Verfahren zur Anfragemodifikation im Information Retrieval

Diplomarbeit

Studiengang Bibliothekswesen

Fakultät für Informations- und Kommunikationswissenschaften

Fachhochschule Köln

vorgelegt von:

Matthias Nagelschmid
Matr.-Nr.: 11047228

am 22.09.2008

bei Prof. Dipl.-Math. Winfried Gödert und Prof. Dr. Klaus Lepsky

Abstract

Für das Modifizieren von Suchanfragen kennt das Information Retrieval vielfältige Möglichkeiten. Nach einer einleitenden Darstellung der Wechselwirkung zwischen Informationsbedarf und Suchanfrage wird eine konzeptuelle und typologische Annäherung an Verfahren zur Anfragemodifikation gegeben. Im Anschluss an eine kurze Charakterisierung des Fakten- und des Information Retrieval, sowie des Vektorraum- und des probabilistischen Modells, werden intellektuelle, automatische und interaktive Modifikationsverfahren vorgestellt. Neben klassischen intellektuellen Verfahren, wie der Blockstrategie und der „Citation Pearl Growing"-Strategie, umfasst die Darstellung der automatischen und interaktiven Verfahren Modifikations-möglichkeiten auf den Ebenen der Morphologie, der Syntax und der Semantik von Suchtermen. Darüber hinaus werden das Relevance Feedback, der Nutzen informetrischer Analysen und die Idee eines assoziativen Retrievals auf der Basis von Clustering- und terminologischen Techniken, sowie zitationsanalytischen Verfahren verfolgt. Ein Eindruck für die praktischen Gestaltungsmöglichkeiten der behandelten Verfahren soll abschließend durch fünf Anwendungsbeispiele vermittelt werden.

Schlagwörter

Anfragemodifikation ; Informationsbedarf ; Information Retrieval ; assoziatives Retrieval ; Relevance Feedback ; semantische Umfeldsuche

Inhaltsverzeichnis

Abbildungs- und Tabellenverzeichnis

1. Einführung

Mit der zunehmenden Verfügbarkeit von Informationen gewinnt auch der Prozess des Information Retrieval an Bedeutung. Durch das World Wide Web wurde der Zugang zu Informationssystemen vereinfacht, für den Vorgang des Suchens und Findens von Informationen muss der Nutzer jedoch mit dem jeweiligen Informationssystem vertraut sein. Er tritt in einen Dialog, indem er dem System eine Suchanfrage zur Verarbeitung übergibt. Liefert die Anfrage ein Ergebnis, durch das sich der Informationsbedarf nicht befriedigen lässt, besteht die Möglichkeit, die Anfrage anders zu formulieren, um ein besseres Ergebnis zu erzielen. Die vielfältigen Möglichkeiten und Abläufe solcher Anfragemodifikationen sind Gegenstand dieser Arbeit.

Das Formulieren und Reformulieren von Suchanfragen kann eine große Hürde darstellen und den Zugang zu Informationen erschweren. Diese und andere Informationsbarrieren sind Hindernisse, die den freien Fluss von Informationen hemmen[1]. Neben den Barrieren, die einen eher allgemeinen, gesellschaftlichen Charakter haben (beispielsweise die politisch-ideologische Barriere) und solchen Barrieren, die die Anbieter von Informationsdiensten ihren Nutzern selbst auferlegen (beispielsweise die Finanzierungsbarriere), stellt das eingangs beschriebene Problem eine kommunikative Barriere zwischen Nutzer und Retrievalsystem, oder allgemeiner: zwischen Mensch und Computer dar.

Die Herabsetzung solcher kommunikativer Barrieren ist ein Teilbereich der Information Retrieval-Forschung, der letztlich dem Zweck dient, den Informationsfluss zum Nutzer zu verbessern. Die Verfahren zur Anfragemodifikation, die diesen Informationsfluss wesentlich gestalten und bestimmen, sollen im Folgenden in ihren verschiedenen methodischen Ansätzen dargestellt werden. Dazu wird einleitend eine Typisierung von Suchanfragen und Anfragemodifikationen gegeben, sowie eine knappe Charakterisierung der klassischen Retrievalmodelle. Die theoretische Darstellung der Modifikationsverfahren wird im Anschluss durch Anwendungsbeispiele illustriert.

---

[1] Vgl. Engelbert, Heinz: Der Informationsbedarf in der Wissenschaft, 1976, S. X

2. Typisierung von Suchanfragen und Anfragemodifikationen

In diesem Kapitel wird der Blick auf verschiedene Typen von Suchanfragen und Anfragemodifikationen gelenkt. Um deutlich zu machen, wie ein Informationsbedarf, eine Suchanfrage und deren Modifikation aufeinander aufbauen, wird zunächst das Verhältnis zwischen diesen drei Teilprozessen eines Suchprozesses dargestellt. Von besonderem Interesse ist dabei die Frage, wann und warum Anfragemodifikationen im Laufe eines Suchprozesses notwendig werden können. Schließlich wird der Begriff der Anfragemodifikation auf die Bedeutung festgelegt, die im Kontext dieser Arbeit maßgeblich sein soll und eine sich daraus ergebende Typisierung von Modifikationsverfahren vorgestellt.

2.1 Verhältnis zwischen Informationsbedarf und Suchanfrage

Das Formulieren eines Informationsbedarfs und einer Suchanfrage sind zwei verschiedene intellektuelle Leistungen, die getrennt voneinander betrachtet werden müssen. Die Befriedigung eines Informationsbedarfs besteht zunächst darin, Klarheit über ein bestimmtes Objekt oder Phänomen zu erlangen[2]. Häufig ist jedoch im Voraus nicht bekannt, welche Informationen im Einzelnen dazu benötigt werden. Dies macht die Überführung eines Informationsbedarfs in eine Suchanfrage schwierig. Frants et al. erklären dieses Problem folgendermaßen:

„We want information, we want it very much [...] but unfortunately we do not always know exactly what we want. This occurs because we only imagine the final product, the result of our physiological desire, and do not imagine exactly what information can lead to this result [...].“[3]

Demnach ergibt sich die Schwierigkeit, nach den jeweils geeigneten Informationen zu suchen daraus, dass der Mensch (im Folgenden Nutzer genannt) auf die übergeordnete Bedürfnisbefriedigung fixiert ist und es ihm schwerfällt, sich eine Vorstellung von den einzelnen Informationen zu machen, die zur Lösung seines eigentlichen Problems zusammengetragen werden müssen. Ob und unter welcher intellektuellen Anstrengung es dem Nutzer

---

[2] Vgl. Frants, Valery; Shapiro, Jacob; Voiskunskii, Vladimir: Automated Information Retrieval. Theory and Methods, 1997, S. 36

[3] Ebd.

gelingt, sich über den Informationsbedarf bewusst zu werden und in Form einer Suchanfrage zu formulieren, wird unter anderem bestimmt durch sein kognitives Modell. Dazu zählen beispielsweise individuelle Vorkenntnisse, die Wahrnehmung der Außenwelt, der sozio-ökonomische Hintergrund und Sprachkenntnisse[4].

Ebenso entscheidend ist die thematische Kompexität und Abgrenzbarkeit des Informationsbedarfs. Ein Informationsbedarf, der durch die Ermittlung eines einfachen Faktums befriedigt werden kann, bedarf einer geringeren intellektuellen Anstrengung, als ein Informationsbedarf, der sich auf einen komplexen Sachverhalt bezieht[5].

Dies wird anschaulich, wenn dazu entsprechende Fragestellungen konstruiert werden. Beispielsweise bringt ein Bibliotheksbesucher, der sich nach den Öffnungszeiten der Bibliothek erkundigt, damit einen eher einfachen Informationsbedarf zum Ausdruck. Dagegen verfügt der Bibliotheksbesucher, der nach Interpretationsmöglichkeiten eines bestimmten belletristischen Werkes verlangt, über einen komplexeren Informationsbedarf[6].

Anhand der beiden Beispiele zeichnen sich zwei Typen von Fragen ab, die im bibliothekarischen Auskunftsdienst als Faktenfragen und Sachverhaltsfragen bezeichnet werden[7]. Im Umfeld des Information Retrieval ist diese Unterscheidung nicht üblich. Zwar unterscheidet das Webretrieval zwischen Navigations-, Informations- und Transaktionsfragen[8], jedoch lässt sich damit der Zusammenhang zwischen Informationsbedarf und Suchanfrage nicht deutlich machen. Dieser Typisierung folgend, wären beide oben genannten Beispiele, trotz ihrer Verschiedenheit den Informationsfragen zuzuordnen.

Eine Ursache für diese abweichenden Sichtweisen liegt darin begründet, dass im Information Retrieval seit der Entwicklung intuitiv bedienbarer, grafischer Benutzeroberflächen und der zunehmenden Verfügbarkeit von

---

[4] Vgl. Stock, Wolfgang G.: Information Retrieval. Informationen suchen und finden, 2007, S. 57
[5] Vgl. Frants; Shapiro; Voiskunskii: Automated Information Retrieval, S. 37
[6] Für weitere Beispielfragen zur Unterscheidung von Informationsbedarfen vgl. Stock: Information Retrieval, S. 51
[7] Vgl. Plassmann, Engelbert; Rösch, Hermann; Seefeldt, Jürgen; Umlauf, Konrad: Bibliotheken und Informationsgesellschaft in Deutschland. Eine Einführung, 2006, S. 198
[8] zitiert nach Stock: Information Retrieval, S. 407

Informationssystemen durch das World Wide Web davon ausgegangen wird, dass ein Nutzer eine Suchanfrage unmittelbar an das Retrievalsystem richtet. Dagegen teilt der Nutzer eines Auskunftsdienstes seinen Informationsbedarf einem Auskunftsbibliothekar mit, der diesen Bedarf in eine Suchanfrage umformuliert und schließlich auch die Suche durchführt. Dieses Prinzip, das von einem Vermittler zwischen Nutzer und Retrievalsystem ausgeht, entspricht auch der früheren Sichtweise im Information Retrieval. Lancaster bezeichnet das Problem der Anfrageformulierung wie folgt:

„[...] the person needing information must convey his need to a member of staff of the information center by telephone, letter or personal visit.“[9]

Obwohl in dieser Arbeit von der zeitgemäßen Situation ausgegangen wird, in der eine unmittelbare Interaktion zwischen Nutzer und Retrievalsystem stattfindet, soll im Folgenden die eigentlich nicht adäquate Unterscheidung zwischen Faktenfragen und Sachverhaltsfragen beibehalten werden, da sie mit der sich nun anschließenden Unterscheidung der Informationsbedarfe besser korrespondiert, als der rein aktionsorientierte Ansatz der Navigations-, Informations- und Transaktionsfragen.

Die Unterscheidung der Informationsbedarfe ist unproblematischer. Diese werden nach einer einschlägigen Einteilung in „konkreten Informationsbedarf“ und „problemorientierten Informationsbedarf“ unterschieden[10].

2.1.1 Konkreter Informationsbedarf

Der konkrete Informationsbedarf führt zu einer Faktenfrage und ist thematisch eindeutig abgrenzbar. Die Suchanfrage lässt sich durch exakte Terme formulieren, die mit dem Informationsbedarf inhaltlich korrespondieren. Die Faktenfrage kann durch Ermittlung des entsprechenden Faktums erschöpfend beantwortet werden, der Informationsbedarf ist damit befriedigt. Dabei besteht das Suchergebnis entweder aus einer Dokumentenmenge, aus der die

---

[9] Lancaster, Frederick W.: Information Retrieval Systems: Characteristics, Testing and Evaluation, 1979, S. 147

[10] Vgl. Frants, Valery; Brush, Craig B.: The need for information and some aspects of information retrieval systems construction. In: Journal of the American Society for Information Science 39(1988)2, S. 86-91 und Frants; Shapiro; Voiskunskii: Automated Information Retrieval, S. 37-38

Fakteninformation entnommen werden muss, oder das Informationssystem bietet einen elaborierteren Ansatz, beispielsweise ein Frage-Antwort-System. In solchen Systemen ist es möglich, durch das Retrieval einzelner Textpassagen die Fakteninformation als Antwort zu übermitteln[11]. Dies wird realisiert durch die intellektuelle Faktenextraktion aus den Dokumenten während des Indexierens, oder durch automatisierte Verfahren, die sich derzeit allerdings noch in einem experimentellen Stadium befinden[12].

2.1.2 Problemorientierter Informationsbedarf

Der problemorientierte Informationsbedarf führt zu einer Sachverhaltsfrage und ist thematisch nicht eindeutig abgrenzbar. Die Suchanfrage lässt sich durch verschiedene terminologische Varianten formulieren, die mit dem Informationsbedarf nicht übereinstimmen müssen. Das Suchergebnis besteht aus einer Dokumentenmenge, durch die der Informationsbedarf noch nicht befriedigt wird, denn dazu ist eine inhaltliche Auseinandersetzung mit den Dokumenten notwendig. Liegt das Suchergebnis vor, kann zunächst lediglich eine Relevanzeinschätzung vorgenommen werden, um solche Dokumente zu selektieren, in denen man Informationen vermutet, die zur Befriedigung des Informationsbedarfs relevant sein könnten. Schließlich kann sich bei der Auswertung der Informationen der Informationsbedarf selbst verändern, so dass wieder eine neue Suchanfrage formuliert werden muss[13].

Die „Fachgruppe Information Retrieval“ der „Gesellschaft für Informatik“ spricht in diesem Zusammenhang von „vagen Anfragen“:

„Vage Anfragen sind dadurch gekennzeichnet, dass die Antwort a priori nicht eindeutig ist. Hierzu zählen neben Fragen mit unscharfen Kriterien insbesondere auch solche, die nur im Dialog iterativ durch Reformulierung (in Abhängigkeit von den bisherigen Systemantworten) beantwortet werden

---

[11] Vgl. Stock: Information Retrieval, S. 501-502

[12] Vgl. ebd.

[13] Zur Unterscheidung zwischen konkretem und problemorientierten Informationsbedarf vgl. Frants; Shapiro; Voiskunskii: Automated Information Retrieval, S. 36-40 und Stock: Information Retrieval, S. 51-52

können; häufig müssen zudem mehrere Datenbasen zur Beantwortung einer einzelnen Anfrage durchsucht werden."[14]

2.2 Konzept der Anfragemodifikation

Wenn ein Suchprozess erfolgreich verlaufen soll, muss die Anfrage mit dem Informationsbedarf möglichst deckungsgleich sein. Je größer die Abweichung zwischen der Anfrage und dem Informationsbedarf ist, desto erfolgloser wird die Suche verlaufen[15]. Mit dieser Differenz nimmt auch die Wahrscheinlichkeit zu, dass eine Anfragemodifikation notwendig wird. Dabei sind zwei theoretische Szenarien möglich: entweder wird die Anfrage im Verhältnis zum Informationsbedarf zu allgemein oder zu spezifisch formuliert.

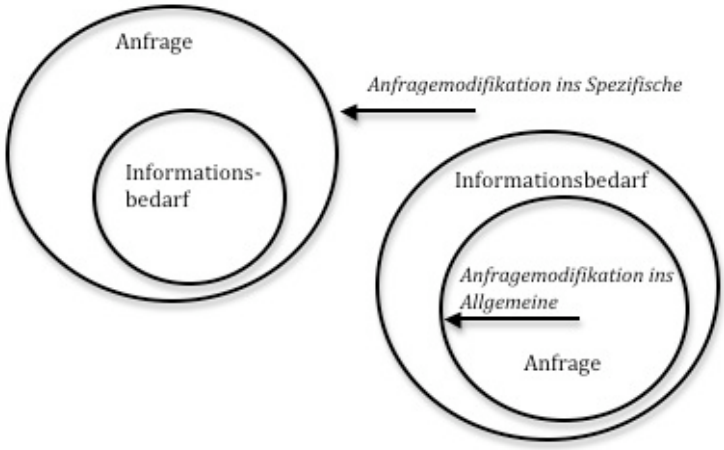

Abbildung 1: Verhältnis zwischen Informationsbedarf und Anfrage
Quelle: Lancaster: Information Retrieval Systems, S. 148-149. Die Abbildung wurde verändert.

Das erste Szenario liefert eine zu große Ergebnismenge, die tatsächlich relevanten Dokumente, die sich mit dem Informationsbedarf decken, müssen aus der Menge der irrelevanten Dokumente herausgefiltert werden. Dazu bedarf es einer Anfragemodifikation ins Spezifische, um die Precision zu verbessern. Die Spezifizierung der Anfrage kann dabei sowohl auf begrifflicher wie auf terminologischer Ebene erfolgen. Eine terminologische Spezifizierung wäre beispielsweise die Disambiguierung eines homonymen Suchterms. Eine begriffliche Spezifizierung entspräche dem Austausch eines allgemeinen Oberbegriffs gegen einen spezifischeren Unterbegriff.

---

[14] zitiert nach Fuhr, Norbert: Information Retrieval. Skriptum zur Vorlesung im SS 2006, S. 5-6
[15] Vgl. Lancaster: Information Retrieval Systems, S. 147

Beim zweiten Szenario ist der Informationsbedarf größer als die dazugehörige Anfrage. Die Ergebnismenge fällt zu gering aus, viele relevante Dokumente wurden nicht gefunden. Um den Recall zu verbessern, geht die Modifikation der Anfrage nun ins Allgemeine. Hier kann umgekehrt der Unterbegriff gegen den allgemeinen Oberbegriff getauscht werden, oder auf terminologischer Ebene eine mögliche Synonymie eines Suchterms berücksichtigt werden.

Wie Abbildung 1 zeigt, kann eine Anfragemodifikation sowohl als Precision-steigernde Maßnahme, wie auch als Recall-steigernde Maßnahme durchgeführt werden, wobei das Hinzufügen wie auch das Entfernen von Suchtermen in die bzw. aus der Anfrage notwendig sein kann. Die Entscheidung, einen Suchterm gegen einen anderen zu tauschen, weil er allgemeiner oder spezifischer ist, oder weil er für terminologische Klarheit sorgt, bezeichnet Bates als „Taktik“[16]. Bates hat 29 solcher Taktiken entworfen, jede Taktik ist ein Instrument, das auf ein bestimmtes Suchproblem angewendet werden kann. Sie systematisiert der Funktion nach die Taktiken in die vier Gruppen „monitoring“, „file structure“, „search formulation“ und „term tactics“, zur letzteren Gruppe gehören beispielsweise die Taktiken „super“ und „sub“, die durch Einbeziehung eines Ober- bzw. Unterbegriffs eine Suche wie in Abbildung 1 dargestellt steuern können. Die Verwendung mehrerer synonymer Suchterme entspricht der Taktik „parallel“, die Disambiguierung bezeichnet Bates als „pinpoint“, beide Vorgehensweisen sind den „search formulation“-Taktiken zugeordnet[17].

Die unter „monitoring“ zusammengestellten Taktiken sind Kontroll-mechanismen, die sicherstellen, dass eine modifizierte Anfrage den Nutzer tatsächlich näher an den Informationsbedarf führt und nicht etwa noch weiter abdriftet[18]. Die „file structure“-Taktiken zielen schließlich auf den Zugang und die Arbeit an den jeweiligen Informationsmitteln ab. Ein Informationsbedarf kann aufgrund seiner Komplexität möglicherweise nur durch die gezielte Suche in verschiedenen Datenbanken befriedigt werden, so dass die Anfrage auf einfachere Teilfragen heruntergebrochen werden muss, die dann den

---

[16] Bates, Marcia J.: Information search tactics. In: Journal of the American Society for Information Science 30(1979)4, S. 208

[17] Vgl. ebd.

[18] In der englischsprachigen Literatur wird dieses Problem allgemein als „query drift“ bezeichnet.

entsprechenden Informationsmitteln zugeführt werden. Diese Taktik nennt Bates „select“[19].

Anhand solcher Taktiken lässt sich die Anfragemodifikation in ein Konzept bringen, dass auf der Beschreibung des Suchverhaltens des Nutzers basiert.

2.2.1 Beschreibung des Suchverhaltens

Wie im Abschnitt 2.1.2 bereits dargestellt wurde, treten Anfragen mit unscharfen Kriterien vor allem im Zusammenhang mit einem problemorientierten Informationsbedarf auf. Demnach ist dann die Vagheit der Anfrage die Ursache für eine Modifikation, mit dem Ziel, eine größere Menge relevanter Dokumente zu finden.

Dagegen lassen sich Antworten auf Faktenfragen, die aus einem konkreten Informationsbedarf hervorgehen, durchaus bereits vor der Suche eindeutig definieren. Solche Anfragen sind von dem Problem der Vagheit nicht betroffen. In diesem Zusammenhang haben Anfragemodifikationen eher den Charakter einer Fehlerkorrektur, beispielsweise wenn eine Suche ergebnislos verläuft, weil die Anfrage so formuliert wurde, dass sie durch das Retrievalsystem nicht richtig verarbeitet oder interpretiert werden konnte. Der Nutzer ist dann gezwungen, die Anfrage solange zu modifizieren, bis ein Suchergebnis vorliegt, durch das sich das gewünschte Faktum ermitteln lässt. Doch die Korrektur eines falsch geschriebenen Suchterms oder einer falsch eingesetzten boole’schen Verknüpfung bedarf keiner Taktik. Bates bezeichnet dies als bloße „Suchschritte“, als kleinste beschreibbare Einheit eines Suchverhaltens[20].

Eine Taktik kann aus einer Abfolge von Suchschritten bestehen, ein Suchschritt ist demnach unterhalb der Taktik anzusiedeln. Komplexere Nutzerüberlegungen, die unter Umständen aus der Abarbeitung mehrerer Suchschritte und Taktiken bestehen, werden als „Strategeme“ bezeichnet. Bates charakterisiert ein Strategem als Teilprozess des gesamten Suchprozesses, beispielsweise die Entscheidung, ein bestimmtes Informationsmittel systematisch zu durchsuchen, unter Anwendung einer Reihe

---

[19] Vgl. Bates: Information search tactics, S. 208

[20] Bates, Marcia J.: Where should the person stop and the information search interface start? In: Information Processing & Management 26(1990)5, S. 578

geeigneter Taktiken. Oberhalb des Strategems folgt schließlich die „Strategie“ als übergeordneter Plan, durch den sich die Suche umfassend beschreiben lässt[21].

Für einen problemorientierten Informationsbedarf ist die Anfragemodifikation auf jeder der vier Ebenen des Suchverhaltens möglich. Die Korrektur einer unplausiblen Eingabe erfordert einen neuen Suchschritt, schlechte Ergebnismengen erfordern neue Taktiken und Strategeme. Es ist allerdings ebenso vorstellbar, dass der Nutzer durch die initiale Anfrage bereits eine Menge relevanter Dokumente findet, die ihm als ausreichend erscheint und kein Bedarf an einer Anfragemodifikation besteht, selbst wenn es sicher ist, dass noch nicht alle relevanten Dokumente gefunden wurden. Er würde damit einer Suchstrategie folgen, die keinen Wert auf einen hohen Recall legt, ein Nutzerverhalten, das mitunter sogar als Regelfall angenommen wird[22]. Der klassische Anfragedialog, der sich aus der linearen Abfolge von initialer Anfrage, Systemantwort und Anfragemodifikation zusammensetzt, wird durch solche Annahmen in seiner praktischen Bedeutung abgeschwächt.

In ihrem „Berrypicking“-Modell stellt Bates sogar die These auf, dass Nutzer häufig mit einem spezifischen Ausschnitt ihres Informationsbedarfs eine Suche beginnen und bereits nach der Sichtung der ersten gefundenen Dokumente, neue Informationen als Impulse und Ideen für gänzlich neue Suchanläufe in verschiedenen Quellen nutzen[23]. Dabei geht es nicht mehr darum, eine einzelne Anfrage so zu modifizieren, dass eine zufriedenstellende Ergebnismenge erzielt wird. Vielmehr entsprechen die verschiedenen Suchanläufe einzelnen Stationen, die zurückgelegt werden auf einer nicht vorhersagbaren Strecke, die über verschiedene Quellen führt. Dabei sind der Informationsbedarf und die sich daraus ableitende Suchanfrage einem ständigen Wandel unterworfen, der sich aus einer permanenten Verlagerung der Interessen ergibt[24].

---

[21] Vgl. ebd.
[22] Vgl. Harman, Donna: Relevance feedback and other query modification techniques. In: Frakes, William (Hrsg.): Information Retrieval. Data Structures & Algorithms, 1992, S. 241
[23] Bates, Marcia J.: The design of browsing and berrypicking techniques for the online search interfaces. In: Online Review 13(1989)5, S. 409-410
[24] Vgl. ebd.

Dagegen vereinfacht Efthimiades den Suchprozess bewusst auf nur zwei Schritte, nämlich den der Formulierung der initialen Anfrage und den der Reformulierung der Anfrage[25]. Durch die Formulierung der initialen Anfrage wird eine verbindliche Strategie festgelegt, die in der anschließenden Reformulierung unter Berücksichtigung der bis dahin erhaltenen Ergebnisse nur noch angepasst wird.

Die Vermutung, dass der von Bates beschriebene Ansatz nur einer Aneinanderreihung verschiedener Suchprozesse gleichkommt, die im einzelnen wieder einem Anfragedialog entsprechen, wie Efthimiades ihn beschreibt, wird von Bates bestritten und darauf verwiesen, dass das „Berrypicking"-Modell keine statischen Anfragen und allgemeingültigen Suchstrategien kennt[26]. Vielmehr seien Anfragen dynamisch, sie würden sich im Laufe eines Suchprozesses weiterentwickeln, anstatt einer gründlichen Auswertung einer Ergebnismenge würden Informationen eher punktuell entnommen, um auf dieser Grundlage die Suchstrategie immer wieder zu verändern und auf verschiedene Quellen anzuwenden[27].

Der Hypertext scheint Bates für dieses Nutzerverhalten wie geschaffen[28]. Durch die netzartige Struktur, die durch Verknüpfungen („Hyperlinks") hergestellt wird, lässt sich die im „Berrypicking"-Modell beschriebene, über mehrere Stationen verlaufende Suche komfortabel realisieren.

Es stellt sich in diesem Zusammenhang die Frage, inwieweit das Navigieren im Hypertext, das mittlerweile zu einem selbstverständlichen Standard geworden ist, Funktionen der Anfragemodifikation im klassischen Anfragedialog ersetzt hat. In vielen bibliografischen Datenbanken oder Online-Katalogen ist es beispielsweise üblich, in der sogenannten Vollanzeige eines bestimmten Werkes, die Verfasserangabe durch einen Hyperlink zu hinterlegen, so dass dieser Verknüpfung nur gefolgt werden muss, um alle nachgewiesenen Titel des Verfassers in Listenform zu sehen. Ohne eine solche Verknüpfung müsste

---

[25] Efthimiades, Efthimis: Query expansion. In: Williams, Martha (Hrsg.): Annual review of information science and technology 31(1996), S. 122

[26] Bates: The design of browsing and berrypicking techniques for the online search interfaces, S. 411-412 und S. 413

[27] Vgl. ebd., S. 421

[28] Ebd.

zum Anfragedialog zurückgekehrt und eine Anfragemodifikation durchgeführt werden. Eine Veränderung des Informationsbedarfs kann dabei nicht unterstellt werden, denn die Annahme, dass die übrigen Werke des Verfassers ähnlich sind zu dem zuvor gefundenen Werk, ist nicht abwegig[29].

Das Springen im Hypertext zwischen verlinkten und damit in irgendeiner Beziehung stehenden Inhalten ist ein iterierter Vorgang, der, ebenso wie die Anfragemodifikation, auf die Befriedigung eines Informationsbedarfs abzielt. Doch anstatt des Schemas des linearen Anfragedialogs, steht hinter der Navigation eher das Leitbild der Wissenserkundung[30]. Die Wissenserkundung geht einher mit der Aufwertung des Hypertextes durch besondere Visualisierungs- und Interaktionstechniken, die allesamt zu weitläufigem Navigieren anregen sollen[31]. Dabei ist es naheliegend, dass bei einer allzu weitläufigen Navigation, die über mehrere Verknüpfungen zu immer neuen Suchstationen gelangt, der Suchprozess schließlich eine eigene Dynamik im Sinne des „Berrypicking“ annehmen und sich soweit von der initialen Anfrage entfernen wird, dass die Analogie zwischen linearem Anfragedialog und Wissenserkundung nicht mehr aufrechterhalten werden kann. Wann sich dieser Übergang vollzieht, hängt vor allem davon ab, wie präzise der Nutzer seinen ursprünglichen Informationsbedarf verfolgt oder „erkundet“.

