Statistische Methoden der VWL und BWL.* Pearson Studium, 2005

[209] SCHMITT, E.: Measuring Web Success / Forrester Report. 1999. – Forschungsbericht. `www.forrester.com` (Zugriff: 08.08.2006)

[210] SCHONBERG, E. ; COFINO, T. ; AL., R. H.: Measuring success. In: *Commun. ACM* 43 (2000), Nr. 8, S. 53–57

[211] SCHUMPETER, J.A. ; SCHUMPETER, J.A. (Hrsg.): *Capitalism, Socialism and Democracy*. Harper & Brothers, 1942

[212] SCHWICKERT, A. C.: Geschäftsmodelle im Electronic Business - Bestandsaufnahme und Relativierung / Justus-Liebig-Universität Gießen. 2004 (2). – Arbeitspapiere WI

[213] SCHWICKERT, A. C. ; WENDT, P.: Controlling Kennzahlen fuer Web Sites / Justus-Liebig-Universität Gießen. 2000 (2). – Arbeitspapiere WI. Arbeitspapiere WI

[214] SHAHABI, C. ; ZARKESH, A.M. ; ADIBI, J. ; SHAH, V.: Knowledge discovery from users Web-page navigation. In: *Proc. of 7th Seventh International Workshop on Research Issues in Data Engineering*, 1997

[215] SHEN, D. ; SUN, J. ; YANG, Q. ; CHEN, Z.: A comparison of implicit and explicit links for web page classification. In: *In Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland, May 23 - 26, 2006). WWW '06*, ACM Press, New York, NY, 2006, S. 643–650

[216] SIRMAKESSIS, S. ; SIRMAKESSIS, S. (Hrsg.): *Text Mining and Its Applications*. Springer, 2004

[217] SIRMAKESSIS, S. ; SIRMAKESSIS, S. (Hrsg.): *Adaptive and Personalized Semantic Web. Studies in Computational Intelligence*. Springer, 2006

[218] SKIERA, B. ; LAMBRECHT, A.: *Erlösmodelle im Internet*. 2000. – Vorgesehen für das Buch "Produktmanagement", herauszugeben von Sönke Albers und Andreas Herrmann

[219] SKIERA, B. ; SOUKHOROUKOVA, A. ; GÜNTHER, O. ; WEINHARDT, C.: Internetökonomie. In: *WI-Wirtschatftsinformatik* 48 (2006), S. 1–2

[220] SOFTWARE, Freshwater: *The Insider's Guide to Web Monitoring*. White Paper,

[221] SPILIOPOULOU, M.: Web Usage Mining for Web Site Evaluation. In: *Communications of the ACM* 43 (2000), S. 127–134

[222] SPILIOPOULOU, M.: Web-Mining - Ein Erfahrungsbericht. In: HIPPNER, H. (Hrsg.) ; KÜSTERS, U. (Hrsg.) ; MEYER, M. (Hrsg.) ; WILDE, K.D. (Hrsg.): *Handbuch Data Mining im Marketing*, Vieweg, 2001, S. 899–903

[223] SPILIOPOULOU, M.: Web Usage Mining: Data Mining über die Nutzung des Web. In: HIPPNER, H. (Hrsg.) ; KÜSTERS, U. (Hrsg.) ; MEYER, M. (Hrsg.) ; WILDE, K.D. (Hrsg.): *Handbuch Data Mining im Marketing*, Vieweg, 2001, S. 489–510

[224] SPILIOPOULOU, M. ; FAULSTICH, L. C.: WUM: A Tool for Web Utilization Analysis. In: *LNCS: The World Wide Web and Databases* 1590 (1999), S. 184–203

[225] SPILIOPOULOU, M. ; POHLE, C.: Data Mining for Measuring and Improving the Success of Web Sites. In: *Data Min. Knowl. Discov.* 5 (2001), Nr. 1-2, S. 85–114

[226] SPILIOPOULOU, M. ; POHLE, C. ; FAULSTICH, L.: Improving the Effectiveness of a Web Site with Web Usage Mining. (1999), S. 142–162