Bereits Anfang der 1990er Jahre, als der Hypertext nocht nicht mit der heutigen Selbstverständlichkeit genutzt wurde, hat Kuhlen das Navigieren im Hypertext als ein Browsing bezeichnet, das bestenfalls in eine nicht klar abzugrenzende Richtung weist, aber nicht zielgerichtet ist. Kuhlen unterscheidet präziser zwischen dem gerichteten Browsing mit Mitnahmeeffekt, dem gerichteten Browsing mit Serendipity-Effekt, dem ungerichteten Browsing und dem assoziativen Browsing[32].

---

[29] Vgl. ebd., S. 412

[30] Vgl. Gödert, Winfried: Navigation und Konzepte für ein interaktives Retrieval im OPAC. Oder: von der Informationserschliessung zur Wissenserkundung. In: AKMB-news 10(2004)1, S. 28

[31] Vgl. ebd. und Trunk, Daniela: Semantische Netze in Informationssystemen. Verbesserung der Suche durch Interaktion und Visualisierung, 2005, S. 27-40

[32] Kuhlen, Rainer: Hypertext. Ein nicht-lineares Medium zwischen Buch und Wissensbank, 1991, S. 128-132

Legt man diese Abstufung als Orientierung zugrunde, dann würde sich bereits dass gerichtete Browsing mit Serendipity-Effekt von der initialen Anfrage soweit entfernen, dass dies nicht mehr als Modifikation, sondern als Verlagerung des Informationsbedarfs interpretiert werden muss.

Abgesehen von den Arbeiten Bates' waren diese Überlegungen in der technisch dominierten Information Retrieval-Forschung bisher von nachgeordnetem Interesse. In diesem Kontext wird häufig von einem linearen Anfragedialog ausgegangen, in dem Anfragemodifikationen als intellektuelle, automatisierte oder halbautomatisierte (interaktive) Verfahren verstanden werden, mit dem Ziel, ein vorliegendes Suchergebnis im Hinblick auf die Precision oder den Recall zu verbessern. Da in dieser Arbeit konkrete Verfahren benannt und vorgestellt werden sollen, muss für eine kohärente Darstellung diesem technisch bestimmten Konzept der Anfragemodifikation gefolgt werden. Mögliche Defizite bei der Entwicklung von Retrievalsystemen, die auf diese Einseitigkeit zurückzuführen sind, können im Rahmen der Arbeit nicht behandelt werden, stattdessen wird die Notwendigkeit einer stärker interdisziplinär ausgerichteten Perspektive auf das Information Retrieval in der abschließenden Betrachtung in Abschnitt 7 herausgestellt.

2.2.2 Typisierung von Modifikationsverfahren

In der Interaktion zwischen Nutzer und Retrievalsystem bestehen für den Ablauf einer Anfragemodifikation vier theoretische Möglichkeiten: entweder wird die Modifikation durch den Nutzer manuell durchgeführt, oder automatisch durch das Retrievalsystem, oder durch den Nutzer mit Unterstützung des Retrievalsystems, oder durch das Retrievalsystem mit Unterstützung durch den Nutzer[33].

Daraus ergeben sich drei Verfahrensklassen der Anfragemodifikation, nämlich intellektuelle Verfahren (die allein durch den Nutzer durchgeführt werden), automatische Verfahren (die allein durch das Retrievalsystem durchgeführt werden) und interaktive Verfahren (die im Dialog zwischen Nutzer und Retrievalsystem durchgeführt werden).

---

[33] Vgl. Efthimiades: Query expansion, S. 122

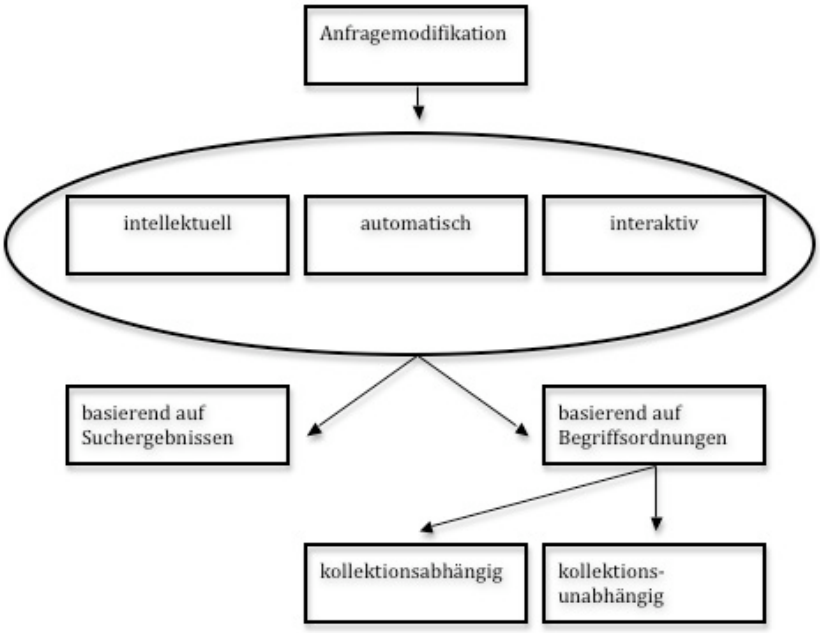

Abbildung 2: Verfahrensklassen der Anfragemodifikation
Quelle: Efthimiades: Query expansion, S. 124. Die Abbildung wurde verändert.

In Abschnitt 2.2 wurde bereits auf das Konzept der Anfragemodifikation eingegangen, das die initiale Anfrage um neue Suchterme erweitert oder bereits verwendete Suchterme entfernt. Abbildung 2 verweist zunächst auf die Frage, wie neue Suchterme gewonnen werden sollen und nennt zwei mögliche Quellen: aus den Dokumenten, die als Suchergebnis der initialen Anfrage vorliegen oder aus Begriffsordnungen[34] anhand derer die Dokumente deskribiert wurden.

Auf die Dokumente, die als Suchergebnis vorliegen, stützen sich Verfahren wie das Relevance Feedback (vgl. Abschnitt 5.3) oder informetrische Analysen (vgl. Abschnitt 5.4). Dagegen sind die aus Begriffsordnungen gewonnenen Terme unabhängig vom Suchergebnis, stattdessen muss unterschieden werden zwischen kollektionsabhängigen und kollektionsunabhängigen Begriffs-ordnungen (vgl. Abschnitt 5.2).

Insgesamt ergeben sich für die Gestaltung von automatischen und interaktiven Verfahren viele Möglichkeiten: neue Suchterme können im Rahmen eines interaktiven Verfahrens in ihrem themenspezifischen Zusammenhang

[34] Efthimiades spricht von „knowledge structures“. Da Dokumentationssprachen Instrumente der Inhaltserschließung sind und kein Wissen strukturieren und repräsentieren, wurde „Begriffsordnung“ als neutralere Benennung vorgezogen. So auch bei Stock: Information Retrieval, S. 477

dargestellt und zur Anfrageerweiterung vorgeschlagen oder durch ein automatisches Verfahren hinzugefügt werden. Gleiches gilt für das Entfernen von ursprünglichen Suchtermen.

Dabei ist die Abgrenzung zwischen automatischen und interaktiven Verfahren problematisch, da häufig nur in der Dialoggestaltung Unterschiede bestehen, während die theoretischen Hintergründe, beispielsweise die Techniken zur Gewinnung neuer Suchterme, unverändert sind. Wird der Nutzer in den Modifikationsprozess einbezogen, etwa explizit, indem ihm neue Suchterme vorgeschlagen werden, oder implizit, durch Auswahl relevanter Dokumente beim Relevance Feedback, wird von einem interaktiven Verfahren gesprochen. Dagegen handelt es sich um ein automatisches Verfahren, wenn das Retrievalsystem die Modifikation völlig selbstständig durchführt. Die intellektuellen Verfahren stehen eindeutig abseits von den beiden anderen Verfahrensklassen: hier muss der Nutzer jeden einzelnen Such- und Modifikationsschritt selbstständig planen und durchführen.

Ruthven weist jedoch darauf hin, dass grundsätzlich jedes Retrievalsystem als ein interaktives System verstanden werden kann, da schließlich jedes Retrievalsystem dem Nutzer irgendeine Form von Interaktion abverlangen muss. Um eine Unterscheidung zwischen der „gewöhnlichen“ Interaktivität, die zur Bedienung eines Computerprogramms in jedem Fall notwendig ist, und der „fortgeschrittenen“ Interaktivität als Idee zur Dialoggestaltung im Information Retrieval vorzunehmen, spricht Ruthven nur dann von „interaktivem Information Retrieval“, wenn das jeweilige System ausdrücklich im Zusammenhang mit der Entwicklung neuer Interaktionsformen, der Evaluation bestehender Interaktionsformen oder der Erforschung von nutzertypischen Interaktionsformen eingesetzt wird[35].

Dieser Sichtweise folgend, würde es sich um „automatisches Information Retrieval“ oder „automatische Verfahren zur Anfragemodifikation“ handeln, wenn der Aspekt der Automatisierung im Zentrum des Interesses stünde. Da sich jedoch einzelne, interaktive Schritte unter Umständen auch automatisieren lassen, ohne dass deshalb aber von einem „automatischen Information

[35] Ruthven, Ian: Interactive information retrieval. In: Cronin, Blaise (Hrsg.): Annual Review of Information Science and Technology 42(2008), S. 45

Retrieval“ gesprochen werden könnte, ist es nicht möglich den Zustand eindeutig zu benennen, an dem der Qualitätssprung von einem interaktiven zu einem automatischen Retrievalsystem – oder umgekehrt – stattfindet. Da kein Kriterium zur trennscharfen Unterscheidung gefunden werden kann, werden automatische und interaktive Modifikationsverfahren in Abschnitt 5 dieser Arbeit gemeinsam behandelt.

3. Retrievalmodelle

Die in Abbildung 2 eingeführten Verfahrensklassen sind in der Praxis nicht selbstverständlich gegeben. Sie sind abhängig vom jeweiligen Retrievalsystem, von der zugrundeliegenden Datenbasis, sowie deren Indexierung. Fakten- und Information Retrieval gehen dabei unterschiedlich vor. Die Unterschiede werden in diesem Abschnitt dargelegt, gefolgt von einer Charakterisierung der beiden grundlegenden Modelle des Information Retrieval, dem Vektorraummodell und dem probabilistischen Modell. Zuvor wird die verwendete Terminologie erläutert.

Datenbasis

Als Datenbasis werden die Informationen bezeichnet, die in einer Datenbank in maschinenlesbarer Form abgelegt sind[36]. Dabei werden „Informationen“ in dieser Arbeit einschränkend als Texte verstanden, auf audielle oder visuelle Informationen wird nicht eingegangen. Je nach Schema der Datenbank, ist die Datenbasis unterschiedlich stark oder schwach in einzelne Felder strukturiert.

Dokumentdatei, *invertierte Datei*

In der Dokumentdatei sind die Dokumentationseinheiten gespeichert, die die einzelnen Objekte oder Datensätze der Datenbank darstellen und die häufig vereinfachend als „Dokumente“ bezeichnet werden. Da die Datenbasis strukturiert ist, lassen sich die Dokumente innerhalb der Dokumentdatei anhand der Feldinhalte sequentiell anordnen, in einer bibliografischen Datenbank beispielsweise alphabetisch nach einem Feld, das Verfassernamen enthält. Soll auf ein bestimmtes Dokument zugegriffen werden, dann muss ebenfalls sequentiell Dokument für Dokument geprüft werden. Dies kann einen erheblichen Aufwand an Suchschritten bedeuten und die Suche sehr zeitintensiv machen. Retrievalsysteme arbeiten deshalb häufig mit invertierten Dateien, sogenannten Indizes. Die invertierte Datei bzw. der Index ordnet bestimmte Feldinhalte, meist in alphabetischer Reihenfolge. Bei der Eingabe eines Suchbegriffs wird der entsprechende Anfangsbuchstabe im Index

[36] Vgl. Datenbasis. In: Kuhlen, Rainer; Seeger, Thomas; Strauch, Dietmar (Hrsg.): Grundlagen der praktischen Information und Dokumentation. Band 2: Glossar, 2004, S. 22

aufgesucht, so dass sich die aufwendige sequentielle Suche nur noch auf das jeweilige Indexsegment beschränkt. Ist der Suchbegriff innerhalb des Indexsegments gefunden, wird von dort aus auf alle Datensätze verwiesen, in denen der Begriff enthalten ist. Durch diese direkten Verweise ist die sequentielle Ordnung innerhalb der Dokumentdatei dann nicht mehr nötig[37].

Für das Retrieval werden in der Regel mehrere Indizes angeboten, beispielsweise für Verfassernamen, Sachtitel, Schlagwörter oder Abstracts. Als „Basic Index" bezeichnet man die Kombination solcher Felder in einem übergreifenden Index, der eine feldunabhängige Suche ermöglicht.

Informationslinguistik

Die Informationslinguistik entwickelt Verfahren für die Verarbeitung natürlicher Sprache. Die Vielfalt der natürlichen Sprache lässt bei der Indexierung, wie auch bei der Formulierung von Suchanfragen viele Varianten zu, das Retrieval kann aber nur dann erfolgreich sein, wenn Anfrage- und Indexierungsvokabular übereinstimmen. Informationslinguistische Verfahren versuchen eine Unabhängigkeit von der jeweiligen sprachlichen Ausdrucksform herzustellen, damit auch bei einer unterschiedlichen Darstellung eines Sachverhalts in Suchanfrage und Index, das jeweils Gemeinte gefunden wird. Die Informationslinguistik kann dabei von zwei Richtungen wirken: entweder werden die verschiedenen sprachlichen Varianten beim Indexieren auf einen zentralen Indexterm abgebildet, oder eine Anfrage wird beim Retrieval um die verschiedenen sprachlichen Varianten erweitert[38]. Das Letztere ist ein Fall von Anfragemodifikation, dem in den Abschnitten 5.1 und 5.2 nachgegangen wird.

Retrievalmodell

Durch das Retrievalmodell wird bestimmt, auf welche Weise aus einer Dokumentenkollektion die Antwortdokumente auf eine erfolgte Suchanfrage hin ermittelt werden. Jedes Modell folgt einer bestimmten Methode zur Ermittlung des Retrievalstatuswerts, mit dem die Übereinstimmung eines Dokuments mit der Suchanfrage ausgedrückt wird und anhand dessen sich das einzelne

---

[37] Vgl. Salton, Gerard; McGill, Michael: Information Retrieval. Grundlegendes für Informationswissenschaftler, 1987, S. 14-22

[38] Vgl. Nohr, Holger: Grundlagen der automatischen Indexierung. Ein Lehrbuch, 2005, S. 58

Dokument in eine Rangfolge aller Antwortdokumente einordnen lässt[39]. Dagegen soll hier von einem *Retrievalsystem* gesprochen werden, wenn eine konkrete Anwendung gemeint ist, die auf einem theoretischen Retrievalmodell basiert.

Textstatistik

Im Kontext des Information Retrieval ist die Textstatistik ein Hilfsmittel, mit dem sich anhand statistischer Häufigkeitsmerkmale Aussagen über die inhaltliche Bedeutung einzelner Wörter innerhalb eines Textes ableiten lassen. Die theoretischen Grundlagen dieser Verfahren gehen zurück auf die These des Informatikers Hans Peter Luhn, die besagt, dass die Bedeutung von Texten durch die Signifikanz der darin vorkommenden Wörter statistisch ermittelbar ist[40]. Luhns These basiert auf dem, von dem Sprachwissenschaftler George Zipf erbrachten Nachweis, dass eine konstante Beziehung zwischen der Auftretenshäufigkeit von Wörtern in Texten und dem Rang der Wörter in einer Häufigkeitsliste besteht[41].

Dieser statistische Zusammenhang, der als „Zipfsches Gesetz“ bezeichnet wird, führte mit der These von Luhn zu der Annahme, dass nicht alle Wörter, die in einem Dokument vorkommen, als Indexterme gleichermaßen geeignet sind und dass nicht alle Wörter, die als Indexterme in Frage kommen, hinsichtlich ihrer inhaltlichen Bedeutung die gleiche Wertigkeit besitzen[42]. Die individuelle Wertigkeit eines Indexterms wird deshalb durch Gewichtungen ausgedrückt. Man spricht auch von „entscheidungsstarken“ Indextermen, womit solche Terme gemeint sind, die geeignet sind, um im Retrievalprozess relevante Dokumente zu selektieren und irrelevante Dokumente zurückzuweisen. Geht man von einer Termhäufigkeitsverteilung aus, die in einen hohen, einen mittleren und einen niedrigen Frequenzbereich unterteilt ist, dann wird den Termen des mittleren Frequenzbereichs eine hohe Entscheidungsstärke unterstellt. Niedrigfrequente Terme gelten für den Dokumenteninhalt als zu

---

[39] Vgl. Fuhr, Norbert: Theorie des Information Retrieval I: Modelle. In: Kuhlen, Rainer; Seeger, Thomas; Strauch, Dietmar (Hrsg.): Grundlagen der praktischen Information und Dokumentation. Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis, 2004, S. 207

[40] Vgl. Nohr: Grundlagen der automatischen Indexierung, S. 43

[41] Vgl. ebd., S. 44

[42] Vgl. ebd.

wenig signifikant, hochfrequente Terme sind meist beudeutungsschwache Wortklassen wie Artikel, Pronomen oder Adverbien[43].

Für die Gewichtung eines Terms müssen zwei Maßzahlen berücksichtigt werden: die Häufigkeit, mit der ein Term in einem Dokument auftritt und die Häufigkeit, mit der ein Term in der gesamten Dokumentenkollektion auftritt. Es wird also unterschieden zwischen einer dokumentbezogenen und einer kollektionsbezogenen Termfrequenz, wobei der Termfrequenzansatz besagt, dass ein Indexterm umso aussagefähiger für den Inhalt eines Dokuments ist, je häufiger er in einem Dokument auftritt und je seltener er überhaupt vorkommt[44].

Für die Ermittlung der dokumentbezogenen Termfrequenz bestehen verschiedene Varianten. Ein primitiver Ansatz ist die Bestimmung der absoluten Häufigkeit mit der ein Term in einem Dokument auftritt. Dies führt allerdings zu einer Bevorzugung von längeren Dokumenten, da mit der Länge des Dokuments auch die absolute Häufigkeit der Terme ansteigt. Es wurden deshalb verschiedene Varianten einer relativen Termfrequenz entwickelt. Als TF (term frequency) werden solche Varianten bezeichnet, die die absolute Häufigkeit eines Terms ins Verhältnis setzen zur Gesamtzahl aller Wörter des Dokuments und gegebenenfalls weitere, frei einstellbare Faktoren vorsehen, um die Bedeutung der dokumentbezogenen Termfrequenz insgesamt zu steuern. Als WDF (within-document frequency) wird eine Variante bezeichnet, die das Verhältnis zwischen Term und Dokument ebenfalls relativiert und zusätzlich mit logarithmischen Werten arbeitet. Dadurch können die Termgewichte kleiner und besser interpretierbar gehalten werden[45].

Liegt eine stärker strukturierte Datenbasis vor, wie etwa innerhalb einer bibliografischen Datenbank, dann bietet es sich zudem an, das TF-Gewicht mit positionsspezifischen Faktoren zu verbinden. Damit kann beispielsweise die Bedeutung von Termen aus der Titel- oder Schlagwortkategorie der Datenbank bestimmt werden.

---

[43] Vgl. ebd., S. 47
[44] Vgl. ebd., S. 45
[45] Vgl. Stock: Information Retrieval, S. 321-324

Zur Berechnung der kollektionsbezogenen Termfrequenz hat sich das Verfahren der inversen Dokumenthäufigkeit (IDF) durchgesetzt. Die IDF-Gewichtung beruht wiederum auf dem Termfrequenzansatz, also auf der Annahme, dass die Bedeutung eines Terms proportional zu seiner Auftretenshäufigkeit in einem bestimmten Dokument ist, aber umgekehrt proportional zu seiner Auftretenshäufigkeit in der gesamten Kollektion. Werden beide Werte zueinander ins Verhältnis gesetzt, ergibt sich das IDF-Gewicht[46].

Sind die beiden Einzelschritte zur Ermittlung der dokument- und kollektionsbezogenen Termfrequenz vollzogen, lässt sich der Gewichtungswert eines Terms schließlich durch Multiplikation beider Werte berechnen. Dies wird in der vereinfachenden Schreibweise TF*IDF ausgedrückt, wobei TF für eines der verschiedenen Verfahren der TF-Gewichtung steht.

Die Gewichtungswerte der einzelnen Terme werden schließlich in der invertierten Datei abgelegt, so dass nach der Eingabe einer Suchanfrage die Gewichte der Suchterme ausgelesen werden können. Anhand dieser Informationen wird der Retrievalstatuswert der verschiedenen Dokumente in Abhängigkeit zur Suchanfrage berechnet[47].

Für die Gestaltung von Anfragemodifikationen nach Abbildung 2 lassen sich folgende Mindestanforderungen aufstellen:

Anfragemodifikationen, die auf Begriffsordnungen basieren, sind grundsätzlich immer möglich (beispielsweise für die Formulierung einer Stichwortsuche unter Nutzung eines natürlichsprachlichen Thesaurus). Anhand einer Dokumentationssprache kann dann modifiziert werden, wenn die Dokumentenkollektion durch dieses Vokabular tatsächlich inhaltlich erschlossen und indexiert wurde. Außerdem kann mit Begriffsordnungen gearbeitet werden, wenn eine Datenbasis vorliegt, die umfangreich genug ist, um durch textstatistische Verfahren deskribierende Daten zu gewinnen und zu einem Vokabular zu strukturieren (beispielsweise zu einem Ähnlichkeitsthesaurus oder statistischem Thesaurus).

---

[46] Vgl. Nohr: Grundlagen der automatischen Indexierung, S. 46-47
[47] Vgl. Stock: Information Retrieval, S. 327

Anfragemodifikationen, die auf Suchergebnissen basieren sind dann möglich, wenn ein Retrievalsystem eingesetzt wird, das für jedes Dokument einen individuellen Retrievalstatuswert in Abhängigkeit zur Anfrage ermitteln kann und damit einen Anhaltspunkt liefert, welche Bereiche einer Ergebnismenge für eine Anfragemodifikation beachtet werden sollten.

Wie dieser Vergleichsprozess zwischen Anfrage und Dokument im Einzelnen funktioniert, hängt vom jeweiligen Retrievalmodell ab. Die grundlegenden Retrievalmodelle, aus denen sich schließlich auch ein jeweils eigenes Verständnis der Anfragemodifikation ergibt, sollen deshalb genauer betrachtet werden.

3.1 Faktenretrieval

Bisher wurde verallgemeinernd von Information Retrieval gesprochen, an dieser Stelle muss eine differenziertere Betrachtung eingeführt werden. Auch wenn der Begriff des Information Retrieval häufig als Oberbegriff für sämtliche Retrievalverfahren verwendet wird, grenzt van Rijsbergen Faktenretrieval und Information Retrieval im engeren Sinn streng voneinander ab:

	Faktenretrieval	*Information Retrieval*
Matching	exakt	partiell, best match
Inferenz	Deduktion	Induktion
Modell	deterministisch	probabilistisch
Klassifikation	monothetisch	polythetisch
Anfragesprache	formal	natürlich
Fragespezifikation	vollständig	unvollständig
Gesuchte Objekte	die Fragespezifität erfüllende	relevante
Reaktion auf Datenfehler	sensitiv	insensitiv

Tabelle 1: Abgrenzung zwischen Faktenretrieval und Information Retrieval
Quelle: van Rijsbergen: Information Retrieval, S. 1, zitiert nach Fuhr: Information Retrieval, S. 6

Das Faktenretrieval basiert auf einem „exact match“-Prinzip und prüft das Vorhandensein der durch die Suchanfrage definierten Zeichenketten ab. Dagegen folgt das Information Retrieval im engeren Sinn dem „partial match“-Prinzip und sucht nach Dokumenten, die mit der Suchanfrage wenigstens

teilweise übereinstimmen, wobei die am besten übereinstimmenden Dokumente („best match“) als besonders relevant interpretiert werden[48].

Das Faktenretrieval arbeitet nach einem dichotomen Prinzip: ein Dokument bzw. eine Zeichenkette wird gefunden, wenn eine exakte Übereinstimmung mit der Suchanfrage besteht, andernfalls wird das Dokument nicht gefunden. Alle gefundenen Dokumente gelten als gleichermaßen relevant, alle nicht gefundenen als irrelevant. Für Gewichtungen von Termen oder Dokumenten bleibt dabei kein Raum.

Das Information Retrieval im engeren Sinn wird von van Rijsbergen als probabilistisch charakterisiert, im Gegensatz zum deterministisch ausgerichteten Faktenretrieval. Ferber beschreibt diesen Unterschied anschaulich, wenn er darauf hinweist, dass das Information Retrieval im Gegensatz zum Faktenretrieval „immer nur nach den relativ besten Lösungen und Antworten suchen kann und im Allgemeinen keine eindeutige beste Lösung existiert“[49].

Während beim Faktenretrieval alle Objekte, die das gleiche Attribut aufweisen, zu einer Klasse zusammengefasst werden (monothetische Klassifikation), ist es im Information Retrieval sinnvoller, solche Objekte zusammenzufassen, die nur über einen Teil aller Attribute einer Klasse verfügen (polythetische Klassifikation)[50]. Ein Objekt ist dabei ein Dokument, unter einem Attribut ist in diesem Zusammenhang ein bestimmtes Merkmal zu verstehen, beispielsweise das Auftreten eines bestimmten Terms in einem Text oder in einem Feld eines Datensatzes.

Die Anfragesprache im Faktenretrieval hat einen formalen Charakter, sie wird aus einem vorgegebenen Vokabular unter Berücksichtigung einer Syntax so zusammengesetzt, dass sie den Informationsbedarf vollständig abdeckt. In der Regel arbeitet das Faktenretrieval mit einer boole’schen Verknüpfungslogik.

---

[48] Vgl. van Rijsbergen, Cornelis J.: Information Retrieval, 1979, S. 1. Online: http://www.dcs.gla.ac.uk/Keith/Preface.html [Abrufdatum: 05.06.2008, Datei: Chapter1.pdf]
[49] Ferber, Reginald: Information Retrieval. Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web, 2003, S. 185
[50] Vgl. van Rijsbergen: Information Retrieval, S. 2

Bei der Eingabe eines komplexen Sucharguments, das Suchterme mit mehreren boole'schen Operatoren verknüpft, werden die Terme zunächst in den invertierten Dateien gesucht. Anschließend wird für jeden Term eine Verweisliste auf die jeweiligen Dokumente erstellt. Schließlich werden die Verweislisten den Operatoren der Suchanfrage entsprechend miteinander verknüpft. Die Dokumente, auf die die Bedingungen zutreffen, werden als Suchergebnis unsortiert ausgegeben[51].

Diese Vorgehensweise bietet zwar einen schnellen Zugriff auf die Indizes, andererseits macht es das Faktenretrieval fehleranfälliger, da bereits ein syntaktischer Fehler bei der Eingabe einer Anfrage dazu führt, dass das eigentlich Gemeinte nicht gefunden wird[52]. Dagegen liefert das Information Retrieval immer die relativ besten Antworten zur Anfrage.

Legt man die Abgrenzung von van Rijsbergen zugrunde, dann sind sämtliche Verfahren der Anfragemodifikation, die in irgendeiner Weise mit der Textstatistik und Gewichtungswerten arbeiten, von vornherein auf das Information Retrieval im engeren Sinn festgelegt. Tatsächlich lassen sich Retrievalsysteme aus der Praxis diesen Idealtypen nur schwierig zuordnen. Als „erweiterte boole'sche Systeme" bezeichnet man solche Modelle, bei denen einerseits boole'sche Operatoren zur Anfrageformulierung geboten werden, andererseits aber auch Gewichtungen möglich sind[53]. Fuhr plädiert deshalb dafür, die Gegensätze aus Tabelle 1 als Endpunkte kontinuierlicher Skalen aufzufassen, auf denen es viele mögliche Zwischenlösungen gibt[54].

3.2 Information Retrieval

Bis heute ist das Information Retrieval vor allem durch zwei klassische und in gewisser Weise auch konkurrierende Modelle geprägt: das Vektorraummodell und das probabilistische Modell. Beiden Modellen gemein ist eine intensive Auseinandersetzung und Weiterentwicklung in der Theorie und in experimentellen Umgebungen, der jedoch eine vergleichsweise geringe Bedeutung in der Praxis gegenübersteht.

---

[51] Vgl. Salton; McGill: Information Retrieval, S. 125-126
[52] Vgl. ebd.
[53] Vgl. Stock: Information Retrieval, S. 185-200
[54] Fuhr: Information Retrieval, S. 6

3.2.1 Vektorraummodell

Das Vektorraummodell beruht auf einer geometrischen Sicht auf das Information Retrieval und wurde in den 1960er und 1970er Jahren von dem Informationswissenschaftler und Informatiker Gerard Salton entwickelt.

Dieses Modell arbeitet nicht mit invertierten Dateien, sondern mit einem Vektorraum, in dem sowohl die Anfrage, wie auch die Dokumente als Vektoren dargestellt und durch ein Ähnlichkeitsmaß die Übereinstimmung der einzelnen Dokumentvektoren zum Anfragevektor ermittelt werden. Je näher ein Dokumentvektor am Anfragevektor liegt, als desto relevanter wird er angenommen und im anschließenden Relevance Ranking entsprechend hoch platziert.

Dazu wird die Dokumentenkollektion in einer Matrix erfasst, deren Koordinaten aus den Dokumenten und den darin enthaltenen Termen bestehen. Es lässt sich dann, unter Berücksichtigung der durch die Textstatistik erzeugten Termgewichtungen der Punkt im Vektorraum ermitteln, an dem der jeweilige Dokumentvektor sich ausrichtet. Ebenso wird mit der Anfrage verfahren, durch deren einzelne, gegebenenfalls individuell gewichtete Terme sich die Position des Anfragevektors bestimmen lässt. Die Terme spannen den Vektorraum auf und bestimmen dessen Dimensionalität[55].

Besondere Auswirkungen auf die Retrievalergebnisse hat das verwendete Ähnlichkeitsmaß, durch das Dokument- und Anfragevektor miteinander verglichen werden und woraus die Relevanz des Dokuments zur Anfrage abgeleitet wird. Ein einfaches Maß ist das Skalarprodukt, bei dem die Koordinaten der Vektoren komponentenweise multipliziert und anschließend addiert werden. Das Produkt der beiden Vektorskalare ist vollständig, wenn beide Vektoren genau aufeinanderliegen und nimmt immer weiter ab, je größer der Winkel zwischen beiden Vektoren ist. Stehen die Vektoren senkrecht aufeinander, nimmt das Skalarprodukt den Wert 0 an. In diesem Fall gilt das Dokument als vollständig irrelevant. Ferber weist auf die Gefahr von Verzerrungen hin, die sich aus der Linearität des Skalarprodukts ergeben und dazu führen, dass längere Dokumente bei einer zugrundeliegenden

[55] Vgl. Salton; McGill: Information Retrieval, S. 128-129

Termgewichtung nach Häufigkeit, unverhältnismäßig hohe Ähnlichkeitswerte erhalten. Ein anderer verzerrender Effekt kann eintreten, wenn alle indexierten Dokumente aus formalen Gründen die gleiche Länge haben, unabhängig von ihrer inhaltlichen Wertigkeit[56]. Um diesen Effekten entgegenzuwirken, gibt es zahlreiche alternative Ähnlichkeitsmaße zum Skalarprodukt, die für Verzerrungen weniger anfällig sind, beispielsweise das Cosinus-Maß, das Dice-Maß oder das Jaccard-Maß. Ferber gibt einen Überblick über alternative Ähnlichkeitsmaße und betont, dass das geeignete Maß für eine bestimmte Dokumentenkollektion nur durch empirische Untersuchungen zu ermitteln ist[57].

Da jeder Term durch eine Dimension im Vektorraum repräsentiert wird und die einzelnen Dimensionen voneinander unabhängig sind, folgt daraus auch die Unabhängigkeit der Terme. Dokumente werden im Vektorraum als lineare Kombination ihrer Terme dargestellt, ohne das ein Zusammenhang zu Termen hergestellt wird, die semantisch ähnlich oder sogar identisch sind zu den Dokumenttermen. Informationslinguistische Verfahren können dieses Problem entschärfen, indem zumindest bedeutungsgleiche Terme zu einem Indexterm zusammengeführt werden. Eine komplexere Lösung für dieses Problem ist das „Latent Semantic Indexing“, ein Verfahren, das unter Abschnitt 5.2.3 behandelt wird.

3.2.2 Probabilistisches Modell

Die älteste Idee eines Retrievalmodells, in dem die Relevanzeinschätzung erstmals nicht vollständig auf den Nutzer verlagert, sondern mit systemseitiger Unterstützung durchgeführt wird, ist das probabilistische Modell[58], das auch den in Tabelle 1 genannten Eigenschaften des Information Retrieval entspricht. Die folgende Darstellung bezieht sich auf eine häufig als „klassisch“ bezeichnete Variante, die 1976 von Robertson und Sparck-Jones eingeführt wurde[59].

---

[56] Ferber: Information Retrieval, S. 73

[57] Ebd., S. 72-80

[58] Der erste Artikel, der die Möglichkeiten eines probabilistischen Retrievals diskutiert, stammt von Maron und Kuhns aus dem Jahr 1960. Vgl. dazu Stock: Information Retrieval, S. 354

[59] Robertson, Stephen; Sparck-Jones, Karen: Relevance weighting of search terms. In: Journal of the American Society for Information Science and Technology 27(1976)3, S. 132

Dem probabilistischen Modell liegt die Einsicht zugrunde, dass kein Retrievalsystem mit Sicherheit vorhersagen kann, welche Dokumente ein Nutzer als relevant einstufen wird. Es wird deshalb mit Wahrscheinlichkeiten gearbeitet[60]. Wie beim Faktenretrieval, wird für die Relevanz eine dichotome Ausprägung angenommen, das heißt, es liegt entweder Relevanz oder Irrelevanz vor. Die Leistung des Modells besteht nun in einem probabilistischen Relevance Ranking der Suchergebnisse, das sich allerdings allein auf den Grad der Wahrscheinlichkeit bezieht, mit dem ein Dokument relevant ist und die Ergebnismenge dementsprechend nach absteigender Wahrscheinlichkeit anordnet. Eine darüber hinausgehende Einschätzung der Relevanz als solches findet nicht statt[61].

Realisiert wird das probabilistische Modell auf der Grundlage der Wahrscheinlichkeitstheorie. Das Retrievalsystem muss eine Einschätzung der Relevanz eines Dokuments zu einer gegebenen Anfrage vornehmen. Dieser Ausdruck wird häufig als $P_q = (rel \mid x)$ dargestellt, wobei q für die Anfrage und x für das Dokument steht. Wird die Irrelevanz eines Dokuments zu einer Anfrage als $P_q = (\overline{rel} \mid x)$ dargestellt, dann kann ein Dokument als relevant angenommen werden, wenn die Wahrscheinlichkeit der Relevanz größer ist als die Wahrscheinlichkeit der Irrelevanz. Es werden dann nur solche Dokumente als Ergebnis ausgegeben, für die der Zusammenhang $P_q = (rel \mid x) > P_q = (\overline{rel} \mid x)$ gilt. Dieses Prinzip kann durch Schwellenwerte ergänzt werden, wenn beispielsweise nur solche Dokumente berücksichtigt werden sollen, die mit einer bestimmten Wahrscheinlichkeit relevant sind. In diesem Fall müsste die Differenz zwischen beiden Wahrscheinlichkeitswerten größer sein, als der jeweilige Schwellenwert. Beide Wahrscheinlichkeitswerte können durch das Bayestheorem ermittelt werden, der Retrievalstatuswert eines Dokuments ergibt sich dann aus dem Quotienten beide Werte[62].

Das entscheidende Kriterium, das zu einer hohen Relevanzwahrscheinlichkeit führt, ist die Übereinstimmung zwischen Dokumenteninhalt und Anfrage.

---

[60] Vgl. Robertson, Stephen: The probability ranking principle in IR. In: Sparck-Jones, Karen; Willett, Peter (Hrsg.): Readings in Information Retrieval, 1997, S. 281

[61] Vgl. ebd.

[62] Vgl. Ruthven, Ian; Lalmas, Mounia: A survey on the use of relevance feedback for information access systems. In: The Knowledge Egineering Review 18(2003)2, S. 102

Anders als im Vektorraummodell, findet die Ähnlichkeitsmessung zwischen Dokument und Anfrage nicht anhand von Repräsentanten wie etwa Vektoren statt, sondern durch den Abgleich zwischen den Indextermen eines Dokuments und den Suchtermen der Anfrage. Es wird dabei vereinfachend angenommen, dass die Verteilung der Indexterme unabhängig ist und somit beispielsweise kein Zusammenhang zwischen dem gemeinsamen Auftreten zweier Terme in einem Dokument besteht. Wie die Textstatistik lehrt, ist diese Annahme unrealistisch, jedoch lässt sich die Komplexität der notwendigen Berechnung dadurch reduzieren. Robertson und Sparck-Jones haben zwei Unabhängigkeitsannahmen formuliert[63]:

- die Verteilung der Terme in den relevanten Dokumenten ist ebenso unabhängig, wie die Verteilung der Terme in allen Dokumenten, oder
- die Verteilung der Terme in den relevanten Dokumenten ist ebenso unabhängig, wie die Verteilung der Terme in den irrelevanten Dokumenten.

Im ersten Fall wird die Wahrscheinlichkeit, mit der ein Term in einem relevanten Dokument auftritt verglichen mit der Wahrscheinlichkeit, mit der er in der gesamten Kollektion auftritt. Im zweiten Fall wird die Wahrscheinlichkeit, mit der ein Term in einem relevanten Dokument auftritt verglichen mit der Wahrscheinlichkeit mit der er in den irrelevanten Dokumenten auftritt.

Für das Ranking der Dokumente anhand ihrer Relevanzwahrscheinlichkeit zu einer gegebenen Anfrage lassen sich ebenfalls zwei Strategien unterscheiden. Entweder wird die wahrscheinliche Relevanz allein abhängig gemacht vom Auftreten der Suchterme in den Dokumenten oder vom Auftreten und Nicht-Auftreten der Suchterme in den Dokumenten[64].

Werden diese beiden Ranking-Prinzipien in Beziehung gesetzt zu den beiden Unabhängigkeitsannahmen, so ergeben sich vier Kombinationsmöglichkeiten, also vier verschiedene Funktionsweisen der Termgewichtung im probabilistischen Retrieval, die sich mengentheoretisch wie folgt herleiten lassen:

---

[63] Robertson; Sparck-Jones: Relevance weighting of search terms, S. 132
[64] Vgl. ebd., S. 133

		Relevanz		
		Dokument ist relevant	Dokument ist irrelevant	
Indexierung	Term t ist enthalten	r	$n - r$	n
	Term t ist nicht enthalten	$R - r$	$N - n - R + r$	$N - n$
		R	$N - R$	

Tabelle 2: Kontingenztabelle für Term t
Quelle: Robertson; Sparck-Jones: Relevance weighting of search terms, S. 131

Dabei entspricht r der Anzahl der relevanten Dokumente, die Term t enthalten, n der Anzahl aller Dokumente, die Term t enthalten, R der Anzahl aller relevanten Dokumente zu einer gegebenen Anfrage und N der Anzahl aller Dokumente der Kollektion. Aus den in Tabelle 2 dargestellten Mengen haben Robertson und Sparck-Jones vier verschiedene Funktionen zur Termgewichtung entwickelt, auf die hier nicht im Detail eingegangen werden soll. Die Gewichtungsfunktion, die in einem Testverfahren die besten Ergebnisse lieferte, wird in Abschnitt 5.2.2 im Zusammenhang mit der Gestaltung des Relevance Feedback aufgegriffen.

Das Relevance Feedback hat im probabilistischen Retrieval eine besondere Bedeutung. Da die Mengen R und r nicht bekannt sind, müssen sie geschätzt werden. Dazu bedarf es Informationen des Nutzers in Form von Relevanzbeurteilungen. Dieses Feedback bezüglich der Relevanz von Dokumenten ist aber nur dann möglich, wenn bereits ein erstes Suchergebnis vorliegt. Es wird deshalb für den Ablauf der initialen Anfrage vereinfachend angenommen, dass alle Terme mit der gleichen Wahrscheinlichkeit in einem Dokument aus der Menge R oder aus der Menge $N - R$ vorkommen und das alle Dokumente, in denen ein eingegebener Suchterm nicht vorkommt, zunächst der Menge $N - R$ zuzuordnen sind[65]. Anschließend wird das Suchergebnis durch Feedback-Verfahren schrittweise verbessert.

[65] Vgl. Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier: Modern Information Retrieval, 1999, S. 32-33

Zwar wurde das probabilistische Modell in der Theorie des Information Retrieval intensiv diskutiert, dennoch konnte keine Verbesserung der Retrievalleistung gegenüber anderen Modellen nachgewiesen werden. Das probabilistische Modell wurde fast ausschließlich in experimentellen Umgebungen eingesetzt und ist in der Praxis noch weniger präsent als das Vektorraummodell[66].

[66] Vgl. Lewandowski, Dirk: Web Information Retrieval. Technologien zur Informationssuche im Internet, 2005, S. 86

4. Intellektuelle Modifikationsverfahren

Die intellektuellen Verfahren der Anfragemodifikation grenzen sich von den beiden anderen in Abbildung 2 eingeführten Verfahrensklassen insofern ab, als sie sich allein durch das Suchverhalten des Nutzers beschreiben lassen – die Modifikation findet schließlich rein nutzerseitig statt. Dazu muss an dieser Stelle nochmals auf die von Bates formulierte Unterscheidung zwischen der Suchstrategie und den kleinteiligeren Strategemen, Taktiken und Suchschritten verwiesen werden.

Die im Folgenden darzustellenden, intellektuellen Modifikationsverfahren, die dem Bibliothekswissenschaftler Charles Bourne zugeschrieben werden, werden von Bates als Suchstrategien interpretiert:

„For example the ‚building-block' approach by Charles Bourne [...] is a strategy. [...] This strategy constrasts with Bourne's ‚citation pearl-growing' strategy [...]."[67]

Harter, auf dessen Ausführungen sich die Darstellung der Blockstrategie und der „Citation Pearl Growing"-Strategie in diesem Abschnitt stützt, folgt der Einschätzung von Bates und unterscheidet weiterhin:

„A search strategy is an overall plan or approach for a search problem, while a search tactic or heuristic is a move made to advance a particular strategy."[68]

Dabei ist es offensichtlich, dass sich zwar jede Anfrage nutzerseitig beliebig modifizieren lässt, dass dies jedoch nicht immer vor dem Hintergrund einer durchdachten Strategie geschieht, sondern häufig willkürlich oder einer spontan getroffenen Entscheidung folgend, also eher im Sinne des „Berrypicking"-Modells.

Die von Harter zusammengestellten Heuristiken entsprechen im Wesentlichen Bates' Suchtaktiken, die bereits im Abschnitt 2.2 angesprochen wurden. Neben einer Reihe von Recall- oder Precision-steigernden Maßnahmen (etwa: Muss

---

[67] Bates, Marcia J.: Search techniques. In: Williams, Martha E. (Hrsg.): Annual Review of Information Science and Technology 16(1981), S. 143

[68] Harter, Stephen: Online Information Retrieval. Concepty, Principles, and Techniques, 1986, S. 170

ein Suchterm disambiguiert werden oder muss eine Synonymie berücksichtigt werden? Ist die Anfrageformulierung über- oder unterspezifiziert?) nennt Harter auch „Personal Heuristics“, die dem Nutzer zu Flexibilität und der Bereitschaft zur iterierten Suche raten:

„Be flexible; stay loose; be willing to look at a search in more than one way.“[69]

Tatsächlich ist die Bereitschaft zur Wiederholung der Suche und zur Erprobung neuer Formulierungen die Grundvoraussetzung, um Anfragemodifikationen im Sinne umfassender Suchstrategien, wie der Blockstrategie oder der „Citation Pearl Growing“-Strategie vorzunehmen. Beide Konzepte sind arbeits- und zeitintensiv, allerdings nutzen sie die logische Klarheit der boole’schen Operatoren, die, wenn ihre Bedeutung verstanden wurde, die präzise Formulierung und Steuerung einer Suche ermöglichen.

4.1 Blockstrategie

Die Blockstrategie gilt als die am weitesten verbreitete und eingesetzte Suchstrategie[70]. In der englischsprachigen Literatur ist sie bekannt als „building block search“, im deutschsprachigen Raum wird sie gelegentlich auch als „Komponentenzerlegung“ bezeichnet[71].

Zunächst wird der vorliegende Informationsbedarf auf seine einzelnen inhaltlichen Facetten (bzw. Blöcke oder Komponenten) hin analysiert und in diese zerlegt. Anschließend werden für jede inhaltliche Facette Terme ausgewählt, die diese repräsentieren. Diese Terme können zueinander synonym oder quasi-synonym oder wenigstens semantisch ähnlich sein. Schließlich umfasst jede Facette eine Reihe von Termen, die für die Suche als äquivalent gelten sollen[72]. Die Zusammenführung der geeigneten Terme zu einer Facette ist ein aufwendiger Prozess, wenn die vorliegende Dokumentenkollektion nicht durch normiertes Vokabular erschlossen wurde. Die in Frage kommenden Terme können dann nur vermutet oder durch

---

[69] Vgl. ebd., S. 195-202
[70] Vgl. ebd., S. 172
[71] So bei Kolke, Ernst-Gerd vom: Online-Datenbanken. Systematische Einführung in die Nutzung elektronischer Fachinformation, 1996, S. 125
[72] Vgl. Harter: Online Information Retrieval, S. 172

stichprobenartige Testrecherchen ermittelt werden. Wurde die Kollektion durch einen Thesaurus erschlossen, ist besonders auf präkombinierte Deskriptoren zu achten. Es besteht dann die Gefahr, dass deren einzelne Bestandteile verschiedenen Facetten zugeordnet werden.

Sind für jede Facette die geeigneten Terme ermittelt, werden diese durch ein boole'sches ODER miteinander verknüpft, so dass jede Facette eine Vereinigungsmenge bildet.

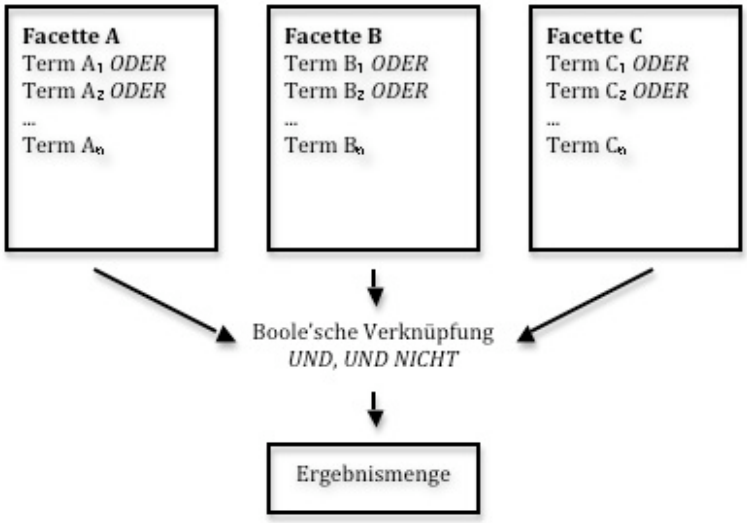

Abbildung 3: Blockstrategie
Quelle: Harter: Online Information Retrieval, S. 173. Die Abbildung wurde verändert.

Die einzelnen Facetten werden anschließend durch ein UND oder UND NICHT miteinander verknüpft und als Suchanfrage an das Retrievalsystem übergeben. Nach der Auswertung der Ergebnismenge kann die Anfrage gegebenenfalls modifiziert werden. Als Recall-steigernde Taktik kommt beispielsweise die Ergänzung der einzelnen Facetten um weitere Terme in Frage, aber auch die vollständige Entfernung einer Facette. Die Precision lässt sich steigern durch das Entfernen allgemeiner oder ambiger Terme, sowie durch den expliziten Ausschluss einer Facette durch eine UND NICHT-Verknüpfung[73].

4.2 Variationen der Blockstrategie

Die Anwendung der Blockstrategie führt zu umfangreichen Suchanfragen mit zahlreichen Verknüpfungen. Dabei steigt bei einem komplexen

[73] Vgl. ebd., S. 174-175

Informationsbedarf, der auf zahlreiche einzelne Facetten abgebildet werden muss, die Wahrscheinlichkeit, dass eine Schnittmenge aller Facetten kein Ergebnis liefert, weil ein zu hoher Grad der Spezifität erreicht wurde. In diesen Fällen bieten sich verschiedene Variationen der Blockstrategie an. Harter nennt drei Abwandlungen: „most specific concepts first“, „fewest postings first“ und „successive fractions“[74].

Diese drei Suchstrategien beginnen mit Anfragen, die auf einen höheren Recall abzielen, um die Ergebnismenge dann schrittweise zu reduzieren, bis ein für den Nutzer akzeptabler Umfang erreicht ist. Sie unterscheiden sich nur in der Art der Zusammenstellung einer ersten Facette für die initiale Anfrage. Der „most specific concept first“-Ansatz verwendet die Facette mit den spezifischsten Termen zur initialen Anfrage, „fewest postings first“ verwendet die Facette, die die kleinste Ergebnismenge liefert. In der Regel ist die Facette der spezifischsten Terme auch diejenige, die die kleinste Ergebnismenge liefert[75].

Die „successive fractions“-Strategie beginnt mit einer Facette, in der formale Kategorien der bibliografischen Beschreibung zusammengefasst werden, beispielsweise der Dokumententyp, Sprachen oder Erscheinungsjahre. Darauf aufbauend, wird die Ergebnismenge schrittweise durch nachfolgende Facetten eingeschränkt, wie bei den beiden anderen Vorgehensweisen. Der Start mit formalen Kriterien bietet sich an bei einem thematisch schwierig abzugrenzenden Informationsbedarf[76].

4.3 „Citation Pearl Growing“-Strategie

Die „Citation Pearl Growing“-Strategie, die im deutschsprachigen Raum auch als „Zitatsuche“ bekannt ist, verläuft entgegengesetzt zur den verschiedenen Variationen der Blockstrategie. Die Ausgangssituation ist ein einzelnes oder einige wenige Dokumente, die dem Informationsbedarf exakt entsprechen. Aus diesen Dokumenten werden geeignete Terme entnommen, anhand derer die Anfrage modifiziert werden kann, so dass neben den bereits bekannten

---

[74] Vgl. ebd., S. 177
[75] Vgl. ebd., S. 177-180
[76] Vgl. ebd.

Dokumenten auch neue relevante Dokumente gefunden werden. Als neue Suchterme eignen sich etwa Deskriptoren oder frei vergebene Schlagwörter und Stichwörter aus dem Dokumententitel oder Abstract. Mit den auf diese Weise gefundenen neuen Dokumenten kann ebenso verfahren werden, um die Anfrage ein weiteres Mal zu modifizieren und wieder neue Terme zu erhalten. Die Anfragemodifikationen werden systematisch durchgeführt, indem die neuen Suchterme wiederum einzelnen Facetten zugeordnet werden und innerhalb der Facetten mit einem boole'schen ODER verknüpft werden. Auf diese Weise wird der Recall schrittweise verbessert, bis ein Suchergebnis mit dem gewünschten Umfang vorliegt. Die Bezeichnung „Pearl Growing" verweist metaphorisch auf das schichtweise Wachstum einer Perle[77].

Die „Citation Pearl Growing"-Strategie bietet sich an, wenn es aufwendig ist, die in Frage kommenden Suchterme für die Blockstrategie zusammenzuführen. Dies kann der Fall sein, wenn die Dokumentenkollektion nicht durch normiertes Vokabular erschlossen wurde oder wenn die Suche sich auf ein Thema erstreckt, in dem keine verbindliche, allgemein akzeptierte Terminologie vorherrscht.

[77] Vgl. ebd., S. 183-184

5. Automatische und interaktive Modifikationsverfahren

In diesem Abschnitt werden die umfangreichen Möglichkeiten der automatischen und interaktiven Anfragemodifikation behandelt. Der Funktionsumfang der dargestellten Verfahren erstreckt sich von elementaren linguistischen Ansätzen der Worterkennung, über die vielfältigen Modelle, die das semantische Umfeld einer Anfrage erschließen und dem, auf Suchergebnissen operierenden Relevance Feedback, bis hin zur Arbeit mit Musterdokumenten, die für ein Retrieval ähnlicher Dokumente eingesetzt werden.

Durch die Beschreibung der einzelnen Modelle wird jeweils ein Überblick über die grundlegende Funktion und den theoretischen Hintergrund gegeben. Eine strikte Trennung zwischen automatisch und interaktiv arbeitenden Verfahren ist dabei nicht möglich, da in vielen Fällen beide Wege realisierbar sind und auch kombiniert werden können. Dies gilt insbesondere für die Verfahren der semantischen Umfeldsuche.

Ein erschöpfender Überblick über sämtliche erprobten und unerprobten Vorgehensweisen kann ebenfalls nicht geleistet werden. Die hier behandelten Verfahren können vielmehr als unterschiedliche „Denkrichtungen“ verstanden werden, deren Ursprung teilweise bis in die 1960er Jahre zurückreicht und in die die Information Retrieval-Forschung seitdem vorgedrungen ist.

5.1 Morphologische und syntaktische Analyse

Bedingt durch die Vielfalt der natürlichen Sprache können gleiche Sachverhalte unterschiedlich formuliert werden. Dabei ist die Verwendung der einzlnen sprachlichen Variationen unvorhersehbar, was dazu führt, dass ein Nutzer nie sicher sein kann, wie bestimmte Sachverhalte als Anfrage zu formulieren sind, um ein optimales Suchergebnis zu erhalten. Ferber bezeichnet dies als „Problem der Vergleichbarkeit von Inhalten“ und nennt zwei mögliche Herangehensweisen[78]:

[78] Ferber: Information Retrieval, S. 40 und Lewandowski: Web Information Retrieval, S. 104

- Versuche, die natürliche Sprache so zu repräsentieren und zu verarbeiten, dass inhaltliche Ähnlichkeiten erkennbar werden;
- Versuche, die zulässigen Mittel zur inhaltlichen Beschreibung so einzuschränken, dass sie Ähnlichkeiten abbilden.

Der zweite Ansatz bezieht sich auf die Erschließung durch kontrolliertes Vokabular, etwa durch Schlagwortlisten oder Dokumentationssprachen. Der erste Ansatz bezieht sich auf informationslinguistische Verfahren.

Wie bereits in Abschnitt 3 angedeutet wurde, befasst sich die Informationslinguistik mit der Entwicklung von Verfahren, durch die eine Unabhängigkeit von verschiedenen sprachlichen Ausdrucksformen gleicher Sachverhalte ermöglicht wird, sowie deren Implementierung in Informationssystemen. Letztlich soll der Nutzer in den Stand versetzt werden, seinen Informationsbedarf in eigenen Worten zu beschreiben und dem Retrievalsystem den Abgleich zwischen dem Anfragevokabular und dem Indexierungsvokabular zu überlassen. Die Aufgabe „inhaltliche Ähnlichkeiten erkennbar zu machen“ findet also systemseitig statt und meint die Ähnlichkeiten zwischen dem individuellen, durch eine Anfrage repräsentierten Wissen des Nutzers und den spezifischen Inhalten der einzelnen Dokumente.