[227] SPLILIOPOULOU, M.: The laborious way from data mining to web mnining. In: *International Journal of Computer Systems, Science and Engineering* 14 (1999), Nr. 2, S. 113–126

[228] SPOOL, J. ; SCHROEDER, W. ; SCANLON, T. ; SNYDER, C.: Web Sites that Work: Designing with Your Eyes Open. In: *Chi 98* (1998), April, S. 147–148

[229] SRIVASTAVA, J. ; COOLEY, R. ; DESHPANDE, M. ; TAN, P.-N.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. In: *SIGKDD Explorations* 1 (2000), Nr. 2, S. 12–23

[230] STÄHLER, P. ; STÄHLER, P. (Hrsg.): *Geschäftsmodelle in der digitalen Ökonomie*. Lohmar, 2001

[231] STAHLKNECHT, P. ; HASENKAMP, U. ; STAHLKNECHT, P. (Hrsg.) ; HASENKAMP, U. (Hrsg.): *Einführung in die Wirtschaftsinformatik*. Springer, 2005

[232] STAUSS, B.: Perspektivenwandel: Vom Produkt-Lebenszyklus zum Kundenbeziehungs-Lebenszyklus. In: *Thexis* 2 (2000), S. 15–18

[233] STAUSS, B.: Grundlagen und Phasen der Kundenbeziehung: Der Kundenbeziehungs-Lebenszyklus. In: HIPPNER, H. (Hrsg.) ; WILDE, K.D. (Hrsg.): *Grundlagen des CRM - Konzepte und Gestaltung*, Gabler, 2006, S. 423–442

[234] STERNE, J. ; STERNE, J. (Hrsg.): *Web Metrics: Proven Methods for Measuring Web Site Success*. Wiley, New York, 2002

[235] STOLZ, C.: Benutzeranalyse - Umsetzung des Apriori Algorithmus zur Sequenzanalyse / Katholische Universtität Eichstätt-Ingolstadt, Lehrstuhl für Wirtschaftsinformatik. 2004. – Forschungsbericht

[236] STOLZ, C. ; BARTH, M.: Website Performance Analysis - Success Assessment of Information Driven Website on User Traces. In: *International Journal of Information Technology and Web Engineering* 2 (2007), July - September 2007, Nr. 3, S. 37–52

[237] STOLZ, C. ; BARTH, M.: Website Success Assessment. In: *invited for Web Intelligence and Agent Systems: An International Journal* (2007)

[238] STOLZ, C. ; BARTH, M. ; VIERMETZ, M. ; WILDE, K.D.: Searchstrings Revealing User Intent - A Better Understanding of User Perception. In: *ICWE 2006, Stanford, Palo Alto*, 2006

[239] STOLZ, C. ; GEDOV, V. ; SEIPEL, D. ; NEUNEIER, R. ; SKUBACZ, M.: Matching Web Site Structure and Content. In: *The 13th Int. WWW Conf. Proceedings* ACM, ACM, 2004, S. 286–287

[240] STOLZ, C. ; GEDOV, V. ; YU, K. ; NEUNEIER, R. ; SKUBACZ, M.: Measuring Semantic Relations of Web Sites by Clustering of Local Context. In: *LNCS: ICWE 2004, Munich*, Springer, 2004, S. 182–186

[241] STOLZ, C. ; VIERMETZ, M. ; SKUBACZ, M. ; NEUNEIER, R.: Guidance Performance Indicator - Web Metrics for Information Driven Web Sites. In: *IEEE Intl. Conf. Web Intelligence 2005, Proc.*, 2005

[242] STOLZ, C. ; VIERMETZ, M. ; SKUBACZ, M. ; NEUNEIER, R.: Improving Semantic Consistency of Web Sites by Quantifying User Intent. In: *Springer LNCS: Int. Conf on Web Engineering, ICWE 2005, Sydney* (2005)

[243] STRAUB, D. W. ; HOFFMAN, D. L. ; WEBER, B. W. ; STEINFIELD, C.: Measuring e-Commerce in Net-Enabled Organizations: An Introduction to the Special Issue. In: *Info. Sys. Research* 13 (2002), Nr. 2, S. 115–124