Einen ersten, elementaren Schritt in diese Richtung bietet die Morphologie, die Lehre der inneren Struktur von Wörtern und der Bildung von Wortklassen[79]. Sowohl im Faktenretrieval, wie auch im Information Retrieval werden Terme zunächst als reine Zeichenketten erfasst. Diese werden zwar von Interpunktionszeichen, anderen Sonderzeichen und Ziffern bereinigt, so dass es sich mit einer gewissen Wahrscheinlichkeit um natürlichsprachliche Wörter handelt, dabei steht jedoch jedes Wort als individuelle Zeichenkette nur für sich selbst[80].

Um das Retrieval von der Zeichenkettenorientierung auf ein wortorientiertes Niveau anzuheben, muss die Identifikation und Zusammenführung verschiedener Wortformen möglich sein. Die Eingabe eines Suchterms in einer beliebig flektierten Form führt dann zum Retrieval aller Dokumente, in denen

---

[79] Vgl. Nohr: Grundlagen der automatischen Indexierung, S. 64
[80] Ferber: Information Retrieval, S. 37

eine Wortform enthalten ist, die mit dem Suchterm die gleiche Grundform oder Stammform teilt. Dazu werden die Terme einer morphologischen Analyse zugeführt, mit der sich die Grundform, bzw. die Stammform und die jeweiligen, flektierten Wortformen einander zuordnen lassen.

Eine morphologische Analyse kann rein algorithmisch oder wörterbuchbasiert durchgeführt werden. Die algorithmische Lösung enthält eine Reihe allgemeiner, sequentiell abzuarbeitender Regeln, welche Zeichen zu entfernen oder zu ersetzen sind, um aus einer flektierten Form eine Grundform oder Stammform zu erzeugen. Für eine morphologisch komplexe Sprache wie das Deutsche kommen diese Lösungen allerdings nicht in Frage; morphologisch komplexe Sprachen lassen sich aufgrund ihrer Unregelmäßigkeit in der Bildung der Wortformen nicht durch allgemein gültige Regeln beschreiben. Für solche Sprachen muss auf wörterbuchbasierte Verfahren zurückgegriffen werden[81]. Dabei wird mit Listen gearbeitet, deren Eingang die einzelnen, möglichen Wortformen und deren Ausgang die jeweils zu indexierende Grundform oder Stammform darstellt. Wörterbuchbasierte Verfahren sind arbeits-, zeit- und kostenintensiver als die algorithmischen Verfahren, ermöglichen aber individuelle Lösungen für sprachliche Unregelmäßigkeiten und sind daher weniger fehleranfällig[82].

In der Praxis sind zwei Realisierungen möglich: entweder werden die verschiedenen Wortformen auf einen zentralen Term, in der Regel die Grundform oder Stammform, abgebildet, wobei die Suchterme einer Anfrage auf die gleiche Weise behandelt werden[83]. Alternativ besteht die Möglichkeit, die Anfrage um alle Flexionsformen der verwendeten Suchterme anzureichern[84]. Diese Vorgehensweise geht von einer Vollformindexierung aus.

---

[81] Vgl. Nohr: Theorie des Information Retrieval II. Automatische Indexierung. In: Kuhlen, Rainer; Seeger, Thomas; Strauch, Dietmar (Hrsg.): Grundlagen der praktischen Information und Dokumentation. Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis, 2004, S. 220-221

[82] Vgl. ebd.

[83] Vgl. Ferber: Information Retrieval, S. 42

[84] So praktiziert durch die Metasuchmaschinen *LexiQuo* und *LexiLib*, vgl. 6.1, S. 73-78

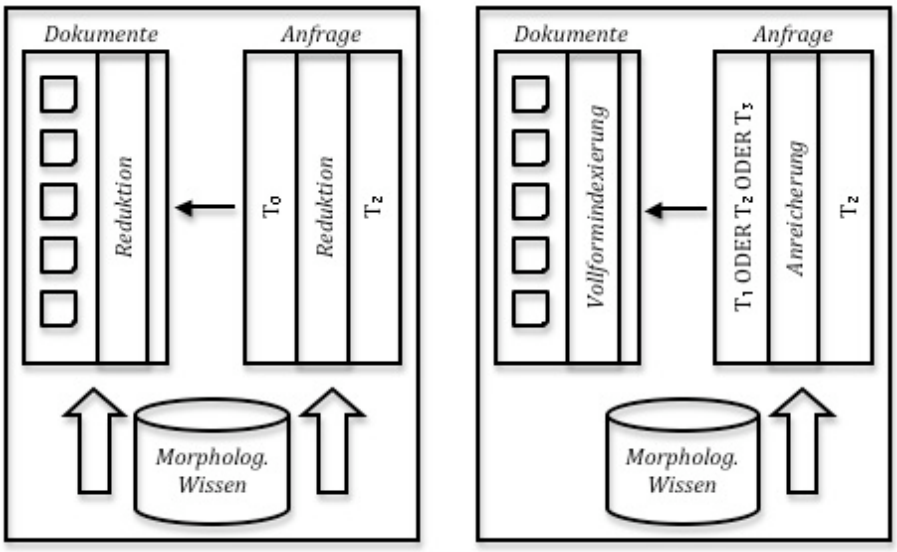

Abbildung 4: Anfragemodifikation durch informationslinguistische Verfahren
Quelle: Frei nach Ferber: Information Retrieval, S. 42

Abbildung 4 zeigt ein boole'sches Retrieval nach den Termen T_1, T_2 und T_3, die verschiedene Wortformen mit der gemeinsamen Grundform T_0 darstellen. Links ist eine Indexierung der Grundform dargestellt, daher muss die Eingabe eines Suchterms in flektierter Wortform (in diesem Fall T_2) ebenfalls auf die entsprechende Grundform T_0 reduziert werden. Sowohl für die Indexierung, wie für die Anfragemodifikation muss auf morphologisches Wissen zurückgegriffen werden. Rechts ist eine Vollformindexierung dargestellt, daher muss die Suche mit T_2 um alle weiteren, möglichen Wortformen (T_1 und T_3) angereichert werden. Algorithmische Lösungen kommen bei diesem Verfahren nicht in Frage, die Zusammenführung aller Flexionsformen anhand einer einzelnen Wortform ist nur durch wörterbuchbasierte Verfahren möglich[85]. Die Vereinigungsmenge aus T_1, T_2 und T_3 bildet die modifizierte Anfrage.

Neben der Flexion sind die Komposition und die Derivation weitere Teilaspekte der Morphologie. Die Berücksichtigung dieser Wortfügungen und -ableitungen setzt bereits auf der semantischen Ebene der Sprache an. Komposita sind für das Retrieval insofern problematisch, als ihre Verwendung nicht vorhersehbar ist. Eine Literaturrecherche zum Thema „Anfragemodifikation" kann mit ebendiesem Kompositum erfolgreich sein, aber ebenso mit den Termen „Anfrage" und „Modifikation" oder der Adjektiv-Substantiv-Verbindung

[85] Vgl. Ferber: Information Retrieval, S. 45

„modifizierte Anfrage“. Je nach dem, welche der beiden Strategien aus Abbildung 4 angewendet wird, werden sowohl das Kompositum, wie auch dessen Bestandteile indexiert, bzw. als Suchterme einer Anfrage hinzugefügt. Bei der Zerlegung eines Kompositums muss eine Überidentifizierung vermieden werden, also eine Zerlegung in Bestandteile, die zum Kompositum in keinem Bezug mehr stehen (beispielsweise die Zerlegung von „Anfrage“ in „An“ und „Frage“). Dieser Effekt kann durch einen einfachen Algorithmus vermieden werden, der eine wörterbuchbasierte Identifikation des längstmöglichen Bestandteils durchführt. Der Term wird dabei von rechts nach links eingelesen, da die hintere Konstituente eine reguläre, der jeweiligen Wortklasse entsprechende Endung aufweist, während die vordere Konstituente häufig auf eine Fugung endet, so dass die Identifikation zusätzlich erschwert würde[86]. Adjektiv-Substantiv-Verbindungen können als Wortklassen-Muster definiert und isoliert werden. Schließlich kann das Adjektiv in ein Substantiv überführt werden (etwa „modifiziert“ zu „Modifikation“), um einen substantivischen Index- oder Suchterm zu erhalten[87].

5.2 Semantische Umfeldsuche

Durch die Berücksichtigung des semantischen Umfelds eines Suchterms wird das Retrievalniveau von der Wortorientierung auf die Begriffsorientierung gesteigert. Dazu bedarf es ein über die Morphologie hinausgehendes, semantisches Wissen, welches Aspekte der Synonymie, der Homonymie, sowie hierarchische Begriffsstrukturen berücksichtigen muss. Lässt sich die Anreicherung einer Anfrage durch synonyme oder quasi-synonyme Suchterme durch Hinzuziehung eines Synonym-Wörterbuchs noch vergleichsweise einfach realisieren, wird zur Klärung von Homonymie oder der Erfassung eventueller Ober- und Unterbegriffe zu einem gegebenen Term der Rückgriff auf semantische Relationen zwischen den Termen notwendig. Um „in das

---

[86] So praktiziert bei der Software „Lingo“, vgl. Lepsky, Klaus; Vorhauer, John: Lingo – ein open source System für die Automatische Indexierung des Deutschen. In: ABI-Technik 26(2006)1, S. 18-29 und der Software „Morphy“, vgl. Lezius, Wolfgang; Rapp, Reinhard; Wettler, Manfred: A freely available morphological analyzer, disambiguator and context sensitive lemmatizer for german. In: Proceedings of the 17th International Conference on Computational Linguistics – Volume 2, 1998, S. 743-748

[87] Vgl. Lepsky; Vorhauer: Lingo – ein open source System für die Automatische Indexierung des Deutschen

semantische Umfeld einer Suchanfrage vorzustoßen“[88] bedarf es eines semantischen Netzes, anhand dessen sich eine Anfrage reformulieren lässt, entweder automatisch oder im Dialog mit dem Nutzer. Letzterer Vorgehensweise gibt Feldman den Vorzug. Als „adding the user to the system“ bezeichnet sie die Disambiguierung homonymer Suchterme und die Auswahl von Synonymen zur Erweiterung der Anfrage durch den Nutzer, da nur dieser das von ihm Gemeinte zuverlässig benennen könne[89].

Im Folgenden werden die Möglichkeiten der Anfragemodifikation anhand solcher Netze diskutiert, wobei unterschieden werden muss zwischen paradigmatisch und syntagmatisch relationierten Netzen. Paradigmatische Relationen bestehen unabhängig von konkreten Dokumentenkollektionen, die semantischen Beziehungen sind „fest verdrahtet“[90]. Paradigmatisch relationierte Netze können daher als „kollektionsunabhängige“ Begriffsordnungen bezeichnet werden. Dagegen besteht eine syntagmatische Relation zwischen zwei Termen, die gemeinsam in einem Dokument oder einer Dokumentenkollektion auftreten[91]. Syntagmatische Relationen sind kontextabhängig, bezogen auf eine Dokumentenkollektion also „kollektionsabhängig“.

Abschließend wird das Verfahren der „latent semantischen Indexierung“ vorgestellt, eine alternative Form des Vektorraummodells, die eine semantische Umfeldsuche ermöglicht und dabei völlig ohne semantisches Wissen arbeitet.

5.2.1 Kollektionsunabhängige Begriffsordnungen

Zu diesem paradigmatisch relationierten Ordnungssystemen zählen Dokumentationssprachen wie Thesauri und Klassifikationen, ebenso wie lexikalische Datenbanken, die kein dokumentationssprachliches, sondern ein natürlichsprachliches Umfeld abbilden. Die verschiedenen Systeme lassen sich

---

[88] Gödert, Winfried; Lepsky, Klaus: Semantische Umfeldsuche im Information Retrieval in Online-Katalogen, 1997, S. 9

[89] Feldman, Susan: Find what I mean, not what I say. Meaning-based search tools. In: Online 24(2000)3, S. 50 und S. 54

[90] Vgl. Stock, Wolfgang G.; Stock, Mechtild: Wissensrepräsentation. Informationen auswerten und bereitstellen, 2008, S. 68

[91] Vgl. ebd.

sowohl anhand ihres Verwendungszwecks, wie auch anhand bestimmter Merkmale voneinander unterscheiden.

Dokumentationssprachen sind ihrem Verwendungszweck nach Instrumente der Inhaltserschließung, die zum Indexieren, Speichern und Wiederauffinden von Dokumenten dienen. Die Anzahl der Relationsarten, die in Dokumentationssprachen berücksichtigt werden, ist begrenzt. Klassifikationen berücksichtigen meist nur hierarchische Relationen, da sich diese durch den hierarchischen Aufbau von selbst ergeben. Thesauri berücksichtigen neben hierarchischen Relationen auch assoziative Relationen zwischen verwandten Deskriptoren, sowie Äquivalenzrelationen zwischen Nicht-Deskriptoren und Deskriptor[92].

Dagegen erfassen lexikalische Datenbanken ein universelles, natürlich-sprachliches Vokabular, das sprachliches Wissen durch eine Vielzahl semantischer Relationen differenziert berücksichtigt.

Als Vertreter der natürlichsprachlichen Thesauri soll die lexikalische Datenbank *WordNet* am Ende dieses Abschnitts vorgestellt werden. Zunächst sollen jedoch Dokumentationssprachen am Beispiel von *UMTHES* (Umwelt-Thesaurus) verfolgt werden.

5.2.1.1 Dokumentationssprachliche Thesauri am Beispiel *UMTHES*

Der Thesaurus *UMTHES*[93] ist die Dokumentationssprache des Umweltbundesamtes (UBA), einer nachgeordneten Behörde des Bundesministeriums für Umwelt, Naturschutz und Reaktorsicherheit (BMU). Im Juni 2008 umfasst dieses umfangreiche Vokabular 10.700 Deskriptoren und 27.500 Nicht-Deskriptoren und wird zur inhaltlichen Erschließung mehrerer Literaturdatenbanken aus den Bereichen Umweltbelastung und Umweltschutz eingesetzt[94].

---

[92] Vgl. Betram, Jutta: Einführung in die inhaltliche Erschließung. Grundlagen – Methoden – Instrumente, 2005, S. 127-147

[93] Online: http://www.umweltbundesamt.de/uba-info/dokufabib/thesdownload.htm [Stand: 25.04.2007, Abrufdatum: 07.07.2008, Datei: udkalfa.pdf]

[94] Vgl. Homepage des Umweltbundesamtes. Online: http://www.umweltbundesamt.de/uba-info/dokufabib/thes.htm [Stand: Juni 2008, Abrufdatum: 07.07.2008, Datei: thes.htm]

Die Möglichkeiten, die sich durch das Ausnutzen der paradigmatischen Beziehungen eines Begriffs zu seinem semantischen Umfeld für die Anfragemodifikation ergeben, sollen anhand eines Beispiels aus dem *UMTHES*-Vokabular dargestellt werden. Dabei soll an dieser Stelle keine Einschätzung über den tatsächlichen Nutzen von erweiterten Retrievalfunktionen in *UMTHES*-Datenbanken abgegeben werden. Der Thesaurus wird hier allein aufgrund seines Umfangs und des dichten Relationsgefüges als Beispiel herangezogen.

Bei einer Suche in einer, durch diesen Thesaurus erschlossenen Dokumentenkollektion, könnte bei Verwendung des Suchterms „Verkehrsmittel“ die Anfrage durch das in Abbildung 5 dargestellte, semantische Umfeld erweitert werden.

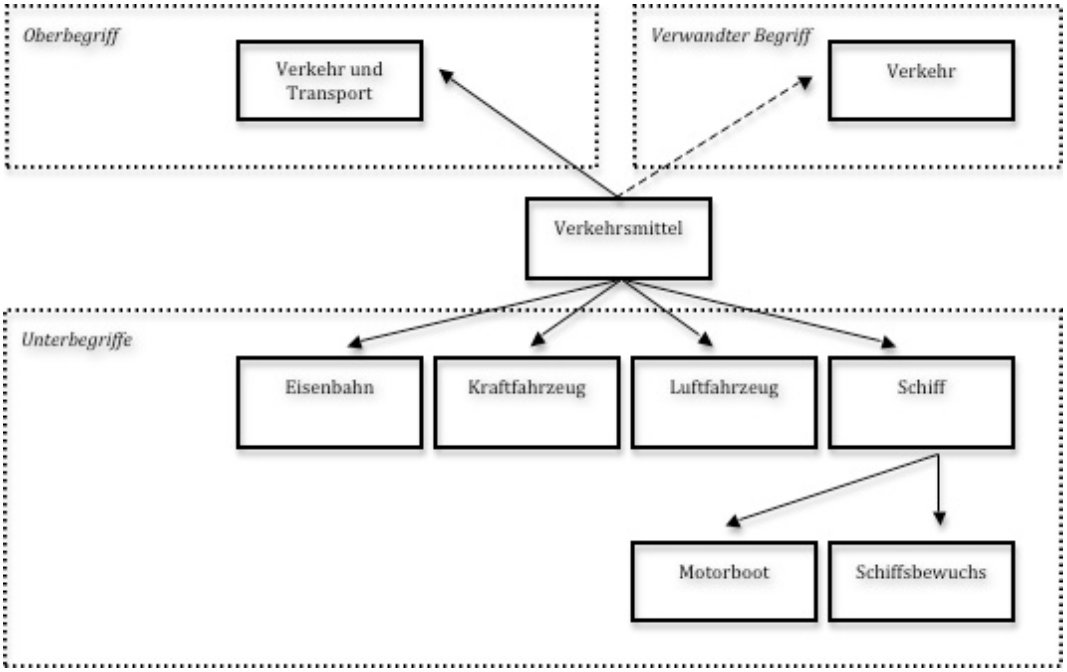

Abbildung 5: Erfassung des semantischen Umfelds
Quelle: Eigene Darstellung mit Vokabular aus dem Umwelt-Thesaurus *UMTHES*. Die Abbildung zeigt nur einen Ausschnitt aus dem Relationsgefüge des Deskriptors „Verkehrsmittel“.

Das unmittelbare semantische Umfeld des Deskriptors „Verkehrsmittel“ besteht aus acht Unterbegriffen, von denen nur vier Unterbegriffe in Abbildung 5 exemplarisch dargestellt sind, sowie über einen Oberbegriff und einen verwandten Begriff. Jeder der vier abgebildeten Unterbegriffe verfügt wiederum über weitere Unterbegriffe, dies ist in Abbildung 5 anhand des untergeordneten

Deskriptors „Schiff“ exemplarisch durch die wiederum untergeordneten Deskriptoren „Motorboot“ und „Schiffsbewuchs“ dargestellt.

Die semantische Ähnlichkeit zwischen dem Deskriptor „Verkehrsmittel“ und den zugeordneten Unterbegriffen beruht auf einer generischen Beziehung. Die Deskriptoren „Eisenbahn“, „Kraftfahrzeug“, „Luftfahrzeug“ und „Schiff“ erben als Hyponyme alle Merkmale des Oberbegriffs „Verkehrsmittel“, des Hyperonyms. Zusätzlich verfügt jedes Hyponym über mindestens ein weiteres, wesenskonstitutives Merkmal, wodurch es sich von den übrigen Deskriptoren innerhalb der Begriffsreihe abgrenzt[95]. Bei der Verwendung des Suchterms „Verkehrsmittel“ ist die Anfragemodifikation durch die Einbeziehung der unmittelbaren Unterbegriffe insofern sinnvoll, als ein gewisses Maß an semantischer Ähnlichkeit besteht.

Für den Deskriptor „Schiff“ sind die beiden Unterbegriffe „Motorboot“ und „Schiffsbewuchs“ abgebildet. Der Nutzen einer Anfragemodifikation zum Suchterm „Verkehrsmittel“ anhand dieser bereits zwei Hierarchieebenen tiefer liegenden Begriffsreihe, muss davon abhängig gemacht werden, ob Transitivität vorliegt[96]. Für den Deskriptor „Motorboot“ kann die Transitivität durch die Folgerung bestätigt werden:

„Wenn ein Motorboot ein Schiff ist und ein Schiff ein Verkehrsmittel, dann ist ein Motorboot ein Verkehrsmittel.“

Für den Deskriptor „Schiffsbewuchs“ kann dies nicht gelten. Es liegt Intransitivität vor, weil es sich um eine andere Relation handelt. Zwischen „Schiffsbewuchs“ und „Schiff“ besteht eine Meronym-Holonym-Relation.

Für die Anfragemodifikation lässt sich aus diesem Beispiel verallgemeinernd ableiten, dass das Hinzufügen neuer Suchterme zur Anfrage über eine einzelne hierarchische Ebene hinweg möglich ist, über mehrere hierarchische Ebenen dagegen nur dann, wenn Transitivität vorliegt, da von einem semantischen

[95] Vgl. Stock; Stock: Wissensrepräsentation, S. 76
[96] Vgl. ebd., S. 70-72

Bezug zum ursprünglichen Suchterm bei Intransitivität nicht ausgegangen werden kann[97].

Der Erfolg einer Modifikation anhand paradigmatischer Beziehungen hängt wesentlich von der Qualität der Begriffsordnung ab. Hierarchische Relationen müssen nach einer schlüssigen Logik aufgebaut sein, wobei insbesondere auf Transitivität der Relation zu achten ist. Die Umsetzung dieses Anspruchs kann Schwierigkeiten verursachen, insbesondere wenn ein normiertes Vokabular tatsächlich zur inhaltlichen Erschließung eingesetzt wird. Es muss dann abgewägt werden zwischen der Bewahrung der logischen Struktur und pragmatischer Subordinierung der einzelnen Begriffe. Ein weiteres Qualitätsmerkmal ist die Vollständigkeit der Relationierungen. Bei einem beziehungslosen Deskriptor kann die Anfragemodifikation nicht greifen[98].

Grundsätzlich kann dieses Prinzip nur zwischen zwei Zuständen unterscheiden: entweder besteht eine semantische Ähnlichkeit zwischen zwei Termen, oder es besteht keine semantische Ähnlichkeit. Im ersten Fall kann ein als ähnlich identifizierter Term automatisch zur Anfrage hinzugefügt werden oder dem Nutzer als neuer Suchterm vorgeschlagen werden. Im zweiten Fall kommt der Term für eine Anfragemodifikation nicht in Frage.

Dieses dichotome Prinzip lässt zu Wünschen übrig, da eine semantische Ähnlichkeit nicht immer in der gleichen Intensität vorliegt. Es ist offensichtlich, dass etwa zwischen „Luftfahrzeug“ und „Flugzeug“ ein intensiverer, semantischer Zusammenhang besteht, als zwischen „Luftfahrzeug“ und „Verkehrsmittel“. In einer Fortentwicklung des beschriebenen Verfahrens würde man daher vom dichotomen zum gewichteten Prinzip übergehen. Darunter versteht man die Kopplung der verschiedenen Relationsarten an bestimmte Gewichtungsfaktoren[99].

Das gewichtete Prinzip lässt sich am besten anhand eines weitern Beispiels schildern. Die Software *RetrievalWare* der Fa. *Convera* berücksichtigt in der Version 8.0 für die verschiedenen Relationsarten eines Vokabulars prozentuale

---

[97] Vgl. ebd., S. 72
[98] Vgl. Gödert; Lepsky: Semantische Umfeldsuche im Information Retrieval in Online-Katalogen, S. 14
[99] Vgl. Stock: Information Retrieval, S. 285-287

Abstufungen, die sich an der jeweils angenommenen semantischen Ähnlichkeit orientieren. Die Standardkonfiguration der Software sieht für hierarchische Relationen zu Unterbegriffen 80 Prozent, für hierarchische Relationen zu Oberbegriffen 50 Prozent, für Assoziationsrelationen 40 Prozent und für Äquivalenzrelationen 100 Prozent vor[100].

Legt man diese Gewichtungen dem Beispiel aus Abbildung 5 zugrunde, würde sich zwischen dem Suchterm „Verkehrsmittel“ und den einzelnen Unterbegriffen eine semantische Ähnlichkeit von jeweils 80 Prozent ergeben. Angenommen es lägen transitive Relationen vor und eine Anfrageerweiterung käme auch über mehrere Hierarchieebenen in Frage, dann würde die semantische Ähnlichkeit zu den zwei Ebenen tiefer liegenden Unterbegriffen nur noch 40 Prozent betragen. Ebenfalls mit 40 Prozent würde der verwandte Begriff „Verkehr“ Berücksichtigung finden.

Die Anfragemodifikation kann bei diesem Modell durch die Definition eines Schwellenwertes gesteuert werden, indem nur solche Terme zur Anfrage hinzugefügt werden, deren semantischer Abstand den Schwellenwert überschreitet[101]. Für interaktive Verfahren ließe sich die Errechnung solcher einfacher Abstandsmaße auch für Zwecke der Visualisierung nutzen. Beispielsweise könnten potentielle Suchterme nicht nur in Listenform, sondern in Form eines semantischen Netzes dargestellt werden, wobei die semantische Ähnlichkeit durch die Größe der Knoten oder die Länge der Kanten repräsentiert werden könnte und dem Nutzer ein intuitives Verständnis für das Relationsgefüge des Vokabulars möglich wäre.

5.2.1.2 Natürlichsprachliche Thesauri am Beispiel *WordNet*

Das Projekt *WordNet* hatte seinen Ursprung 1985 an der Princeton University in den USA, wo es zunächst als ein gewöhnliches Online-Wörterbuch aufgebaut werden sollte. Im Laufe der Zeit wurde *WordNet*, basierend auf psycholinguistischen Grundsätzen, zu einer lexikalischen Datenbank der

---

[100] Vgl. Bayer, Oliver et al.: Evaluation of an ontology-based knowledge-management-system. A case study of Convera RetrievalWare 8.0. In: Information Services & Use 25(2005), S. 189

[101] Vgl. Stock: Information Retrieval, S. 286

englischen Sprache weiterentwickelt[102]. *WordNet* unterscheidet zwischen den Wortklassen Substantiv, Verb, Adjektiv und Adverb. Im Gegensatz zu einem dokumentationssprachlichen Thesaurus werden in *WordNet* keine Vorzugsbenennungen festgelegt. Als kleinste semantische Einheiten werden sogenannte „sets of synonyms" (Synsets) erfasst, in denen alle Benennungen für einen bestimmten Begriff zusammengeführt werden[103].

Auch in *WordNet* haben semantische Beziehungen eine zentrale Bedeutung für die Strukturierung des Vokabulars. Dabei bestehen Beziehungen in *WordNet* nicht nur zwischen einzelnen Synsets, sondern auch zwischen einzelnen Wörtern innerhalb eines oder mehrer Synsets. Die Synonymie, die Miller explizit als die wichtigste Beziehung in *WordNet* hervorhebt[104], besteht zwangsläufig zwischen allen Wörtern, die in einem Synset zusammengefasst sind.

Die Antonymie, also die gegensätzliche Bedeutung, ist die vorherrschende Relation zwischen adjektivischen Synsets[105]. Antonymie ist allerdings ebenso in anderen Wortklassen möglich. Es muss außerdem zwischen der kontradiktorischen und der konträren Antonymie unterschieden werden. Die kontradiktorische Antonymie liegt vor, wenn es sich um Gegensatzpaare handelt, zwischen denen keine weiteren Abstufungen möglich sind. Ein Beispiel aus *WordNet* für kontradiktorische Antonymie in der Klasse der Adjektive ist „right" und „wrong" im Sinne eines moralischen Urteils. Dagegen handelt es sich um konträre Antonymie, wenn zwischen dem Begriffspaar noch weitere Abstufungen möglich sind, etwa führt das Synset „right" im Sinne einer politisch-weltanschaulichen Orientierung neben „left" auch „center" als Antonyme an (vgl. Abbildung 6)[106].

---

[102] Vgl. Miller, George et al.: Introduction to WordNet: an on-line lexical database. In International Journal of Lexicography 3(1990)4, S. 236

[103] Vgl. ebd., S. 240

[104] Ebd., S. 241. Miller unterscheidet präziser zwischen lexikalischen Relationen, die zwischen einzelnen Wörtern bestehen und semantischen Relationen, die zwischen Synsets bestehen. Diese Unterscheidung wird hier nicht übernommen, sondern vereinfachend von semantischen Relationen ausgegangen.

[105] Vgl. ebd., S. 242

[106] Recherchiert in WordNet Search Version 3.0. Online: http://wordnet.princeton.edu/perl/webwn [Abrufdatum: 02.06.2008]

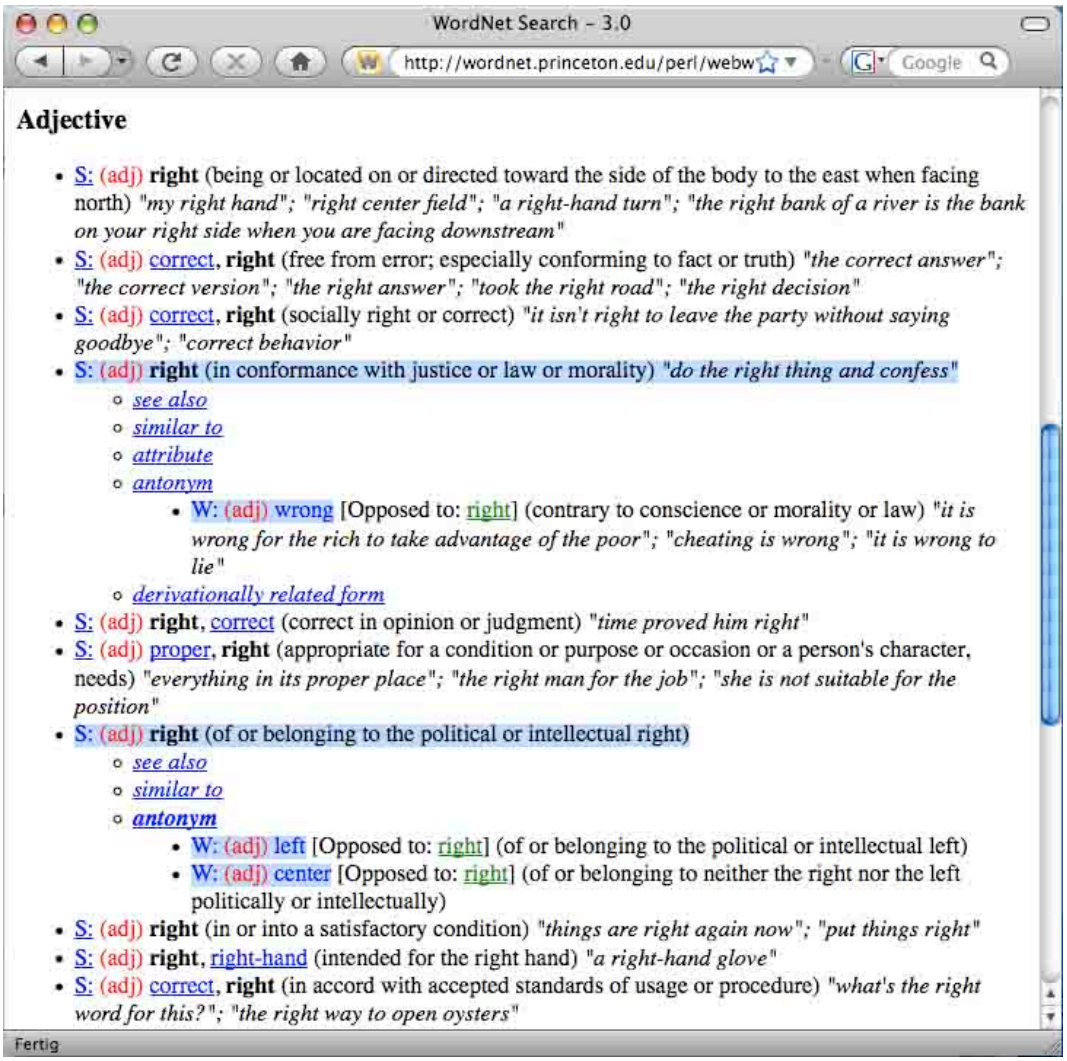

Abbildung 6: Antonymie in *WordNet* Search 3.0
Quelle: N.N. 5a

Stock weist auf den Nutzen kontradiktorischer Antonymie für die Modifikation der Anfrage hin und schlägt vor, die Anfrage um eine Negation des jeweiligen Gegenbegriffs zu erweitern. Bezogen auf das Beispiel aus Abbildung 6 ist es jedoch zweifelhaft, ob Dokumente mit dem Inhalt „not wrong“ tatsächlich dem „right“ entsprechen, auf das der Nutzer jeweils aus ist. Bei der konträren Antonymie verschärft sich diese Vagheit noch weiter, so dass Stock vorschlägt, zumindest Relationen der konträren Antonymie für die Anfragemodifikation außer Acht zu lassen[107].

Darüber hinaus stehen – wie bei dokumentationssprachlichen Thesauri – hierarchische Strukturen zwischen den Synsets in Form von Hyponym-

[107] Stock: Information Retrieval, S. 281

Hyperonym- und Meronym-Holonym-Relationen für die Anfragemodifikation zur Verfügung.

Da *WordNet* aber nicht mit einer Dokumentationssprache verglichen werden kann, kann es auch nicht ohne weiteres zur inhaltlichen Erschließung und Indexierung eingesetzt werden. Mitunter wird nämlich ein einzelner Begriff auf mehrere Synsets abgebildet, die dann nur noch eine minimale Bedeutungsdifferenz aufweisen (beispielsweise führt der Begriff „library“ zu fünf substantivischen Synsets, wovon sich vier Synsets auf „library“ im Sinne einer Büchersammlung beziehen, so dass zwischen ihnen eine semantische Differenzierung kaum noch möglich ist[108]). Ein derart ausdifferenziertes Vokabular für die Indexierung zu gebrauchen, kann beim Retrieval „den Nutzer zur Verzweiflung führen“[109], wenn er nach Eingabe des Suchterms „library“ aufgefordert wird, sich zwischen „einem Raum, in dem Bücher aufbewahrt werden“, „einem Gebäude, in dem Bücher aufbewahrt werden“, „einem Aufbewahrungsort für Bücher“ und „einer Büchersammlung“ zu entscheiden.

Eine Lösung besteht darin, überflüssige Synsets zu löschen, oder zusammenzufassen[110]. Wann jedoch ein Synset überflüssig ist, kann mit Sicherheit nur in Abhängigkeit vom jeweiligen Anwendungsfall entschieden werden. Außerdem muss die Konsistenz gewahrt bleiben, das heißt, eine einmal getroffene Entscheidung, wann ein Synset entfernt oder fusioniert werden muss, muss auf alle vergleichbaren Fälle ebenso angewendet werden. Obgleich es bereits Regelsammlungen für die Zusammenfassung semantisch nahezu identischer Synsets für das Retrieval gibt[111], wäre die Durchführung doch immer noch mit einem erheblichen intellektuellen Aufwand verbunden.

Vor diesem Hintergrund erscheint es sinnvoller, natürlichsprachliche Thesauri wie *WordNet* als reine Orientierungshilfen bei einem Freitextretrieval einzusetzen. Beim Durchsuchen von Sachtiteln, Abstracts oder gar Volltexten ist der Nutzer ohnehin auf eine Stichwortsuche angewiesen. Bleibt diese in der initialen Anfrage erfolglos, tritt *WordNet* buchstäblich als „Stichwortgeber“ auf,

---

[108] Recherchiert in WordNet Search Version 3.0
[109] Stock: Information Retrieval, S. 282-283
[110] Vgl. ebd.
[111] Vgl. ebd.

indem die eingegebenen Suchterme um die dazu relationierten Synsets erweitert werden.

Mandala et al. geben einen Überblick über die verschiedenen Retrievaltests der 1990er Jahre, in denen *WordNet* zur Anfragemodifikation nach diesem Prinzip eingesetzt wurde. Die Ergebnisse fassen sie allerdings als durchweg negativ zusammen. Die Retrievalleistung konnte in keinem der unterschiedlichen Testverfahren verbessert werden, häufig führte der Einsatz von *WordNet* sogar zu einer Verschlechterung[112]. Um den Ursachen für die schlechten Ergebnisse nachzugehen, haben Mandala et al. in neun Testkollektionen ein Retrieval mit Anfragemodifikation, unter Berücksichtigung der Synonymie und der Hyponym-Hyperonym-Relationen anhand von *WordNet* durchgeführt: ein Suchdurchgang ohne Erweiterung der Anfrage, Erweiterung um Synonyme, Erweiterung um Synonyme und Oberbegriffe, Erweiterung um Synonyme und Unterbegriffe, Erweiterung um Synonyme, Oberbegriffe und Unterbegriffe[113].

Im Ergebnis zeigte sich, dass alle Modifikationsverfahren zwar zu einer Verbesserung des Recall führten, die niedrige Precision die Ergebnisse jedoch unbrauchbar machte. Mandala et al. führten das schlechte Ergebnis auf die folgenden Ursachen zurück: zunächst verhindert die Trennung in Wortklassen die Berücksichtigung semantischer Relationen zwischen Termen verschiedener Wortklassen. Außerdem sind in *WordNet* keine Eigennamen enthalten, was dazu führt, dass auch keine Relationen zwischen Eigennamen oder zwischen Eigennamen und *WordNet*-Vokabular berücksichtigt werden können[114].

In der zitierten Studie argumentieren Mandala et al., man könne der Unvollkommenheit natürlichsprachlicher Begriffsordnungen wie *WordNet* letztlich nur durch Anreicherung mit ergänzendem Vokabular begegnen, das in der Lage ist, semantische Relationen auch zwischen verschiedenen Wortklassen, Eigennamen und anderen, kollektionsspezifischen Termen zu erkennen und gewissermaßen „die Lücken zu schließen“[115]. Solches Vokabular

---

[112] Mandala, Rila; Takenobu, Tokunaga; Hozumi, Tanaka: The use of WordNet in information retrieval. In: Harabagiu, Sanda (Hrsg.): Proceedings of ACL Workshop on the Usage of WordNet in Natural Language Processing Systems, 1998, S. 31-32
[113] Ebd., S. 32-33
[114] Ebd., S. 32
[115] Ebd., S. 36

wird in Abhängigkeit von der Dokumentenkollektion durch automatische Verfahren aufgebaut.

5.2.2 Kollektionsabhängige Begriffsordnungen

Kollektionsabhängige Begriffsordnungen bieten, alternativ oder ergänzend zu dem begrenzten Vorrat paradigmatischer Relationen die Möglichkeit, syntagmatisch relationierte Begriffsordnungen aus der Kollektion heraus zu gewinnen und für eine Anfragemodifikation zu nutzen.

Dazu werden Techniken herangezogen, die unter dem Begriff „Data-Mining“ oder spezifischer „Text-Mining“ zusammengefasst werden. Auf Basis der Termfrequenzen und der Gewichtungen werden unter Nutzung des Vektorraummodells Ähnlichkeiten berechnet[116]. Während im Vektorraum die einzelnen Dokumente mit der Anfrage für ein Retrieval verglichen werden, lassen sich ebenso Ähnlichkeiten von Dokument zu Dokument oder von Term zu Term bestimmen.

Einander ähnliche Dokumente können beim Anlegen der Kollektion in Cluster zusammengefasst werden. Die theoretische Grundlage dafür liefert die Clusterhypothese, die davon ausgeht, dass eng verwandte Dokumente zur gleichen Anfrage relevant sind[117]. Durch Clusterverfahren lassen sich Begriffe zu Gruppen (Clustern) zusammenfassen. Salton et al. führen dazu aus:

„Werden Begriffe zu Gruppen zusammengefasst, erhält man eine strukturierte Begriffsliste oder einen Thesaurus [...]. Thesauri werden normalerweise von Fachwissenschaftlern manuell erstellt, obwohl sich automatische Klassifikationsverfahren, die beispielsweise auf Ähnlichkeitsinformationen zwischen Begriffspaaren aufbauen, prinzipiell verwenden lassen. So lässt sich aus den Verteilungen der Begriffe einer Dokumentation ohne weiteres die Ähnlichkeit zwischen diesen Begriffen bestimmen.“[118]

Diese Ähnlichkeitsinformationen beruhen auf einem häufigen gemeinsamen Auftreten bestimmter Terme. Dazu wird zunächst eine Dokument-Term-Matrix

---

[116] Vgl. Nohr: Theorie des Information Retrieval II: Automatische Indexierung, S. 218-219
[117] Vgl. van Rijsbergen: Information Retrieval, S. 80
[118] Salton; McGill: Information Retrieval, S. 240-241

herangezogen, in der erfasst ist, welche Dokumente welche Terme mit welcher Gewichtung enthalten.

	t_1	t_2	t_3	t_4
d_1	3	0	1	0
d_2	0	2	2	2
d_3	2	2	0	4
d_4	3	0	5	1

	t_1	t_2	t_3	t_4
t_1	-	4	18	11
t_2	4	-	4	12
t_3	18	4	-	9
t_4	11	12	9	-

Abbildung 7: Dokument-Term-Matrix und Term-Term-Korrelationsmatrix
Quelle: Frei nach: Kowalski; Maybury: Information Storage and Retrieval Systems, S. 146-147

Für jeweils zwei Spalten der Dokument-Term-Matrix werden die Spalten in jeder Zeile multipliziert und anschließend aufaddiert. Das Ergebnis ist eine Term-Term-Matrix, die alle Terme zueinander in Beziehung setzt[119]. Anschließend setzt ein Clustering-Algorithmus ein, der anhand der Term-Term-Matrix die Terme den entsprechenden Clustern zuweist. Clusteringverfahren sind vielseitige Instrumente in der Informatik, die für zahlreiche Bedürfnisse und Gegebenheiten existieren, daher ist die Anzahl der einzelnen Algorithmen unüberschaubar. Eine sehr allgemeine Unterscheidung wird zwischen hierarchischem und partitionierendem Clustering getroffen. Zur automatischen Erzeugung von Thesauri wird häufig der Star-Algorithmus herangezogen, ein graphentheoretisches Verfahren, das zum partitionierenden Clustering gezählt wird[120]. Der Ablauf eines graphentheoretischen Clustering lässt sich folgendermaßen beschreiben:

Ausgehend von einer Term-Term-Matrix wie in Abbildung 7 muss zunächst ein Schwellenwert definiert werden, ab wann von einer Ähnlichkeit zweier Terme ausgegangen werden soll. Die Terme, die den Schwellenwert unterschreiten, werden durch eine erste Clusterbildung separiert. Die Termpaare, deren Ähnlichkeit den Schwellenwert überschreitet, werden durch eine Kante miteinander verbunden. Es entsteht ein Ähnlichkeitsgraph, aus dem die

[119] Vgl. Kowalski, Gerald J.; Maybury, Mark T.: Information Storage and Retrieval Systems. Theory and Implementation, 2000, S. 146-147
[120] Vgl. Kürsten, Jens: Systematisierung und Evaluierung von Clustering-Verfahren im Information Retrieval, 2006, S. 49

einzelnen Cluster sich als die am häufigsten miteinander verbundenen Komponenten herausbilden[121]. Zuletzt wird für jedes Cluster ein Zentroid bestimmt. Zentroide sind in der Regel fiktive Terme, die sich aus den Durchschnittswerten des Clusters errechnen. Sowohl der Retrievalprozess, wie auch das Einordnen neuer Terme in die Clusterstruktur kann durch den Abgleich mit den Clusterzentroiden vereinfacht werden[122].

Auf der Grundlage des Termclustering lassen sich schließlich Thesauri erzeugen, die mit dokumentationssprachlichen Thesauri nicht zu vergleichen sind. Automatisch erzeugte Thesauri verfügen nur über eine Relation, nämlich die Koinzidenz der Terme und damit der Themen, die sie repräsentieren[123]. Baeza-Yates und Ribeiro-Neto beschreiben zwei Modelle, die auf diesem Prinzip beruhen: den Ähnlichkeitsthesaurus und den statistischen Thesaurus[124], die im Folgenden näher betrachtet werden sollen.

5.2.2.1 Ähnlichkeitsthesauri

Das Konzept des Ähnlichkeitsthesaurus beruht auf der Idee, Terme aufgrund ihrer Ähnlichkeit zur Anfrage für die Anfragemodifikation heranzuziehen. Dabei werden die statistischen Werte nicht unmittelbar einer Term-Term-Matrix entnommen, der Ähnlichkeitsthesaurus basiert auf Term-Term-Beziehungen, die sich aus dem Vergleich der Terme im Vektorraum ergeben. Die Verhältnisse im Vektorraum werden dazu umgekehrt, die Dokumente der Kollektion bestimmen die Dimensionalität des Raumes, die Terme werden darin als Vektoren repräsentiert[125].

Die einzelnen Gewichtungen der Terme, die die Position des Termvektors bestimmen, errechnen sich aus einer Formel, die der TF*IDF-Formel entspricht, mit dem Unterschied, dass nicht die inverse Dokumenthäufigkeit, sondern die inverse Termhäufigkeit berücksichtigt wird. Dieser „ITF“-Faktor ergibt sich aus

---

[121] Vgl. Nohr: Grundlagen der automatischen Indexierung, S. 57
[122] Vgl. Salton; McGill: Information Retrieval, S. 230-231
[123] Vgl. Stock, Wolfgang G.: Textwortmethode. In: Password 15(2000)7/8, S. 32
[124] Baeza-Yates; Ribeiro-Neto: Modern Information Retrieval, S. 131-137
[125] Vgl. ebd., S. 131

der Anzahl aller Terme der Kollektion im Verhältnis zur Anzahl aller verschiedenen Terme innerhalb eines Dokuments[126].

Ähnlich zur Vorgehensweise in Abbildung 7 wird der Korrelationsfaktor zwischen zwei Termen durch komponentenweises Multiplizieren und Aufaddieren ermittelt, wobei die zugrundeliegenden Gewichtungen sich aus den verkehrten Verhältnissen ableiten (Dokumente werden als Indexterme interpretiert und umgekehrt). Durch den rechenintensiven Vorgang des paarweisen Vergleichens aller Terme entsteht schließlich die Begriffsordnung.

Beim Retrieval wird eine Anfrage als Vektor repräsentiert, so dass sich anhand des Korrelationsfaktors die Ähnlichkeit der Anfrageterme und der gesamten Anfrage zum Vokabular des Ähnlichkeitsthesaurus ermitteln lässt. Die Terme, die die größte Ähnlichkeit zur Anfrage aufweisen, können zur Modifikation genutzt werden[127] – entweder dialogorientiert durch bloßes Vorschlagen, oder, bei entsprechender Parametrierung des gewünschten Ähnlichkeitsgrades und der maximalen Anzahl hinzuzufügender Terme, auch automatisch.

5.2.2.2 Statistische Thesauri

Der statistische Thesaurus geht hervor aus der Zusammenfassung von häufig gemeinsam auftretenden Termen in Klassen. Diese Klassenbildung soll auf entscheidungsstarke Terme beschränkt bleiben, weshalb Baeza-Yates und Ribeiro-Neto ein Verfahren vorschlagen, dass von der Gewinnung niedrigfrequenter Terme ausgeht[128]. Dies steht einerseits im Widerspruch zu den Erkenntnissen der Textstatistik, wonach niedrigfrequente Terme zur Repräsentation eines Dokumenteninhalts zu wenig signifikant sind, andererseits geht es beim Aufbau des statistischen Thesaurus nicht um einzelne Dokumenteninhalte, sondern um die Gewinnung eines kollektionsbezogen aussagefähigen Vokabulars.

Da ein Termclustering der niedrigfrequenten Terme schon allein aufgrund ihres geringen Vorkommens in der Kollektion Schwierigkeiten bereitet, wird

---

[126] Vgl. ebd., S. 132
[127] Vgl. ebd., S. 133
[128] Vgl. ebd., S. 134-137

stattdessen ein Dokumentenclustering durchgeführt, um in Anschluss die gewünschten Terme aus den einzelnen Dokumentenclustern zu entnehmen[129].

In dieser Situation sind möglichst kleine und eng verbundene Cluster erforderlich, damit ein semantischer Bezug der im Cluster enthaltenen Terme nicht zu unwahrscheinlich wird. Der „complete link"-Algorithmus entspricht diesen Anforderungen. Es handelt sich um ein hierarchisches, agglomeratives Verfahren, das heißt, dass das Clustering zu einer hierarchischen Baumstruktur führt, wobei die Cluster zu Beginn aus einzelnen Dokumenten bestehen und schrittweise die ähnlichsten Cluster zusammengefasst werden[130].

Dazu wird die Ähnlichkeit zwischen zwei Clustern ermittelt anhand des Dokumentenpaares, das über die geringste Ähnlichkeit verfügt. Anschließend werden solche Cluster als Paare zusammengefasst, die jeweils über die höchste Ähnlichkeit verfügen, woraus sich kleine, eng verbundene Cluster ergeben. Dieser Vorgang führt schließlich zu einem Strukturbaum, in dem zwei Cluster fusionieren und ein übergeordnetes Cluster bilden, welches wiederum mit einem anderen Cluster zu einem übergeordneten Cluster fusioniert. Je höher in der Hierarchie vorangeschritten wird, desto mehr nimmt die Ähnlichkeit der Cluster ab, da sie immer größer werden und immer mehr Dokumente umfassen[131].

Aus der gegebenen Clusterstruktur werden anschließend die Terme entnommen, die zur Bildung von Thesaurusklassen verwendet werden sollen. Dazu müssen drei Parameter bestimmt werden. Zunächst wird ein Schwellenwert festgelegt, der regelt, welche Cluster zur Bildung der Thesaurusklassen überhaupt in Frage kommen. Der Schwellenwert bezieht sich auf die, während des Clusterings ermittelten Ähnlichkeitsmaße: je höher er angesetzt wird, desto mehr beschränkt sich die Auswahl auf die kleineren Cluster im unteren Bereich der Hierarchie.

Wird der Schwellenwert niedrig gewählt (und damit große Cluster also nicht von vornherein ausgeschlossen), besteht die Möglichkeit durch einen zweiten

---

[129] Vgl. ebd., S. 134-135
[130] Vgl. Kürsten: Systematisierung und Evaluierung von Clustering-Verfahren im Information Retrieval, S. 33
[131] Vgl. Baeza-Yates; Ribeiro-Neto: Modern Information Retrieval, S. 135

Parameter, die die maximale Anzahl der Dokumente in einem Cluster festlegt, zwischen einem größeren und einem kleineren Cluster zu wählen. Die Termextraktion wird durch einen Mindestwert der inversen Dokumenthäufigkeit (MIDF) reguliert, um dem statistischen Thesaurus nur entscheidungsstarke Terme zuzuführen[132].

Zuletzt müssen die Thesaurusklassen gewichtet werden, um bei einem Suchprozess einen Abgleich mit einer Anfrage zu ermöglichen. Dazu wird mittels der einzelnen Termgewichtungen und der Anzahl der Terme pro Thesaurusklasse ein Durchschnittsgewicht errechnet[133].

Der Nutzen eines statistischen Thesaurus für die Anfragemodifikation hängt besonders von der Parametrierung ab, mit der die Granularität der Thesaurusklassen beeinflusst wird. Ein statistischer Thesaurus ist ein dynamisches Instrument, das sich mit jedem neu indexierten Dokument ändert, ebenso wie mit jeder Nutzereinstellung. Konventionelle Thesauri unterscheiden zwischen Deskriptoren als Gebrauchsvokabular und Nicht-Deskriptoren als Zugangsvokabular. Auf die statistischen Thesauri übertragen, kann von temporären Deskriptoren und temporären Nicht-Deskriptoren gesprochen werden[134], wobei letztere den Status der Nicht-Deskiptoren innehaben, weil sie etwa einen temporär definierten MIDF-Wert unterschreiten.

5.2.3 Latent Semantic Indexing

Durch Latent Semantic Indexing ist es möglich, das semantische Umfeld einer Anfrage in die Suche miteinzubeziehen, ohne dass es zu einer Anfragemodifikation im Sinne des Entfernens oder Hinzufügens von Suchtermen kommt. Das Latent Semantic Indexing ist originär kein Verfahren zur Anfragemodifikation, das den iterativen Ablauf eines Suchprozesses unterstützen würde. Es handelt sich vielmehr um eine Weiterentwicklung des klassischen Vektorraummodells, mit der eine semantische Umfeldsuche bereits mit der initialen Anfrage durchgeführt werden kann. Ein besonderes Merkmal

---

[132] Vgl. ebd., S. 136
[133] Vgl. ebd.
[134] Vgl. Stock: Textwortmethode, S. 32

des Verfahrens ist die Fähigkeit, zu einer Anfrage auch solche Dokumente zu finden, in denen kein einziger Anfrageterm enthalten ist[135].

Der theoretische Hintergrund des Latent Semantic Indexing gründet sich wiederum auf die Textstatistik. Es wird das häufige gemeinsame Auftreten von Termen ermittelt und daraus auf eine semantische Ähnlichkeit geschlossen. Das rein statistische Verfahren verzichtet dabei vollständig auf linguistische Hilfsmittel wie Wörterbücher, morphologsiche Analysen oder Instrumente der Wissensrepräsentation wie etwa Begriffsordnungen[136].

Der Ansatz des Latent Semantic Indexing beruht stattdessen auf der Reduzierung der hochdimensionalen Vektorräume auf wesentlich weniger Dimensionen. Dadurch kommt es zu einer Verzerrung des Vektorraums, die für das Retrieval durchaus positive Auswirkungen haben kann. Semantisch identische oder ähnliche Terme, die so über die Kollektion verteilt sind, dass sie zwar nie gemeinsam in einem Dokument auftreten, aber häufig in den gleichen Kontexten, liegen im beschränkt-dimensionalen Vektorraum dicht beieinander und sind zu einem entsprechenden Anfragevektor relevant. Dumais gibt dazu ein Beispiel:

„[...] the words ‚physician' and ‚doctor' never co-occur in a single article, but they are quite similar in the reduced LSA space. This is because they occur in many of the same contexts [...], and when dimension constraints are imposed, the vectors for doctor and physician are near each other in the reduced LSA space."[137]

Durch die Berücksichtigung solcher latent semantischen Zusammenhänge lassen sich viele Phänomene der natürlichen Sprache im Vektorraummodell besser bewältigen. Synonyme Terme können als semantische Einheiten erkannt und durch eine einzige Dimension abgebildet werden, homonyme Terme können durch ihren terminologischen Kontext disambiguiert und der jeweils richtigen Dimension zugeordnet werden. Das Verfahren zielt also darauf ab, in der Kollektion vorkommende Terme zu einzelnen Themen

---

[135] Dumais, Susan: Latent semantic analysis. In: Cronin, Blaise (Hrsg.): Annual Review of Information Science and Technology 38(2004), S. 194
[136] Vgl. ebd., S. 191
[137] Ebd., S. 198-199

zusammenzufassen und anstatt jeden Term, nur noch jedes Thema durch eine Dimension abzubilden.

Der Schlüssel zu einem derart modifizierten Vektorraum ist das mathematische Verfahren der Singulärwertzerlegung. Ausgangspunkt ist eine durch die Textstatistik erstellte, gewöhnliche Term-Dokument-Matrix. Diese Matrix wird in drei spezielle Matrizen gespalten:

Abbildung 8: Singulärwertzerlegung
Quelle: Schenkel; Weikum: Vorlesung „Informationssysteme“, S. 30. Die Abbildung wurde verändert.

Die Term-Themen- und die Themen-Dokument-Matrix beschreiben jeweils den Term- bzw. den Dokumentraum in der, auf eine bestimmte Anzahl von Themen reduzierten Dimensionalität. Die Diagonalmatrix enthält die sogenannten „Singulärwerte“, der Term-Dokument-Matrix, anhand derer die Transformation in einen Raum mit anderer Dimensionalität ermöglich wird[138].

Im beschränkt-dimensionalen Vektorraum lassen sich dann Ähnlichkeiten von Term zu Term, von Dokument zu Dokument, sowie von Term zu Dokument bestimmen. Anfragevektoren werden als Zentroide der Termvektoren verortet, aus denen sie bestehen und können für ein Retrieval mit Dokumentvektoren verglichen werden[139].