[244] STRAUB, D.W. ; HOFFMAN, D.L. ; WEBER, B.W. ; STEINFIELD, C.: Toward New Metrics for Net-Enhanced Organizations. In: *Information Systems Research* 13 (2002), Nr. 3, S. 227–238

[245] SUBRAHMANIAN, V.S. ; SUBRAHMANIAN, V.S. (Hrsg.): *Principles of Multimedia Database Systems*. Morgan Kaufmann Publishers, 1998

[246] SULLIVAN, T.: Reading reader reaction: A proposal for inferential analysis of web server log files. In: *Proc. of the Human Factors and the Web 3 Conference*, 1997

[247] TEEVAN, J. ; DUMAIS, S. ; HORVITZ, E.: Personalizing Search via Automated Analysis of Interests and Activities. In: *SIGIR*, 2005

[248] TELTZROW, M. ; BERENDT, B.: Web-Usage-Based Success Metrics for Multi-Channel Businesses. In: *The 9th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining - WEBKDD Workshop*, 2003

[249] TELTZROW, M. ; GÜNTHER, O.: Web Metrics for Retailers. In: *LNCS: EC-Web 2003* Bd. 2738, 2003, S. 328 – 338

[250] TITTMANN, P. ; TITTMANN, P. (Hrsg.): *Graphentheorie*. Fachbuchverlag Leipzig, 2003

[251] TORKZADEH, G. ; DHILLON, G.: Measuring factors that influence the success of Internet commerce. In: *Information Systems Research* 13 (2002), S. 187–204

[252] TOTZ, C. ; RIEMER, K. ; KLEIN, S.: Web Evaluation. In: LOWRY, P.B. (Hrsg.) ; CHERRINGTON, R.R. (Hrsg.): *The E-Business Handbook*, Watson, 2001, S. 45–66

[253] TSENG, V.S. ; CHANG, J.-C. ; LIN, K.W.: Mining and prediction of temporal navigation patterns for personalized services in e-commerce. In: *SAC '06: Proceedings of the 2006 ACM symposium on applied computing*. New York, NY, USA : ACM Press, 2006, S. 867–871

[254] VIERMETZ, M. ; STOLZ, C. ; GEDOV, V. ; SKUBACZ, M.: Relevance and Impact of Tabbed Browsing Behavior on Web Usage Mining. In: *Web Intelligence 2006*, 2006

[255] VIERMETZ, M. ; STOLZ, C. ; GEDOV, V. ; SKUBACZ, M.: Relevance and Impact of Tabbed Browsing Behavior on Web Usage Mining. In: *invited for Journal of Web Intelligence* (2007)

[256] WADE, M.R. ; NEVO, S.: Development and Validation of a Perceptual Instrument to Measure E-Commerce Performance. In: *International Journal of Electronic Commerce* 10 (2005), Nr. 2, S. 123–146

[257] WANG, Y. ; KITSUREGAWA, M.: Evaluaating Contents-Link Coupled Web Page Clustering for Web Search Results. In: *CIKM'02*, 2002

[258] WEISS, M.S. ; INDURKHYA, N. ; ZHANG, T. ; DAMERAU, F.J. ; WEISS, M.S. (Hrsg.) ; INDURKHYA, N. (Hrsg.) ; ZHANG, T. (Hrsg.) ; DAMERAU, F.J. (Hrsg.): *Text Mining - Predictive Methods for Analyzing Unstructured Information*. Springer, 2004

[259] WELTY, C.: Towards a Semantics for the Web / Dagstuhl Symposium on Semantics for the Web. 2000. – Invited Presentation

[260] WIKIPEDIA: *EM-Algorithmus*. Zugriff: 13.7.2006. `http://de.wikipedia.org/wiki/EM_Algorithmus`

[261] WIKIPEDIA: *Latent Semantic Indexing*. Zugriff: 13.7.2006. `http://de.wikipedia.org/wiki/Latent_Semantic_Indexing`

[262] WINKLER, K. ; SPILIOPOULOU, M.: Employing Text Mining for Semantic Tagging in DIAsDEM. In: *KI Künstliche Intelligenz - Special issue on text mining* (2002), S. 27–29