Entscheidend für die Leistungsfähigkeit des Verfahrens ist die Anzahl der Dimensionen, auf die die Singulärwertzerlegung eine Dokumentenkollektion reduziert. In einem Retrievaltest, in dem 30 Anfragen an eine kleine Testkollektion von 1.033 Dokumenten gestellt wurden, konnte Latent Semantic Indexing einen um 30 Prozent höheren Recall erzielen, als ein Retrievalsystem mit konventionellem Wort-Matching. Die Kollektion enthielt insgesamt 5.831 Terme und wurde durch die Singulärwertzerlegung auf nur noch 90

[138] Vgl. Schenkel, Ralf; Weikum, Gerhard: Vorlesung „Informationssysteme“, Sommersemester 2004, S. 30
[139] Vgl. Dumais: Latent Semantic Analysis, S. 192-193

Dimensionen reduziert. Bei einer zu starken Reduzierung fällt die Leistung allerdings hinter das Niveau des konventionellen Retrieval zurück, bei einer zu geringen Reduzierung gleicht sie sich an[140].

5.3 Relevance Feedback

Das Relevance Feedback bietet im Vergleich zu den bisher vorgestellten Verfahren eine neue Qualität des Dialogs zwischen Nutzer und Retrievalsystem. Der Gegenstand des Dialogs ist nicht mehr die Identifizierung und Berücksichtigung semantisch ähnlicher Suchterme, sondern die Relevanz einzelner Dokumente.

Einleitend lässt sich das Relevance Feedback durch drei Merkmale beschreiben[141]:

- Es erspart dem Nutzer den aufwendigen Vorgang der intellektuellen Modifikation und ermöglicht die Formulierung geeigneter Suchargumente ohne detaillierte Kenntnisse über Retrievalsystem und Dokumentenkollektion.
- Es unterteilt den Suchprozess in eine Abfolge kurzer Suchschritte, die schließlich zum gewünschten Themenkreis führen.
- Es bietet einen kontrollierten Ablauf der Anfragemodifikation, der darauf ausgerichtet ist, je nach Informationsbedarf Terme in ihrer Gewichtung herauf- oder herabzusetzen.

Die Ausgangssituation für ein Relevance Feedback ist eine erste Ergebnismenge, die auf eine initiale Anfrage hin ausgegeben wurde. Der Nutzer hat nun die Möglichkeit, die Ergebnismenge auszuwerten und zu den einzelnen Dokumenten Relevanzurteile abzugeben. Relevante Dokumente werden markiert oder auf irgendeine andere Art gekennzeichnet, je nach den gegebenen Interaktionsmöglichkeiten. Die unmarkierten Dokumente gelten damit als irrelevant. Die bearbeitete Ergebnismenge wird dem Retrievalsystem übergeben, das auf der Grundlage der Relevanzurteile des Nutzers die Anfrage modifiziert und eine zweite Ergebnismenge ausgibt, in der im Idealfall mehr

---

[140] Vgl. ebd., S. 196-197

[141] Vgl. Salton, Gerard; Buckley, Chris: Improving retrieval performance by relevance feedback. In: Journal of the American Society for Information Science 41(1990)4, S. 288

relevante Dokumente enthalten sind. Dabei besitzt die modifizierte Anfrage eine hohe Ähnlichkeit mit den als relevant markierten Dokumenten und eine geringe Ähnlichkeit mit den als irrelevant markierten Dokumenten. Dieser Vorgang lässt sich theoretisch beliebig oft wiederholen[142].

Hinter diesem, aus der Nutzerperspektive beschriebenen Ablauf, steht ein Algorithmus, der geeignete Terme aus den Dokumente der ersten Ergebnismenge entnimmt und sie, entsprechend ihrer Relevanzbeurteilung, für die Anfragemodifikation heranzieht[143]. Terme aus Dokumenten, die als relevant markiert sind, werden zur Anfrage hinzugefügt oder, falls sie in der initialen Anfrage bereits enthalten waren, in ihrer Gewichtung heraufgesetzt. Terme aus Dokumenten, die als irrelevant markiert sind, werden aus der Anfrage entfernt oder in ihrer Gewichtung herabgesetzt. Basierend auf dieser grundsätzlichen Vorgehensweise, gibt es eine Vielzahl verschiedener algorithmischer Varianten, die grob unterteilt werden können in:

- positives Feedback,
- negatives Feedback und
- gemischtes Feedback.

Wobei das positive Feedback nur relevante Dokumente und das negative Feedback nur irrelevante Dokumente zur Anfragemodifikation berücksichtigt. Dagegen berücksichtigt das gemischte Feedback relevante und irrelevante Dokumente[144].

Das Relevance Feedback kann sowohl im Vektorraummodell, wie auch im probabilistischen Modell realisiert werden. Aufgrund der geringen praktischen Bedeutung dieses Verfahrens, bleibt die nachfolgende Darstellung auf eine Beschreibung des grundsätzlichen Ablaufs des Feedback-Prozesses in beiden Modellen beschränkt und ignoriert die umfangreichen Erweiterungen und Fortführungen des Relevance Feedbacks, die von den 1970er Jahren an bis in

---

[142] Vgl. Salton; McGill: Information Retrieval, S. 252-253
[143] Vgl. Salton; Buckley: Improving Retrieval Performance by Relevance Feedback, S. 288
[144] Vgl. Salton; McGill: Information Retrieval, S. 255-256

die 1990er Jahre diskutiert wurden, da diese nie über einen rein experimentellen Charakter hinausgekommen sind[145].

5.3.1 Vektorbasierter Ansatz nach Rocchio

Das Relevance Feedback im Vektorraummodell ist durch die Möglichkeit der grafischen Darstellung intuitiv verständlich. Der Feedback-Prozess führt zu einer Verschiebung des Anfragevektors von einem Bereich zu einem anderen Bereich innerhalb des Vektorraumes.

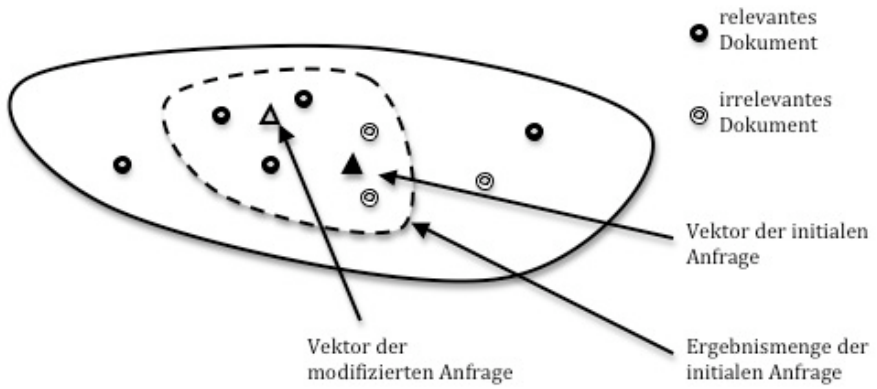

Abbildung 9: Verschiebung des Anfragevektors im Relevance Feedback
Quelle: Salton; McGill: Information Retrieval, S. 255. Die Abbildung wurde verändert.

Abbildung 9 zeigt ein erfolgreich verlaufenes Relevance Feedback. Der Anfragevektor wurde aus dem Umfeld der irrelevanten Dokumente verschoben, in die Nähe der relevanten Dokumente. Im gegebenen Beispiel wird das positive Ergebnis des Feedbacks unterstützt durch die günstige Verteilung der Dokumente im Vektorraum. Relevante und irrelevante Dokumente bilden jeweils eine relativ dichte Menge. Die Differenz zwischen den durchschnittlich relevanten und den durchschnittlich irrelevanten Dokumenten entspricht der Distanz der dazugehörigen Vektoren im Vektorraum, daher lässt sich verallgemeinernd sagen, dass das Relevance Feedback dann besonders erfolgreich ist, wenn sowohl die relevanten, wie auch die irrelevanten Dokumente eine dichte Menge bilden, bei gleichzeitig größtmöglichem Abstand zwischen beiden Mengen[146].

---

[145] Einen umfassenden Überblick über die Entwicklung des Relevance Feedback bieten Ruthven; Lalmas: A survey on the use of relevance feedback for information access systems

[146] Vgl. Salton; McGill: Information Retrieval, S. 152-153

Die modifizierte Anfrage wird durch einen Algorithmus generiert. Dazu wird ein Vektor berechnet, der aus den Unterschieden zwischen den relevanten und irrelevanten Dokumenten abgeleitet wird. Weil dabei nur mit Teilmengen der Dokumente gearbeitet werden kann – die Mengen aller relevanten und irrelevanten Dokumente der Kollektion sind ja unbekannt –, kann eine optimale Anfrage erst dann erreicht werden, wenn die zugrundeliegenden Teilmengen tatsächlich repräsentativ für die Gesamtmengen der Kollektion sind[147]. Aus dieser schrittweisen Annäherung an eine optimale Anfrage ergibt sich der iterative Charakter des Relevance Feedback.

Der Rocchio-Algorithmus gilt als der älteste Ansatz für das Vektor-basierte Relevance Feedback. Rocchio definiert die optimale Anfrage als einen Vektor, der sich aus der Differenz der bekannten relevanten und irrelevanten Dokumente ergibt. Auch der Vektor der initialen Anfrage muss für die Modifikation berücksichtigt werden, da er bereits wichtige Suchterme enthält. So ergibt sich der folgende Zusammenhang[148]:

$$Q_1 = Q_0 + \frac{1}{a}\sum R_i - \frac{1}{b}\sum N_i$$

Dabei entspricht Q_0 dem Vektor der initialen Anfrage, Q_1 dem Vektor der modifizierten Anfrage, a der Anzahl der relevanten Dokumente, b der Anzahl der irrelevanten Dokumente, R_i dem Vektor des $i$ten relevanten Dokuments und N_i dem Vektor des $i$ten irrelevanten Dokuments.

Aus der obenstehenden Formel ergibt sich schließlich die modifizierte Anfrage als eine Summe aus der initialen Anfrage und der Differenz der im Durchschnitt relevanten und irrelevanten Dokumente.

In einer Verfeinerung dieses Verfahrens werden die initiale Anfrage, sowie der Durchschnitt der relevanten und irrelevanten Dokumente mit jeweils individuell einstellbaren Gewichtungsfaktoren ausgestattet. Beispielsweise lässt sich die Formel nach Rocchio für ein negatives oder positives Feedback einstellen, durch die Gewichtung der relevanten, bzw. irrelevanten Dokumente mit dem

---

[147] Vgl. ebd., S. 254

[148] zitiert nach Ruthven; Lalmas: A survey on the use of relevance feedback for information access systems, S. 101

Wert 0, während eine Gewichtung beider Elemente mit einem Wert > 0 einem gemischten Feedback entspricht[149].

5.3.2 Probabilistischer Ansatz nach Robertson und Sparck-Jones

Das Relevance Feedback muss im klassischen probabilistischen Retrieval als fester Bestandteil des Modells verstanden werden. Die Anfragemodifikation ist hierbei keine optionale Erweiterung, sondern ein von vornherein vorgesehener Teilprozess. Wie im Vektorraummodell werden Anfragen und Dokumente durch Vektoren repräsentiert, im probabilistischen Modell tritt jedoch an die Stelle der Ähnlichkeitsmessung zwischen den Vektoren die Abschätzung der Wahrscheinlichkeit, dass ein Dokument für den Nutzer Relevanz besitzt, unter der Bedingung einer gegebenen Anfrage[150].

Anknüpfend an die im Abschnitt 3.2.2 beschriebenen Möglichkeiten der Termgewichtung zur Einschätzung der Relevanzwahrscheinlichkeit im probabilistischen Retrieval, soll an dieser Stelle eine Termgewichtungsfunktion für ein Relevance Feedback vorgestellt werden. Robertson und Sparck-Jones haben vier Gewichtungsfunktionen entwickelt, durch die eine Anfrage modifiziert werden kann[151]. Die Gewichtungsfunktionen erfüllen den gleichen Zweck wie der Rocchio-Algorithmus im vektorbasierten Relevance Feedback: auf der Grundlage einer Menge bekannter und durch den Nutzer als relevant oder irrelevant eingestufter Dokumente wird die Gewichtung der Anfrageterme angepasst.

Die vier Gewichtungsfunktionen wurden abgeleitet aus der in Tabelle 2 dargestellten Dokumentverteilung. Sie setzen die Mengen r, R, n und N auf verschiedene Weise zueinander ins Verhältnis, unter Berücksichtigung der beiden Unabhängigkeitsannahmen und Ranking-Prinzipien (vgl. 3.2.2). Die Darstellung soll hier beschränkt bleiben auf die Funktion, die in einem Retrievaltest im Durchschnitt die besten Ergebnisse erbrachte[152].

---

[149] Vgl. ebd.
[150] Vgl. ebd., S. 102
[151] Robertson; Sparck-Jones: Relevance weighting of search terms, S. 131
[152] Vgl. ebd., S. 135-138

Ausgehend von einem Term t setzt die Funktion die Anzahl der relevanten Dokumente, in denen Term t vorkommt ins Verhältnis zur Anzahl der relevanten Dokumente, in denen Term t nicht vorkommt. Ebenso wird verfahren mit der Anzahl der irrelevanten Dokumente, in denen Term t vorkommt und der Anzahl der irrelevanten Dokumente, in denen Term t nicht vorkommt. Das Gewicht für Term t ergibt sich dann aus dem Quotienten beider Werte. Dieser Zusammenhang wird dargestellt als[153]:

$$w_t = \log \frac{\frac{r}{(R-r)}}{\frac{(n-r)}{(N-n-R+r)}}$$

Der Wert w_t entspricht dem Gewicht für Term t, durch den Logarithmus kann die Spannbreite der einzelnen Termgewichte kleiner und aussagefähiger gemacht werden. Die Funktion berechnet Termgewichte anhand der Verteilung der Terme in den relevanten und irrelevanten Dokumenten. Ein Suchterm, der überwiegend in relevanten Dokumenten enthalten ist, erhält eine entsprechend hohe Gewichtung. Dagegen werden Suchterme, die zum größten Teil in irrelevanten Dokumenten enthalten sind, in ihrer Gewichtung herabgesetzt.

Der Vorteil des probabilistischen Relevance Feedback gegenüber den Verfahren im Vektorretrieval, ist die unmittelbare Ableitung der Neugewichtung der Suchterme anhand der im Feedback-Prozess übermittelten Informationen. Andererseits kann es als Nachteil empfunden werden, dass die Anfragemodifikation nicht direkt auf den Suchergebnissen, also den zuvor gefundenen Dokumenten beruht, wie dies im vektorbasierten Relevance Feedback der Fall ist, sondern indirekt über die Termverteilung in den relevanten und irrelevanten Dokumenten zustande kommt. Salton und Buckley vermuten darin die Ursache dafür, dass das probabilistische Relevance Feedback in der Effektivität insgesamt hinter dem Feedback im Vektorretrieval zurückbleibt[154].

Die von Robertson und Sparck-Jones eingeführte Gewichtungsfunktion ist außerdem beschränkt auf eine Neugewichtung der Terme der initialen Anfrage.

---

[153] Vgl. ebd., S. 131

[154] Salton; Buckley: Improving retrieval performance by relevance feedback, S. 291

Eine Anfrageerweiterung durch die Hinzuziehung neuer Suchterme findet – anders als im vektorbasierten Relevance Feedback – nicht statt. Harman gibt einen Überblick über verschiedene Verfahren zur Suchtermgewinnung für das probabilistische Relevance Feedback und diskutiert die Ergebnisse der dazu durchgeführten Retrievaltests[155].

5.4 Informetrische Rangordnungen und Zeitreihen

Nachdem in den vorangegangenen Abschnitten automatische und interaktive Modifikationsverfahren anhand von Termen und Dokumenten geschildert wurde, folgt nun der Sprung auf die abstraktere Ebene der bloßen Informationen.

Gegenstand der Informetrie ist die quantitative Messung von Informationen. Die Art der dazu berücksichtigten Dokumente geht weit über das Spektrum von Zeitschriften und Monografien hinaus und erstreckt sich auch auf den Bestand von Archiven oder das Internet. Während die Bibliometrie, die Webometrie, die Patentometrie oder die Szientometrie bestimmte Teilmengen dieses Spektrums erfassen, hat die Informetrie einen allumfassenden Anspruch und kann somit als Oberbegriff zu den übrigen Verfahren verstanden werden[156]. Die ältesten informetrischen oder szientometrischen Arbeiten entstanden in den 1920er Jahren, besondere Bedeutung hat der 1963 von Garfield und Sher aufgebaute *Science Citation Index* erlangt, ein Zitationsindex wissenschaftlicher Zeitschriften, für die durch Zitationsanalyse der „Impact-Factor“ als Kennzahl des wissenschaftlichen Einflusses der jeweiligen Zeitschrift bestimmt wird[157].

Zwischen informetrischen Verfahren, die einem analytischen Teilbereich der Metawissenschaft zugeordnet werden können, und Verfahren der Anfragemodifikation als Teilbereich im Information Retrieval, existiert offensichtlich kein direkter Zusammenhang. Der Berührungspunkt zwischen beiden Gebieten liegt bei der Informationsgewinnung. Informationen, die in

---

[155] Harman, Donna: Relevance feedback revisited. In: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development Information Retrieval, S. 1-10

[156] Vgl. Umstätter, Walther: Szientometrische Verfahren. In: Kuhlen, Rainer; Seeger, Thomas; Strauch, Dietmar (Hrsg.): Grundlagen der praktischen Information und Dokumentation. Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis, 2004, S. 237-243

[157] Vgl. ebd.

Datenbanken vorgehalten werden, müssen unter Verwendung geeigneter Retrievaltechniken gewonnen werden. Dies gilt unabhängig davon, ob die Informationen der Informetrie zugeführt werden, oder den Informationsbedarf eines Nutzers decken sollen.

Die Ermittlung von Autoren, die zu bestimmten Themen häufig publizieren, von wissenschaftlichen Werken, die häufig zitiert werden oder von wechselseitiger Beeinflussung verschiedener Forschungsthemen sind typisch informetrische Informationsbedarfe. Die dazu relevanten Informationen sind jedoch nicht in einzelnen Dokumenten enthalten, sondern müssen einer definierten Dokumentenmenge entnommen werden, die Dokumentationseinheiten werden in ihrer Summe „als Ganzes qualifiziert“[158].

Stock nennt vier Retrievaltechniken, die als informetrische Analysen die „Rohdaten“ für die Informetrie liefern[159]:

- Rangordnungen,
- Zeitreihen,
- semantische Netze und
- Informationsflussanalysen.

Die Informetrie benötigt sehr spezifische Daten, von deren Qualität letztlich die Aussagekraft der darauf aufbauenden empirischen Befunde abhängt. Es ist daher unwahrscheinlich, dass das Retrieval der Daten ohne Anfragemodifikation auskommt. Darauf sind auch die vier oben genannten Verfahren ausgerichtet, die auf einer gegebenen Dokumentenmenge (etwa einer Ergebnismenge zu einer initialen Anfrage) operieren und diese anhand statistischer Merkmale neu sortieren, aufbereiten oder reduzieren. Es lässt sich hier eine Verbindung zur den Verfahren des Relevance Feedbacks herstellen, die ebenfalls auf der Basis bereits vorliegender Suchergebnisse operieren. Das Relevance Feedback arbeitet jedoch global, das heißt, auch die modifizierte Anfrage wird an die gesamte Dokumentenkollektion gerichtet, nur so können schließlich neue Dokumente gefunden werden. Informetrische Analysen

---

[158] Vgl. Stock: Information Retrieval, S. 168-169
[159] Ebd., S. 173

arbeiten lokal, sie beziehen sich ausschließlich auf die bereits vorliegende Ergebnismenge, mit dem Ziel, implizite Informationen zu gewinnen.

Von den vier genannten Verfahren sollen hier lediglich Rangordnungen und Zeitreihen vorgestellt werden; die Funktion semantischer Netze gleicht den als „kollektionsabhängigen Begriffsordnungen“ bereits behandelten, statistischen Verfahren (vgl. 5.2.2), das Prinzip der Informationsflussanalysen wird im nachfolgenden Abschnitt im Zusammenhang mit assoziativem Retrieval aufgegriffen. Die Beschreibung der Rangordnungen und Zeitreihen stützt sich auf die umfangreicheren Ausführungen bei Stock[160].

Rangordnungen

Die elementarste Form einer informetrischen Analyse ist die Sortierung einer Ergebnismenge nach einer bestimmten Rangordnung. Sortierschlüssel ist dabei ein durch den Nutzer festzulegendes Feld der Datenbank, außerdem können Parameter wie auf- oder absteigende Sortierung und die Anzahl der zu sortierenden Datensätze festgelegt werden. Dieses einfache Prinzip ist dann von Nutzen, wenn geeignete und aussagefähige Feldinhalte für die Sortierung zur Verfügung stehen und wenn eine umfangreiche Ergebnismenge vorliegt, die eine Sortierung sinnvoll macht.

Zeitreihen

Mit der zuvor erzeugten Rangordnung liegt ein Sachverhalt vor, dessen Entwicklung im Laufe der Zeit durch die Erstellung einer Zeitreihe dargestellt werden kann. Die Dokumentenmenge muss dazu in gleichmäßige, zeitliche Intervalle unterteilt werden, bei bibliografischen Datenbanken kommen dazu in der Regel nur die Erscheinungsjahre der nachgewiesenen Dokumente in Frage. Beispielsweise ließe sich eine zuvor erstellte Rangordnung der, in einer bibliografischen Datenbank am häufigsten nachgewiesenen Verfasser zu einem bestimmten Sachgebiet, zusätzlich nach den Erscheinungsjahren ihrer einzelnen Werke unterteilen, um einen Überblick über das Publikationsgeschehen innerhalb eines definierten Zeitraums zu bekommen. Zeitreihen lassen sich, je nach Funktionalität des Retrievalsystems, in

[160] Stock: Information Retrieval, S. 168-184

Tabellenform oder in grafisch aufbereiteter Form erstellen. Ein Beispiel für eine grafisch aufbereitete Zeitreihe, die allerdings nicht beim Retrieval, sondern „a priori“ bei der Indexierung erzeugt wird, ist die „Publication Timeline“ der kostenfreien, bibliografischen Datenbank *WorldCat Identities* der Fa. *OCLC*, die zu jedem Personendatensatz eingeblendet wird und das Publikationsgeschehen von und über die jeweilige Person anzeigt[161].

Rangordnungen und Zeitreihen sind Rankingverfahren, die in der Praxis eine wesentlich größere Bedeutung haben, als das anspruchsvollere Relevance Feedback. Die Ursachen dafür liegen vor allem in der einfacheren Realisierung. Rangordnungen und Zeitreihen sind nicht auf die Bestimmung der Relevanz durch vektorielle Repräsentationen und probabilistische Schätzungen angewiesen, sie arbeiten auf der Grundlage eines einfachen Faktenretrievals. Die formelle, boole’sche Anfragesprache ermöglicht zudem präzisere Formulierungen für das Retrieval einer vollständigen Dokumentenmenge, aus der eine aussagefähige Rangordnung erzeugt werden soll.

Beide Verfahren sind allerdings auf eine umfangreiche, qualitativ hochwertige Datenbasis angewiesen. Der Zugang zu solchen Datenbanken ist in der Regel kostenpflichtig und wird durch kommerzielle Anbieter kontrolliert, die die Retrievalfunktionalitäten zur Verfügung stellen. Diese sogenannten „Datenbankhosts“ ermöglichen die Erstellung von Rangordnungen oder Zeitreihen durch dialogorientierte, interaktive Verfahren, entweder befehlsgeführt durch die Eingabe einer streng formalisierten Retrievalsprache oder menügeführt durch eine grafische Oberfläche.

5.5 Assoziatives Retrieval

Als „assoziatives Retrieval“ wird eine Verfahrensklasse bezeichnet, die, wie auch die informetrischen Analysen, auf der Ebene der bloßen Informationen arbeitet. Während aber die Informetrie von einer Dokumentenmenge ausgeht, operiert das assoziative Retrieval auf der Grundlage eines einzelnen Dokuments. Dieses einzelne Dokument wird als hochrelevantes „Musterdokument“ deklariert, das dem Informationsbedarf des Nutzers ideal

[161] Online: http://orlabs.oclc.org/Identities [Stand: 2008, Abrufdatum: 14.07.2008]

entspricht. Zu diesem Musterdokument werden dann weitere, ähnliche Dokumente gesucht und dem Nutzer präsentiert.

Ausgangspunkt dieser Vorgehensweise ist die Clusterhypothese (vgl. 5.2.2), derzufolge einander ähnliche Dokumente zur gleichen Anfrage relevant sind. Griffiths et al. weisen darauf hin, dass diese Annahme einfach geprüft werden kann: in einer Dokumentenkollektion, deren relevante und irrelevante Dokumente zu einer gegebenen Anfrage bekannt sind, muss die Ähnlichkeit aller relevanten Dokumente zueinander und aller irrelevanten Dokument zueinander errechnet werden (unter Verwendung eines geeigneten Ähnlichkeitsmaßes im Vektorraummodell). Ist nun die Hypothese zutreffend, dann muss die Ähnlichkeit aller relevanten Dokumente zueinander größer sein, als die Ähnlichkeit der irrelevanten Dokumente[162].

Daneben bestehen alternative Ansätze, die die Ähnlichkeit zwischen zwei oder mehreren Dokumenten nicht auf der Basis der Dokumentvektoren, sondern anhand gemeinsam verwendeter Terminologie, gemeinsamer Referenzen oder gemeinsamer Zitationen ermitteln[163]. Auch diese Verfahren folgen der Idee der Clusterhypothese: ein einzelnes, relevantes Dokument soll den Zugang zu vielen weiteren, relevanten Dokumenten eröffnen.

5.5.1 Clusterbasierter Ansatz

Ananlog zu den unter Abschnitt 5.5.2 beschriebenen Termclustering lässt sich aus einer Dokument-Term-Matrix ebenfalls eine Dokument-Dokument-Matrix erzeugen, aus der ein Algorithmus einander ähnliche Dokumente in Cluster einteilt. Für das Retrieval wird der Vektor der Anfrage mit den Vektoren der Clusterzentroide verglichen, die in diesem Fall keine fiktiven Terme, sondern fiktive oder reale Musterdokumente repräsentieren. Entscheidend für das Zustandekommen der Ergebnismenge ist dann nicht mehr die Ähnlichkeit zwischen der Anfrage und den einzelnen Dokumenten, sondern die Ähnlichkeit

---

[162] Griffiths, Alan; Luckhurst, Claire H.; Willett, Peter: Using interdocument similarity information in document retrieval systems. In: Journal of the American Society for Information Science 37(1986)1, S. 4

[163] Stock: Information Retrieval, S. 485-486

zwischen dem Musterdokument und den Dokumenten des entsprechenden Clusters[164].

Wie für das Termclustering stehen auch für das Dokumentenclustering zahlreiche Algorithmen zur Verfügung. Griffiths et al. haben den Einsatz verschiedener Algorithmen für ein Retrieval anhand von Dokumentähnlichkeiten getestet und sind zu dem allgemeinen und intuitiv naheliegenden Ergebnis gelangt, dass ein Dokumentencluster umso präziser durch ein Musterdokument repräsentiert werden kann, je kleiner es ist, also je weniger Dokumente es umfasst. Die kleinsten möglichen Cluster beinhalten schließlich nur noch ein einzelnes Dokument und das dazu jeweils ähnlichste Dokument, den sogenannten „nächsten Nachbarn“. Dabei kann ein Dokument durchaus in mehreren Clustern vorkommen, da zwei Dokumente in einem Cluster nicht zwingend beide die nächsten Nachbarn des jeweils anderen sein müssen[165].

5.5.2 Terminologischer Ansatz

Durch den terminologischen Ansatz werden Dokumentähnlichkeiten auf der Grundlage gemeinsamer Terminologie bestimmt. Dies entspricht im Wesentlichen der Idee des Relevance Feedback, mit dem Unterschied, dass die Dokumentähnlichkeit nur zu einem einzigen Dokument hin (dem Musterdokument) bestimmt wird.