[263] WIRTZ, B. ; BECKER, D.R.: Erfolgsrelevanz und Entwicklungsperpektiven von Geschäftsmodellvarianten im Electronic Business. In: *Wirtschaftswissenschaftliches Studium* 3 (2002)

[264] WIRTZ, B. ; BECKER, D.R.: Geschäftsmodellansätze und Geschäftsmodellvarianten im Electronic Business. In: *WiSt - Wirtschaftsstudium* 2 (2002), S. 85–90

[265] WROBEL, S. ; MORIK, K. ; JOACHIMS, T.: Maschinelles Lernen und Data Mining. In: GÖRZ, G. (Hrsg.) ; ROLLINGER, C.-R. (Hrsg.): *Handbuch der künstlichen Intelligenz*, Oldenburg, 2003, S. 517–597

[266] XU, G. ; ZHANG, Y. ; MA, J. ; ZHOU, X: Discovering user access pattern based on probabilistic latent factor model. In: *ADC '05: Proceedings of the sixteenth Australian database conference*. Darlinghurst, Australia : Australian Computer Society, Inc., 2005, S. 27–35

[267] XUE, G.-R. ; YANG, Q. ; ZENG, H.J. ; YU, Y. ; CHEN, Z.: Exploiting the hierarchical structure for link analysis. In: *Proc. of SIGIR '05*. New York, NY, USA : ACM Press, 2005, S. 186–193

[268] YEUNG, W.L. ; LU, M.: Functional Characteristics of Commercial Web Sites: A Longitudinal Study in Hong Kong. In: *Information and Management* 41 (2004), S. 483–495

[269] ZAIANE, O.R. ; HAN, J. ; LI, Z.N. ; CHEE, S.H. ; CHIANG, J.: Multimediaminer: a system prototype for multimedia data mining. In: *ACM SIGMOD Intl. Conf. on Management of Data*, 1998, S. 581–583

[270] ZAMIR, O. ; ETZIONI, O.: Web document clustering: a feasibility demonstration. In: *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA : ACM Press, 1998, S. 46–54

[271] ZENG, C. ; XING, C.-X. ; ZHOU, L.Z.: Similarity measure and instance selection for collaborative filtering. In: *WWW '03: Proceedings of the 12th international conference on World Wide Web*. New York, NY, USA : ACM Press, 2003, S. 652–658

[272] ZERDICK, A. ; PICOT, A. ; SCHRAPE, K. ; ARTOPE, A. ; GOLDHAMMER, K. ; LANGE, U.T. ; VIERKANT, E. ; LOPEZ-ESCOBAR, E. ; SILVERSTONE, R.: *Die Internet-Ökonomie: Strategien für die digitale Wirtschaft*. 1998. – European Communication Council Report

[273] ZHONG, N. ; LIU, J. ; YAO, Y. ; ZHONG, N. (Hrsg.) ; LIU, J. (Hrsg.) ; YAO, Y. (Hrsg.): *Web Intelligence*. Springer, 2003

[274] ZHU, J ; HONG, J. ; HUGHES, J.G.: PageCluster: Mining Conceptual Link Hierarchies from Web Log Files for Adaptive Web Site Navigation. In: *ACM Transaction on Internet Technology* 4 (2004), May, Nr. 2, S. 185–2085

[275] ZHU, K. ; KRAEMER, K.L.: e-Commerce Metrics for Net-Enhanced Organizations: Assessing the Value of e-Commerce to Firm Performance in the Manufacturing Sector. In: *Information Systems Research* 13 (2002), S. 275–295

[276] ZHU, Y. ; YE, S. ; LI, X.: Distributed PageRank computation based on iterative aggregation-disaggregation methods. In: *CIKM '05: Proceedings of the 14th ACM international conference on Information and Knowledge Management*. New York, NY, USA : ACM Press, 2005, S. 578–585

[277] ZVIRAN, M. ; GLEZER, C. ; AVNI, I.: User satisfaction from commercial web sites: The effect of design and use. In: *Information & Management* (2005)