Stock beschreibt das Verfahren als Termextraktion aus dem Musterdokument. Für die extrahierten Terme werden anhand der TF*IDF-Formel die Gewichte berechnet, durch die die Terme in ein Ranking gebracht werden können. Die bestplatzierten Terme werden dann als neue Anfrage verwendet (die Anzahl der Terme, die aus dem Ranking für die Anfragemodifikation zu entnehmen sind, muss entweder systemseitig oder nutzerseitig festgelegt werden). Die modifizierte Anfrage ist als natürlichsprachige Suche einfach zu formulieren, die extrahierten Terme können als Abfolge einzelner Wörter oder Phrasen aneinander gereiht werden. Da das Verfahren ausschließlich auf der Termextraktion basiert, wird eine umfangreiche Datenbasis benötigt, aus der

---

[164] Vgl. Griffiths; Luckhurst; Willett: Using interdocument similarity information in document retrieval systems, S. 3
[165] Ebd., S. 7-8

sich geeignete, entscheidungsstarke Terme gewinnen lassen – ideal ist eine Volltextindexierung[166].

5.5.3 Informationsflussanalyse

Eine gänzlich andere Vorgehensweise um Dokumentähnlichkeiten aufzuspüren bieten Informationsflussanalysen. Es handelt sich dabei, ebenso wie bei Rangordnungen und Zeitreihen, um ein originär informetrisches Verfahren, das jedoch auch für ein einfaches, das heißt nicht-informetrisches Retrieval eingesetzt werden kann.

Unter einem Informationsfluss werden formale Zitate verstanden, wie sie vor allem in wissenschaftlichen Arbeiten, Patentschriften und Gerichtsurteilen enthalten sind. Je nach Flussrichtung wird von „Referenzen“ oder von „Zitationen“ gesprochen. Referenzen verlaufen von der zitierenden zur zitierten Literatur, Zitationen verlaufen umgekehrt, von der zitierten zur zitierenden Literatur[167].

Die Bedeutung der Informationsflussanalyse für das assoziative Retrieval beruht auf der Annahme, dass sich einander inhaltlich ähnliche Dokumente durch gemeinsames Vorkommen formaler Zitate ebenso zusammenführen lassen, wie durch clusterbasierte oder terminologische Verfahren. Dabei wird der Ähnlichkeitsgrad zweier Dokumente allein aufgrund der Häufigkeit der gemeinsamen Referenzen oder Zitationen bestimmt, was im Gegensatz zu den clusterbasierten, bzw. terminologischen Verfahren den Vorteil der Sprachunabhängigkeit mit sich bringt[168]. Dokumentähnlichkeiten können damit auch über mehrsprachige Kollektionen ermittelt werden, dies wäre unter Verwendung eines der beiden anderen Verfahren nur mit einem erheblichen Aufwand realisierbar.

Sowohl für die Nutzung der Referenzen, wie auch für die Nutzung der Zitationen wurden jeweils eigene Vorgehensweisen entwickelt: die bibliografische Kopplung und die Kozitation.

---

[166] Vgl. Stock: Natürlichsprachige Suche – More like this! In: Password: 13(1998)11, S. 27

[167] Vgl. Stock: Information Retrieval, S. 179-181 und S. 487

[168] Vgl. Garfield, Eugene: Announcing the SCI compact disc edition: CD-ROM gigabyte storage technology, novel software, and bibliographic coupling make desktop research and discovery a reality. In: Current Comments 11(1988)22, S. 161

5.5.3.1 Bibliografische Kopplung

Bereits in den frühen 1960er Jahren formulierte Kessler das Prinzip der bibliografischen Kopplung. Eine einzelne Referenz, die von zwei Dokumenten zitiert wird, wird definiert als Verbindung zwischen diesen beiden Dokumenten – sie sind bibliografisch aneinander gekoppelt. Basierend auf dieser Beziehung definiert Kessler zwei mögliche Formen der Kopplung[169]:

- Eine Anzahl von Dokumenten gehört einer Gruppe an, wenn jedes Dokument dieser Gruppe mindestens eine gemeinsame Referenz mit irgendeinem anderen Dokument dieser Gruppe enthält.
- Eine Anzahl von Dokumenten gehört einer Gruppe an, wenn jedes Dokument dieser Gruppe mindetens eine gemeinsame Referenz mit einem bestimmten Musterdokument enthält. Die Intensität des Zusammenhangs zwischen dem Musterdokument und jedem Dokument der Gruppe wird gemessen durch die Anzahl der gemeinsamen Referenzen.

Die zweite Kopplungsform lässt sich unmittelbar auf ein assoziatives Retrieval anhand eines Musterdokuments übertragen. Wird jedes Dokument einer Kollektion als Musterdokument zu den restlichen Dokumenten bestimmt, so ergibt sich eine Strukturierung nach Dokumentähnlichkeiten, die vergleichbar ist mit einem Clustering-Verfahren.

Dabei kann das Musterdokument den Zugang zu älteren, wie auch zu aktuelleren Dokumenten eröffnen, lediglich solche Referenzen, die auf Literatur verweisen, die aktueller ist als das Musterdokument, können bei der Ähnlichkeitsbestimmung nicht berücksichtigt werden.

5.5.3.2 Kozitation

Im Gegensatz zur bibliografischen Kopplung geht die Kozitation von einer Ähnlichkeit zwischen zwei Dokumenten aus, wenn sie gemeinsam zitiert

---

[169] Kessler, Meyer M.: Bibliographic coupling netween scientific papers. In: American Documantation 14(1963)1, S. 10

werden. Die dadurch entstehenden Muster unterscheiden sich signifikant von den Mustern, die durch bibliografische Kopplung entstehen[170].

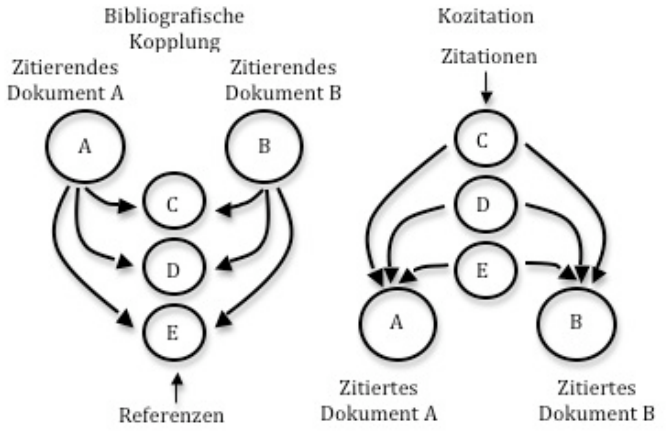

Abbildung 10: Bibliografische Kopplung und Kozitation
Quelle: Garfield: Announcing the SCI compact disc edition, S. 162. Die Abbildung wurde verändert.

Der angenommene Ähnlichkeitsgrad zwischen zwei Dokumenten bestimmt sich durch die Häufigkeit, mit der beide Dokumente gemeinsam zitiert werden. Durch Zitationsindizes wie den bereits erwähnten *Science Citation Index* können solche Zusammenhänge ermittelt werden[171].

Für einen intensiven Zusammenhang zwischen zwei Dokumenten, müssen diese beiden Dokumente von einer entsprechend großen Anzahl von Verfassern in aktuelleren Publikationen häufig gemeinsam zitiert werden. Diese Abhängigkeit macht die Kozitation zu einem dynamischeren Verfahren, das die jeweiligen Interessen eines Fachgebietes stärker reflektiert, als dies durch bibliografische Kopplungen möglich ist.

[170] Vgl. Small, Henry: Co-citation in the scientific literature: a new measure of the relationship between two documents. In: Journal of the American Society for Information Science 24(1973)4, S. 265
[171] Vgl. ebd.

6. Anwendungsbeispiele

In diesem Abschnitt werden die zuvor beschriebenen theoretischen Aspekte der Anfragemodifikation durch konkrete Anwendungsbeispiele illustriert. Produkte, die ähnliche oder vergleichbare Funktionen aufweisen, werden mitunter gemeinsam abgehandelt. Insgesamt decken die ausgewählten Beispiele ein Spektrum von informationslinguistischen Funktionen, verschiedenen Realisierungen von semantischer Umfeldsuche, informetrischen Analysen, sowie assoziativem Retrieval ab.

Es wurden sowohl frei verfügbare, wie auch kommerzielle Angebote aus den Bereichen der Websuchmaschinen, der bibliografischen Datenbanken und der Volltextdatenbanken berücksichtigt.

6.1 *LexiQuo* und *LexiLib*

LexiQuo und *LexiLib* sind Schwesterprojekte, die kooperativ durch die Fa. *Textec* und die Fa. *Media Nova* entwickelt wurden, bei beiden Projekten handelt es sich um frei verfügbare Angebote[172]. *LexiQuo* ist eine Metasuchmaschine, die eine Anfrage an die Suchmaschinen *Yahoo*, *MSN*, *Google* und *Exalead*, sowie an die deutschsprachige *Wikipedia* weiterreicht. *LexiLib* bietet eine Metasuche über drei Bibliothekskataloge an: den Online-Katalog der Deutschen Nationalbibliothek, der die Bestände aus den Standorten Leipzig und Frankfurt am Main beinhaltet, den Online-Katalog („Integrated Catalogue“) der British Library und den Online-Katalog der Library of Congress.

Beide Projekte reichern die Anfrage um flektierte Wortformen zu jedem eingegebenen Suchterm an. Der Funktionsumfang von *LexiQuo* und *LexiLib* ist identisch, so dass im Folgenden *LexiQuo* exemplarisch für beide Projekte dargestellt wird. Die Angaben sind der Homepage der Fa. *Textec* entnommen, die in den Gemeinschaftsprojekten für das linguistische Retrieval verantwortlich ist[173].

---

[172] Online: http://www.lexiquo.net [Abrufdatum: 15.07.2008, Datei: index.html] und http://www.lexilib.de [Abrufdatum: 15.07.2008, Datei: index.html]

[173] Vgl. Homepage der Fa. Textec. Online: http://www.textec.de [Stand: Januar 2008, Abrufdatum: 15.07.2008]

LexiQuo basiert auf der Software „Extrakt“, ein „umfassendes modulares System für die Behandlung von natürlicher (geschriebener) Sprache“[174]. „Extrakt“ wird seit Anfang der 1990er Jahre entwickelt und fand erstmals in dem mehrsprachigen „EMIR“-Retrievalsystem (European Multilingual Information Retrieval) für das Deutsche Anwendung. In der Zwischenzeit unterstützt „Extrakt“ neben dem Deutschen die Sprachen Englisch, Französisch, Italienisch, Dänisch, Niederländisch, Griechisch, Spanisch, Portugiesisch, Polnisch und Latein. Neben dem Aspekt der Mehrsprachigkeit wird durch „Extrakt“ eine morphologische Analyse für *LexiQuo* bereitgestellt, durch die die einzelnen Suchterme einer Anfrage automatisch erweitert werden, entsprechend dem in Abbildung 4 rechts dargestellten Schema.

Bei der Eingabe eines Suchterms wird dessen Grundform ermittelt. Dies geschieht anhand eines Vollformwörterbuchs, das für das Deutsche nach Selbstauskunft von *Textec* etwa eine Million Einträge umfassen soll[175]. Von der identifizierten Grundform ausgehend, lassen sich die übrigen Wortformen des Terms auffinden. Für das Retrieval stehen dann neben der Grundform auch die verschiedenen Wortformen zur Verfügung. Für einen substantivischen Suchterm wären dies alle Formen, die sich aus den Kategorien des Numerus, Genus und Kasus ergeben, bei einem Suchterm einer anderen Wortklasse, etwa einem Adjektiv, sind die zu berücksichtigenden Wortformen entsprechend umfangreicher, da sich in diesem Fall allein durch die Steigerung vom Positiv auf den Komparativ und den Superlativ viele unterschiedliche Formen ergeben können.

Ergänzt wird das System außerdem durch Wörterbücher, in denen für Wortformen mit Umlauten alternative Schreibweisen in nicht-umgelauteter Form hinterlegt sind. Die Zerlegung von Komposita wird ebenfalls durch ein entsprechendes Wörterbuch kontrolliert, die Anzahl der dort erfassten Einträge wird mit etwa einer Million angegeben[176].

---

[174] Ebd.
[175] Ebd.
[176] Ebd.

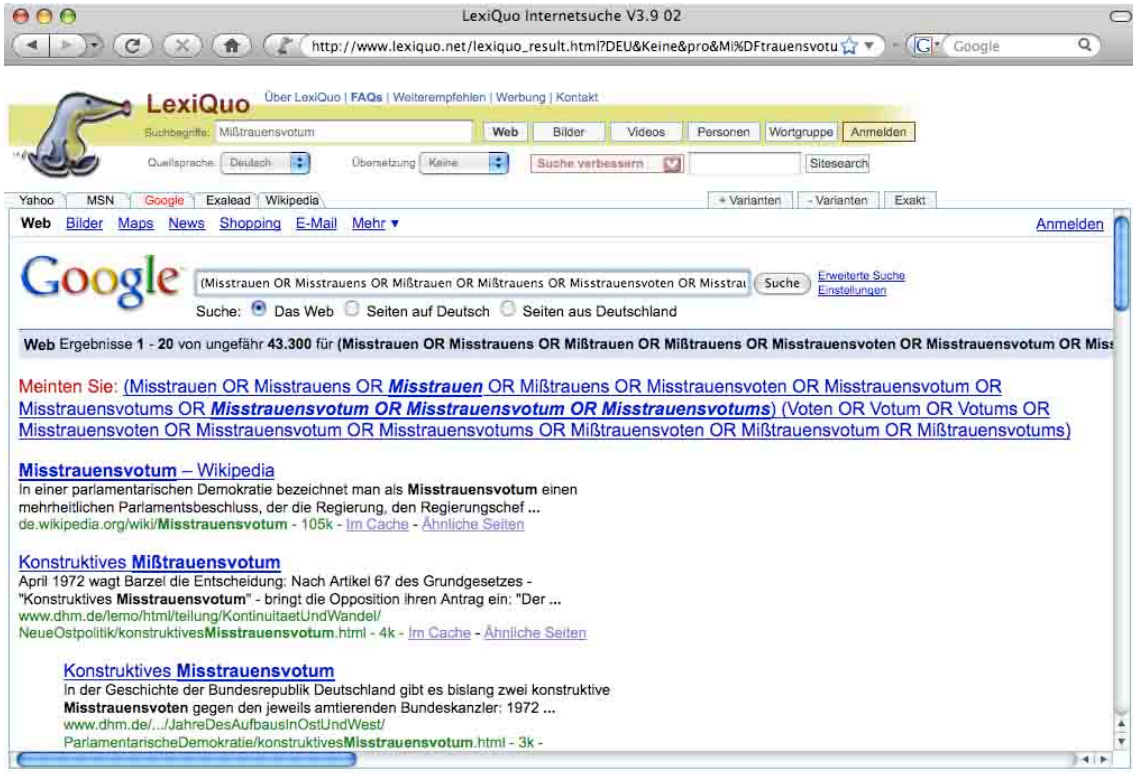

Abbildung 11: Morphologische und syntaktische Analyse bei *LexiQuo*
Quelle: N.N. 6a

Abbildung 11 zeigt eine, durch *LexiQuo* modifizierte und an *Google* weitergeleitete Anfrage. Die initiale Anfrage bestand aus dem einzelnen Suchterm „Mißtrauensvotum“ im Nominativ Singular. In einem ersten Schritt wird der Suchterm als Kompositum erkannt und erfolgreich in „Mißtrauen“ und „Votum“ zerlegt. Die Anfrage in *Google* lautet dann: „(Mißtrauen ODER Mißtrauensvotum)(Votum ODER Mißtrauensvotum)“. Gesucht wird demnach mit zwei Vereinigungsmengen, die sich jeweils aus dem Kompositum und einer Konstituente zusammensetzen, zwischen diesen beiden Mengen verwendet *Google* standardmäßig eine UND-Verknüpfung[177].

Durch einen Klick auf die Fläche „+Varianten“ im rechten oberen Bildbereich wird die Anfrage schließlich auf den in Abbildung 11 dargestellten Umfang ausgedehnt. In beiden Vereinigungsmengen werden sowohl für das Kompositum, wie auch für die jeweilige Konstituente die verschiedenen, flektierten Formen berücksichtigt. Außerdem wird für „Mißtrauen“ die alternative Schreibweise „Misstrauen“ hinzugefügt. Einzig der Suchterm „Votum“ ist auf die

[177] Vgl. Stock: Information Retrieval, S. 152

Pluralform „Voten“ beschränkt, der ebenfalls zulässige, wenn auch weniger gebräuchliche Plural „Vota“ bleibt außen vor.

Die durch „Meinten Sie ...“ eingeleitete *Google*-Rechtschreibkontrolle ist beim Einsatz von *LexiQuo* kontraproduktiv. Die Vorschläge, die *Google* in Abbildung 11 unterbreitet, beispielsweise den dritten Suchterm „Mißtrauen“ durch „Misstrauen“ zu ersetzen, also den Term, der bereits an erster Stelle steht, laufen der morphologischen Analyse von *LexiQuo* zuwider. Der Umfang der Ergebnismenge bei der reinen Kompositumszerlegung betrug etwa 69.100 Treffer, bei der umfassenden, zweiten Suche etwa 43.300 Treffer. Diese Zahlen besitzen wenig Aussagekraft, da sie sich bei wiederholter Eingabe der gleichen Anfrage immer wieder um einige hundert Treffer mehr oder weniger verändern. Überraschend ist, dass das Ergebnis der zweiten, umfassenderen Anfrage mit etwa 40.000 Treffern um mehr als ein Drittel geringer ist, als das der ersten, einfacheren Anfrage, die auf über 65.000 bis knapp 70.000 Treffer kommt. Da die zweite Anfrage zahlreiche, durch ein ODER verknüpfte Varianten verbindet und somit wesentlich offener formuliert ist als die erste Anfrage, könnten für diese Anfrage eigentlich mehr Treffer erwartet werden, tatsächlich verhält es sich umgekehrt. Die Ursachen dafür stehen allerdings weniger im Zusammenhang mit der Anfragemodifikation, sondern sind eher als ein allgemeines Phänomen des Webretrievals zu verstehen, wo derart umfangreiche und schwer interpretierbare Suchergebnisse der Regelfall sind[178].

Auch auf der Ebene der Semantik bietet *LexiQuo* Möglichkeiten der Modifikation. *Textec* verweist auf den Einsatz eines Wörterbuchs des Deutschen mit etwa 90.000 Synonymen und etwa 150.000 Einträgen aus Wortfamilien, also solchen Wörtern, die auf den gleichen Wortstamm zurückgeführt werden können[179]. Die Einbeziehung dieser semantischen Relationen kann per Klick auf das Feld „Suche verbessern“ optional gewählt werden.

---

[178] Vgl. Lewandowski: Web Information Retrieval, S. 15
[179] Vgl. Homepage der Fa. Textec

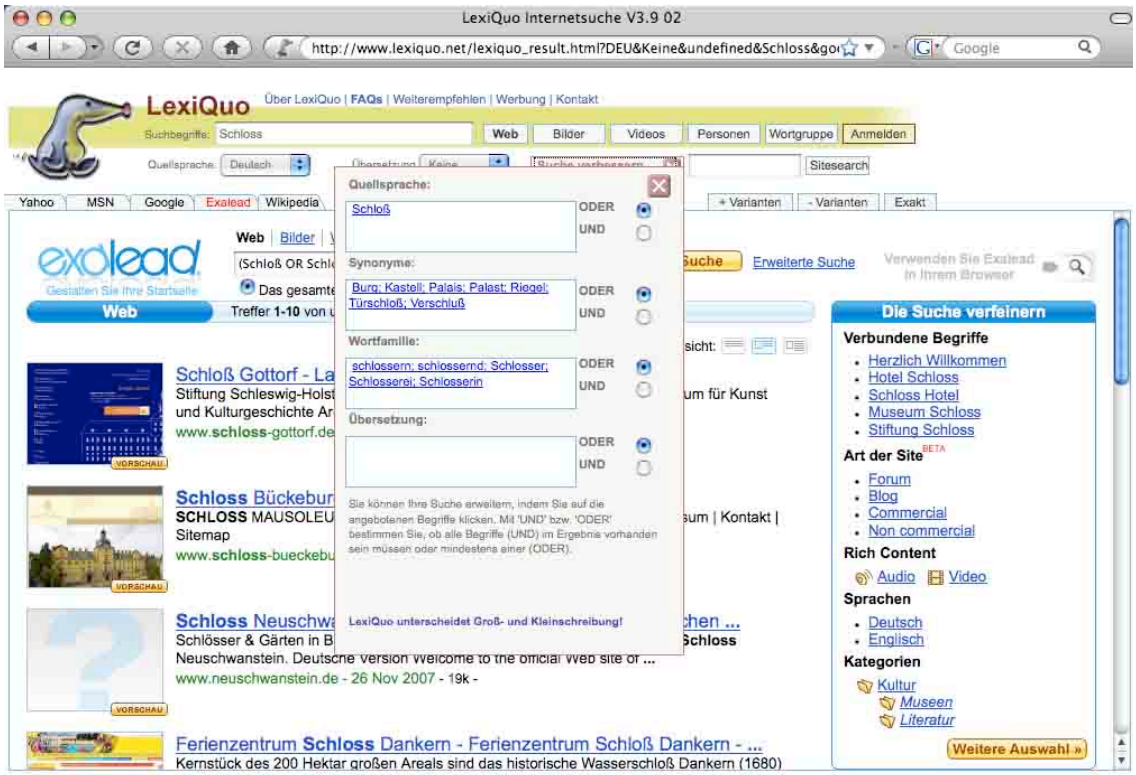

Abbildung 12: Semantische Relationen bei *LexiQuo*
Quelle: N.N. 6b

Für den eingegebenen Suchterm „Schloss“, der in Abbildung 12 an die Suchmaschine *Exalead* geschickt wurde, wird zunächst die Grundform in der Schreibweise „Schloß“ identifiziert und im oberen Fenster „Quellsprache“ angeboten. Die darunter aufgeführten Synonyme „Burg“, „Kastel“, „Palais“, „Palast“ und „Riegel“, „Türschloß“, „Verschluß“ erlauben eine Disambiguierung des Suchterms.

Durch die Zurückführung auf den Wortstamm „schloss“ ergeben sich die im Fenster „Wortfamilie“ stehenden Terme, die als assoziierte Begriffe zu „Schloss“ (im Sinne einer Schließvorrichtung) das semantische Umfeld weiter erschließen. Diese Begriffe sind der Qualität des von der Suchmaschine *Exalead* vorgeschlagenen Vokabulars überlegen. Abbildung 12 zeigt auf der rechten Seite das blauumrandete Fenster „Die Suche verfeinern“, das von *Exalead* generiert wurde. So ist beispielsweise zwischen dem Suchterm „Schloss“ und dem als „verbundener Begriff“ ausgewiesenen „Herzlich Willkommen“ kein semantischer Zusammenhang erkennbar, stattdessen deuten die von *Exalead* vorgeschlagenen Terme auf Werbeangebote hin.

Im Sinne der dialogorientierten Vorgehensweise des „adding the user to the system“-Prinzips (vgl. 5.2) ermöglicht *LexiQuo* das Hinzufügen aller vorgeschlagenen Terme zur Anfrage mit einer UND- bzw. ODER-Verknüpfung.

6.2 *Daffodil*

Daffodil („Distributed agents for user-friendly access of digital libraries“) ist ein Retrievalsystem, das am Fachgebiet Informationssysteme der Universität Duisburg-Essen von 2000 bis 2004 als Teil des DFG-Schwerpunktprogramms V3D2 („Verteilte Verarbeitung und Vermittlung digitaler Dokumente“) entstanden ist[180]. Das System *Daffodil* ermöglicht ein verteiltes Retrieval über mehrere, bibliografische Datenbanken. Die Verbreitung von *Daffodil* bleibt weit hinter der Verbreitung eines universellen Dienstes wie *LexiQuo* zurück, *Daffodil* ist ein experimentelles System, das bislang nur in der Gestalt eines Prototypen realisiert wurde und vor allem zu Evaluationszwecken dient[181].

Der *Daffodil*-Prototyp umfasst die Suche in derzeit sieben bibliografischen Datenbanken (Stand: Juli 2008) in denen Fachliteratur zum Thema Informatik nachgewiesen ist. Es besteht die Möglichkeit, *Daffodil* von der Webseite des Projekts als Java-Anwendung herunterzuladen[182]. Alternativ kann *Daffodil* in einem Browser gestartet und genutzt werden[183]. Die beiden Varianten unterscheiden sich in ihrer Oberfläche, wie auch in ihrem Funktionsumfang erheblich voneinander. Im Folgenden wird die browserunabhängige Java-Anwendung vorgestellt, die über einen größeren Funktionsumfang verfügt.

---

[180] Vgl. Homepage von Daffodil. Online: http://www.daffodil.de [Abrufdatum: 16.07.2008]

[181] Vgl. Klas, Claus-Peter; Kriewel, Sascha; Fuhr, Norbert et al.: Daffodil – Nutzerorientiertes Zugangssystem für heterogene Digitale Bibliotheken. In: Ockenfeld, Marlies (Hrsg.): Leitbild Informationskompetenz. Positionen – Praxis – Perspektiven im europäischen Wissensmarkt, 2005, S. 171

[182] Vgl. Homepage von Daffodil. Online: http://www.daffodil.de [Abrufdatum: 21.07.2008, Datei: http://www.is.informatik.uni-duisburg.de/projects/daffodil/pdaffodil.jnlp]

[183] Vgl. ebd. Online: http://www.daffodil.de [Abrufdatum: 21.07.2008, Datei: http://andy.is.informatik.uni-duisburg.de:8090/daffodil-web]

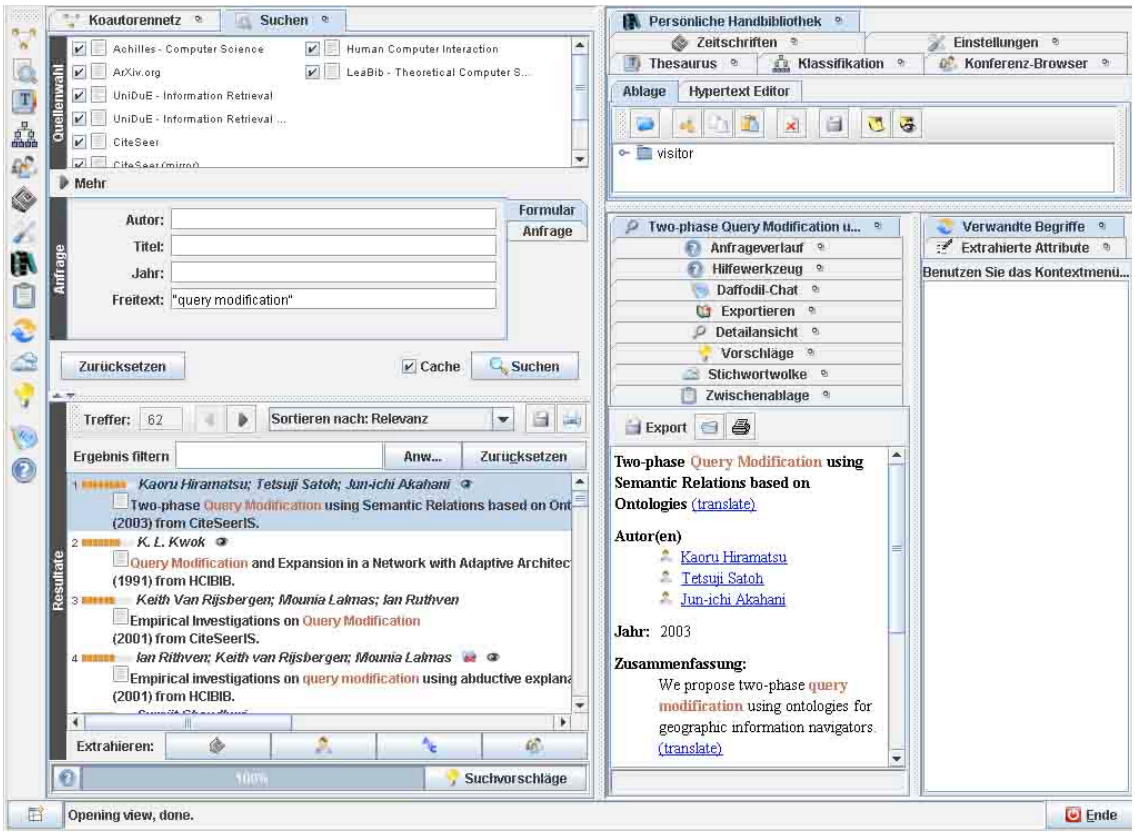

Abbildung 13: *Daffodil*-Desktop

Der *Daffodil*-Desktop verfügt über zwei Spalten. In der linken Spalte können zunächst nach dem Checkbox-Verfahren die Datenbanken ausgewählt werden, die in eine Suche einbezogen werden sollen. In der Mitte der linken Spalte befindet sich das Eingabeformular, im unteren Bereich werden die Suchergebnisse angezeigt.

In der rechten Spalte des Desktops sind die verschiedenen Funktionen angeordnet, die zur Unterstützung einer Suche in *Daffodil* herangezogen werden können. Das System bietet ein umfangreiches Spektrum an Instrumenten zur Anfragemodifikation, von einfachen Plausibilitätskontrollen bei der Eingabe, über einen Thesaurus- und einen Klassifikationsbrowser, bis hin zur automatischen Erzeugung von Vorschlägen zur Verbesserung der Suchergebnisse[184]. Ein assoziatives Retrieval auf terminologischer Grundlage, wie auch anhand von Referenzen und Zitationen ist in *Daffodil* angelegt, allerdings zum Zeitpunkt der Erprobung nicht funktionsfähig.

---

[184] Schaefer, Andreas; Jordan, Matthias; Klas, Claus-Peter et al.: Active support for query formulation in virtual digital libraries: a case study with Daffodil. In: Rauber, Andreas (Hrsg.): Research and Advanced Technologies for Digital Libraries. 9th European Conference. Proceedings, 2005, S. 2

Im Zusammenspiel sollen diese Instrumente dazu beitragen, dass einfache, syntaktische Fehler bei der Eingabe verhindert werden, dass die kognitive Belastung des Nutzers bei der Modifikation der Anfrage möglichst gering gehalten wird und dass der Nutzer Zutrauen zum Konzept der iterierten Suche, wie auch zu den einzelnen formulierten Anfragen fasst[185].

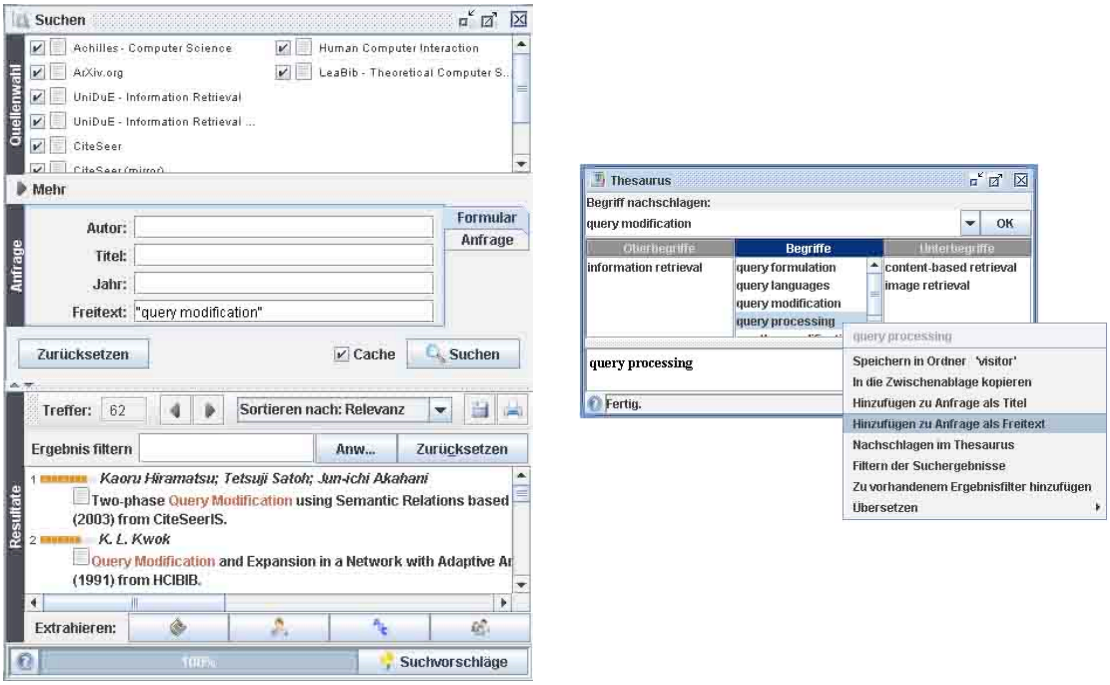

Abbildung 14: *Daffodil*-Suchwerkzeug und Thesaurus-Browser

Abbildung 14 zeigt das Suchwerkzeug und den Thesaurus-Browser. Nach einer initialen Anfrage mit der Phrase „query modification“ kann nach der gleichen Phrase im Thesaurus gesucht werden. Der Deskriptor „query modification“ wird innerhalb seines semantischen Umfelds angezeigt, durch Verwendung des Kontextmenüs kann ein beliebiger Deskriptor zur Anfrage in das Eingabeformular hinzugefügt werden. Entsprechend den Ausführungen in Abschnitt 5.2.1 sollte bei der Anfrageerweiterung vor allem auf die Transitivität der neuen Deskriptoren zum ursprünglichen Suchterm geachtet werden. Neue Suchterme werden automatisch mit einem boole’schen UND an die initiale Anfrage angehängt. Auch eine Erweiterung des Thesaurus um lexikalische

[185] Vgl. ebd.

Datenbanken wie *WordNet* soll möglich sein, ist im hier besprochenen Prototyp derzeit jedoch nicht realisiert[186].

Es besteht außerdem die Möglichkeit kollektionsunabhängige Begriffsordnungen ergänzend oder alternativ zum Thesaurus-Vokabular zu nutzen:

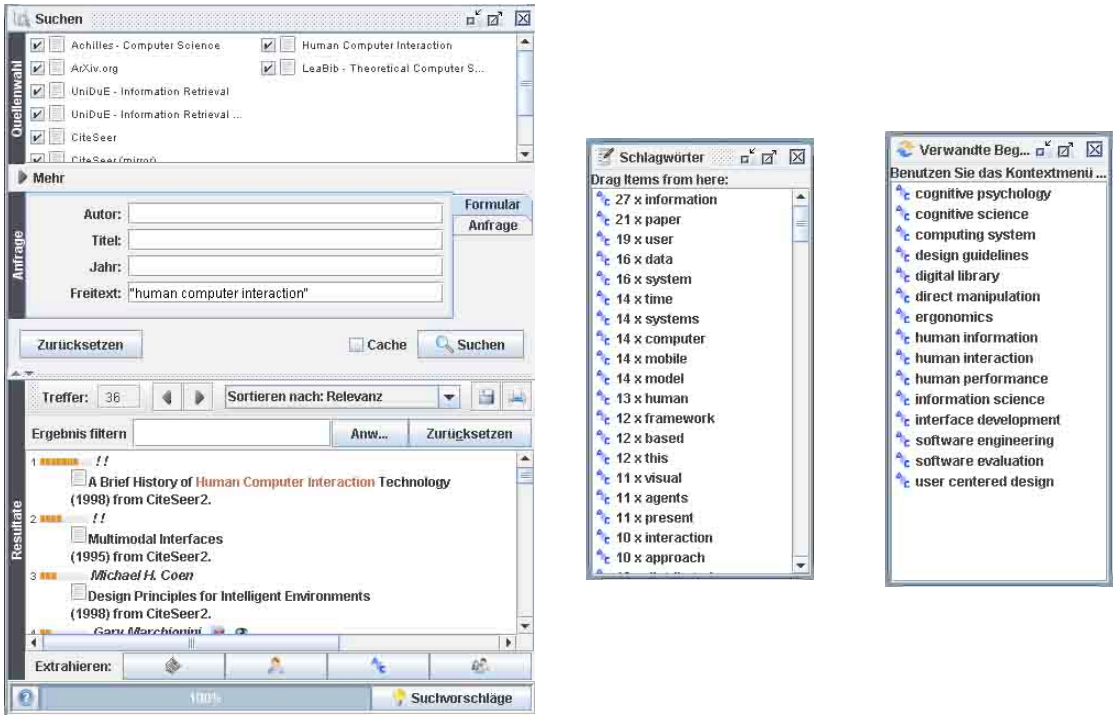

Abbildung 15: Termextraktion und ähnliche Terme bei *Daffodil*

Aus der in Abbildung 15 links dargestellten Ergebnismenge kann *Daffodil* die häufigsten Terme extrahieren und in einer Liste mit der Bezeichnung „Schlagwörter“ nach absteigender Häufigkeit aufführen. Die rechts abgebildete Liste „verwandte Begriffe“ enthält Wörter oder Phrasen, die mit den Suchtermen häufig gemeinsam auftreten. Es erfolgt außerdem ein Abgleich mit dem Thesaurusvokabular, so dass die Liste gegebenenfalls durch Synonyme, sowie Ober- und Unterbegriffe ergänzt wird[187]. Um die hochfrequenten oder die korrelierten Terme zur Anfrage hinzuzufügen, kann sowohl eine entsprechende Funktion im Kontextmenü genutzt werden (ähnlich wie in Abbildung 14), wie auch ein „drag and drop“-Vefahren, mit dem der jeweilige Term direkt in das gewünschte Eingabefeld im Suchformular platziert werden kann.

---

[186] Vgl. Klas, Claus-Peter: Daffodil. Strategische Unterstützung bei der Informationssuche in Digitalen Bibliotheken, 2007, S. 95

[187] Vgl. Schaefer; Jordan; Klas et al.: Active support for query formulation in virtual digital libraries: a case study with Daffodil, S. 8

Während die „verwandten Begriffe“ nach Eingabe einer Suche automatisch generiert werden, vorausgesetzt es liegt eine ausreichend umfangreiche Ergebnismenge vor, werden die häufig vorkommenden Terme nur auf einen Nutzerbefehl hin zusammengestellt. Neben den als „Schlagwörter“ bezeichneten, thematischen Termen besteht die Möglichkeit, Verfassernamen, Zeitschriftentitel oder die Titel von Kongressschriften (in *Daffodil* als „Konferenzen“ bezeichnet) zu extrahieren. Der Nutzen solcher Termextraktionen hängt von dem Umfang und der Qualität der vorliegenden Ergebnismenge ab und letztlich von der konsistenten Erschließung der Literatur in den einzelnen, durchsuchten Datenbanken.

Neben diesen unterschiedlichen Formen semantischer Relationen lassen sich durch das Instrument „Koautorennetz“ einfache, bibliografische Relationen zwischen gemeinsam publizierenden Verfassern ausfindig machen. Wird beispielsweise während des Anfragedialogs ein Verfasser ermittelt, dessen Werke dem Informationsbedarf entsprechen, kann das „Koautorennetz“ weitere, möglicherweise ebenso relevante Verfasser identifizieren.

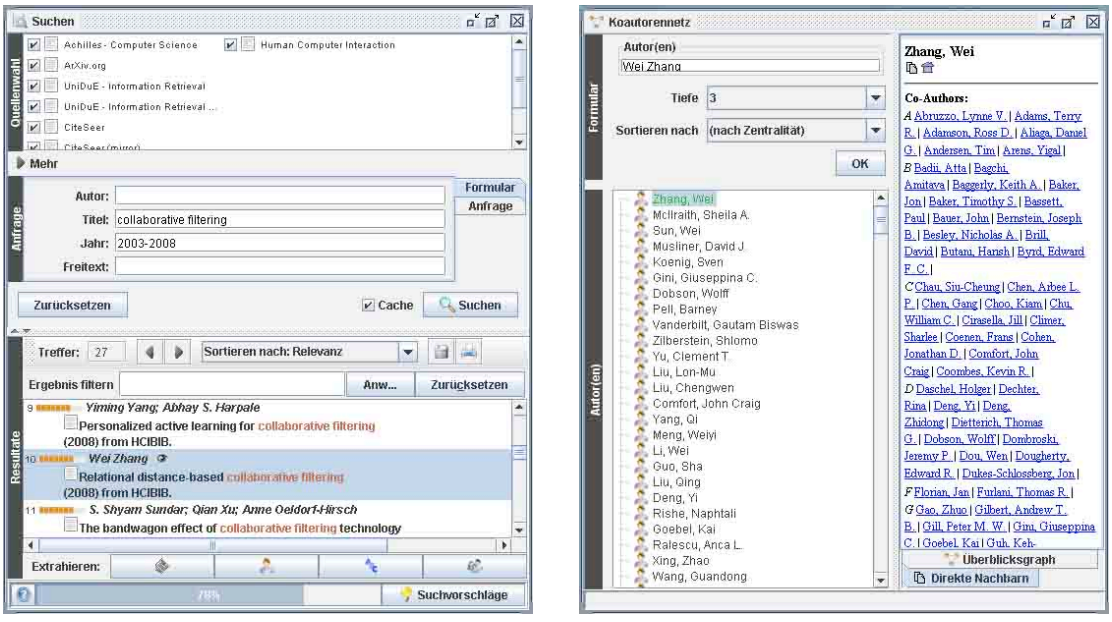

Abbildung 16: Koautorennetz bei *Daffodil*

Per „drag and drop“-Verfahren lässt sich ein beliebiges Dokument aus der Ergebnismenge in das Feld „Autor(en)“ im Fenster „Koautorennetz“ bewegen. Für den oder die Verfasser des Dokuments wird daraufhin ein Netz der Koautoren erstellt. Abbildung 16 rechts zeigt die Koautoren von Wei Zhang in

einem Ranking nach Zentralität angeordnet. Wei Zhang selbst führt das Zentralitätsranking an, als der Autor, der die meisten bibliografischen Relationen auf sich vereinigt. Durch die Visualisierung des Koautorennetzes als Überblicksgraph entsteht ein unmittelbarer Eindruck der Zentralität:

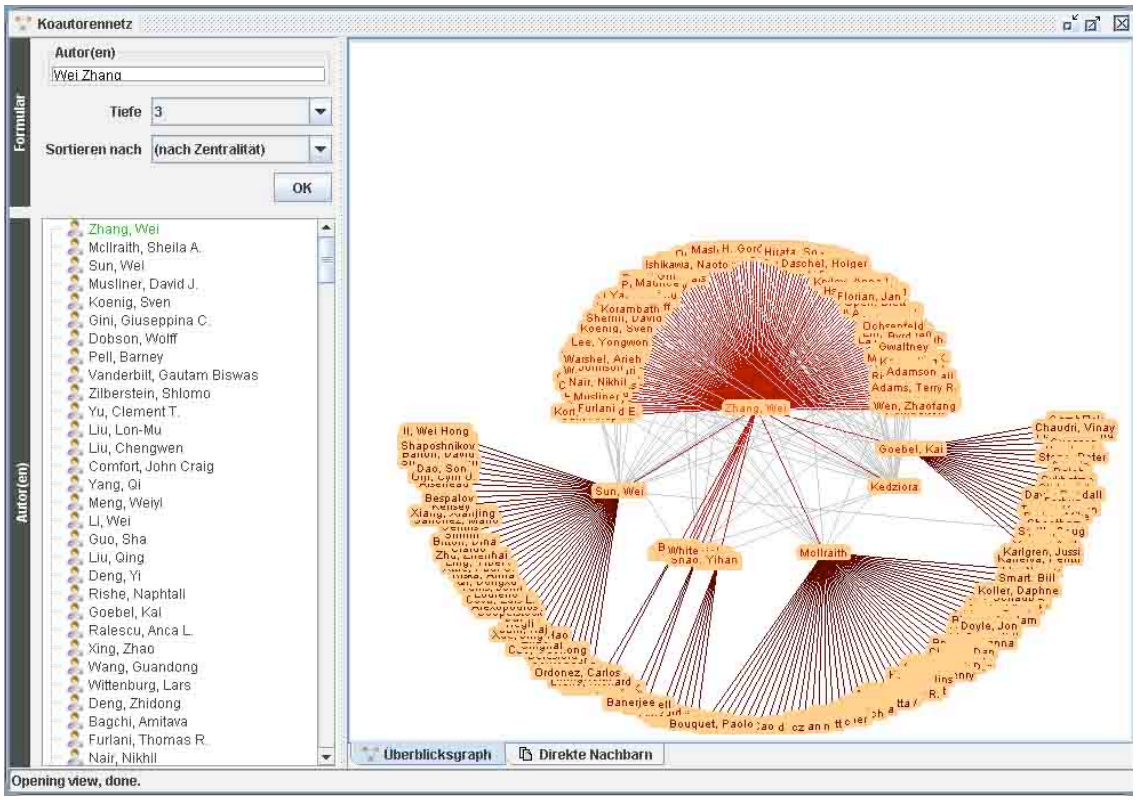

Abbildung 17: Visualisierung bibliografischer Beziehungen bei *Daffodil*

In der unteren Hälfte des Netzes gruppieren sich einige Verfasser um Zhang, für die ebenfalls eine hohe Zentralität festgestellt wurde und die sich aufgrund ihrer engen Beziehung zu Zhang für eine Anfragemodifikation empfehlen. Die *Daffodil*-Entwickler weisen darauf hin, dass durch eine Anpassung des verwendeten Algorithmus auch andere Beziehungsarten zwischen den Verfassern ermittelt und visualisiert werden könnten, etwa Zitationsbeziehungen[188].

Für einen Überblick, welche Aktionen insgesamt für eine Modifikation in Frage kommen, ist in *Daffodil* ein Vorschlagsystem integriert. Diese Funktion, die im Anschluss an eine initiale Anfrage Vorschläge zur Verbesserung der Suchergebnisse auflistet, wird von den *Daffodil*-Entwicklern besonders

[188] Klas: Daffodil, S. 91

hervorgehoben. Bezug nehmend auf Bates' Kategorisierung des Suchverhaltens in Suchschritte, Taktiken, Strategeme und Strategien, wurde ein Instrument entwickelt, das Vorschläge auf der Ebene von Taktiken und Strategemen liefert[189].

Diese „adaptive Suchunterstützung für digitale Bibliotheken" (ASDL) besteht aus drei Modulen: der Beobachtung der Nutzeraktivitäten und der Ergebnismengen, die durch das System ausgegeben werden, einer Technik des „fallbasierten Schließens", wodurch bei einem akuten Problem nach früheren, ähnlichen Problemen und deren Lösungen gesucht und deren Ähnlichkeitsgrad zum vorliegenden Problem ermittelt wird und einem Vorschlaginstrument, das die relevanten Lösungen dem Nutzer präsentiert[190].

Jede Anfrage wird gemeinsam mit der dazugehörigen Ergebnismenge als individuelle Situation gespeichert. Die Informationen, die das ASDL-System zur Beschreibung einer einzelnen Suchsituation auswertet, beinhalten die verwendeten Suchterme in den einzelnen Eingabefeldern, die eingesetzten boole'schen Verknüpfungen, die zur Suche ausgewählten Datenbanken, den Umfang der Ergebnismenge und die benötigte Antwortzeit, eine Liste der im Suchergebnis am häufigsten enthaltenen Terme, Verfassernamen, sowie die vorkommenden Zeitschriften- oder Kongressschriftentitel[191].

Die somit beschriebene Suchsituation wird mit den taktischen bzw. strategischen Vorschlägen verglichen, die in *Daffodil* gespeichert sind[192]. Jeder Vorschlag enthält bereits eine Reihe von dazu passenden Suchsituationen, die ebenfalls anhand der oben genannten Informationen beschrieben sind. Im Folgenden wird jeder einzelne Aspekt der vorgehaltenen Situationen mit dem entsprechenden Aspekt der konkreten Situation verglichen. Der Vergleich von terminologischen Gemeinsamkeiten in der Anfrage oder in der Ergebnismenge erfolgt über ein Ähnlichkeitsmaß im Vektorraum. In der Summe ergibt sich ein

---

[189] Vgl. Kriewel, Sascha; Fuhr, Norbert: Adaptive search suggestions for digital libraries. In: Goh, Dion Hoe-Lian (Hrsg.): Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers, 2007, S. 220-229

[190] Vgl. ebd., S. 222

[191] Vgl. ebd.

[192] Vgl. ebd., S. 225. Die Anzahl der vorgehaltenen Vorschläge wird mit 16 angegeben (Stand: 2007).

Ähnlichkeitsmaß zwischen der konkreten Suchsituation und früheren Suchsituationen, so dass die dazugehörigen Vorschläge nach absteigender Ähnlichkeit dem Nutzer angeboten werden können[193].

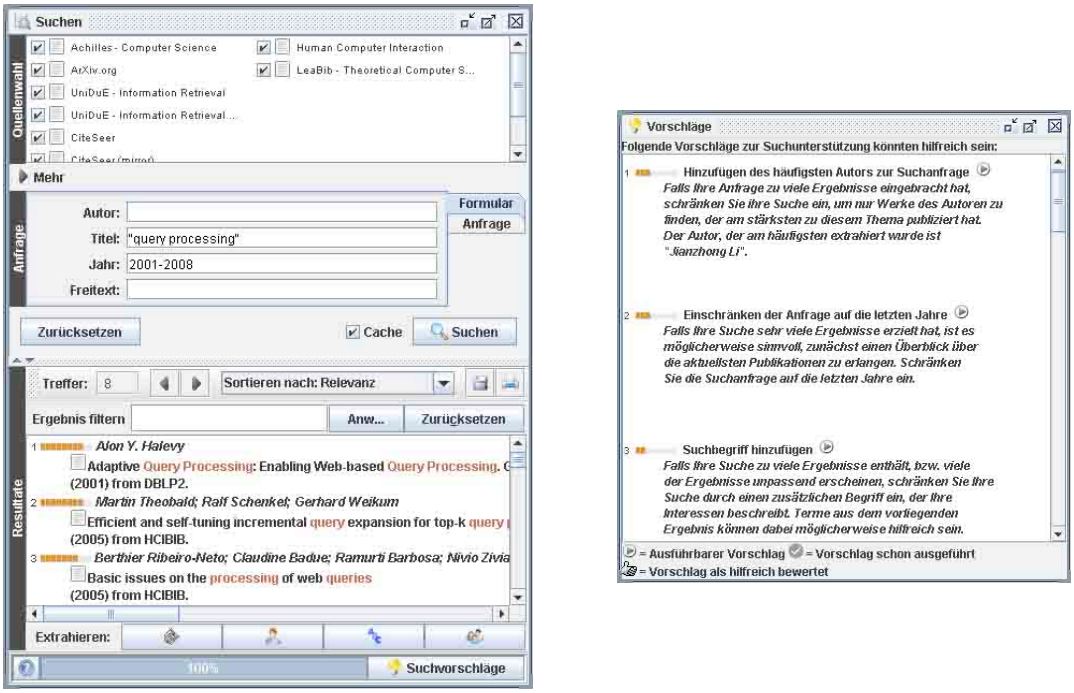

Abbildung 18: Suchvorschläge bei *Daffodil*

Per Klick auf die Schaltfläche „Suchvorschläge“ am rechten Ende der unteren Leiste des Suchwerkzeugs in Abbildung 18 wird eine Liste der Vorschläge nach dem oben beschriebenen Verfahren generiert. Vorschläge, die automatisch abgearbeitet werden können, sind gekennzeichnet durch ein entsprechendes Symbol (vgl. Abbildung 18 rechts) und lassen sich durch doppeltes Klicken ausführen. Wenn der Vorschlag für den Suchprozess nützlich war, kann eine positive Bewertung dazu abgegeben werden. Je häufiger ein Vorschlag positiv bewertet wird, desto höher wird er in späteren, vergleichbaren Situationen im Ranking platziert[194].

[193] Vgl. ebd., S. 222-224
[194] Vgl. ebd., S. 223

6.3 *Dialog* und *STN International*

Dialog und *STN* (Scientific and Technical Information Networks) sind kommerzielle Informationsanbieter, die als sogenannte „Datenbankhosts“ den Zugang zu ausgewählten Datenbanken entgeltlich ermöglichen.

Die *Dialog* Corporation, heute Teil der *Thomson*-Gruppe, besteht seit 1972 und gilt damit als ältester Online-Informationsanbieter. Allein 50 Prozent der angebotenen Datenbanken decken den Bereich von technischer und medizinischer Fachinformation ab, die übrigen 50 Prozent machen Patentdatenbanken, sowie Datenbanken zu allgemeinen Wirtschafts-informationen aus[195]. Die Summe der insgesamt angebotenen Datenbanken wird von *Dialog* derzeit mit etwa 900 beziffert[196]. *Dialog* bietet eine Vielzahl verschiedener Produkte an, die – je nachdem – Zugang zu allen 900 Datenbanken bieten, oder zu bestimmten, thematischen Datenbankclustern und die außerdem über unterschiedliche Retrievalumgebungen mit verschiedenen Funktionen und Interaktionsmöglichkeiten verfügen. Die hier dargestellten Beispiele sind dem Angebot „DialogWeb“ entnommen, eine Kombination aus befehlsgeführter Oberfläche, zu dessen Nutzung die *Dialog*-Retrievalsprache beherrscht werden muss und grafischer Oberfläche, in der auch Interaktionen per Maus möglich sind. Der Zugang erfolgte via World Wide Web durch die Nutzung eines „ONTAP“ (Online Training and Practice)-Kontos. Dieser kostenlose Zugang wird von *Dialog* für Übungs- und Demonstrationszwecke zur Verfügung gestellt. Es stehen damit nur 38 nicht-aktualisierte Datenbanken zur Verfügung, die Retrievalfunktionen können jedoch in vollem Umfang genutzt werden[197].

Der Informationsanbieter *STN International*, der kooperativ durch das FIZ Karlsruhe, den „Chemical Abstracts Service“ (CAS) in Columbus, Ohio, USA und der „Japan Association for International Chemical Information“ (JAICI) in Tokio betrieben wird, tritt mit seiner, ebenfalls technisch-

---

[195] Vgl. Stock, Mechtild; Stock, Wolfang G.: One-Stop-Shops internationaler Fachinformation. Wie gut sind Dialog / DataStar? In: Password: 18(2003)4, S. 22-23

[196] Vgl. Homepage von Dialog. Online: http://www.thomsondialog.com/about [Abrufdatum: 29.07.2008]

[197] Online: http://training.dialog.com/sem_info/ontap_pw.html [Abrufdatum: 29.07.2008, Datei: ontap_pw.html]

naturwissenschaftlichen Ausrichtung als direkter Konkurrent zu *Dialog* auf. Auch *STN* bietet verschiedene Retrievalumgebungen an, durch die der Zugang zu insgesamt etwa 220 Datenbanken ermöglicht wird[198]. Für diese Arbeit wurde durch das FIZ Karlsruhe ein zeitlich befristeter Testzugang zum Angebot „STN on the Web“ bereitgestellt, der nur Zugang zu sechs Trainingsdatenbanken ermöglichte, jedoch den vollen Funktionsumfang der Retrievalsprache „Messenger“ bot.

Sowohl *Dialog* wie auch *STN* bieten durch die befehlsgeführte Bedienung mittels Retrievalsprachen die Möglichkeit, aufbauend auf einer Ergebnismenge exakte informetrische Rangordnungen und Zeitreihen zu erstellen.

Bei *Dialog* lassen sich Rangordnungen durch den „rank“-Befehl erstellen. Ausgehend von einer Ergebnismenge und unter Angabe eines bestimmten Datenbankfeldes, wird der Feldinhalt nach Termhäufigkeit ausgewertet, anschließend werden die Feldinhalte nach absteigender Häufigkeit sortiert und als Ranking ausgegeben.

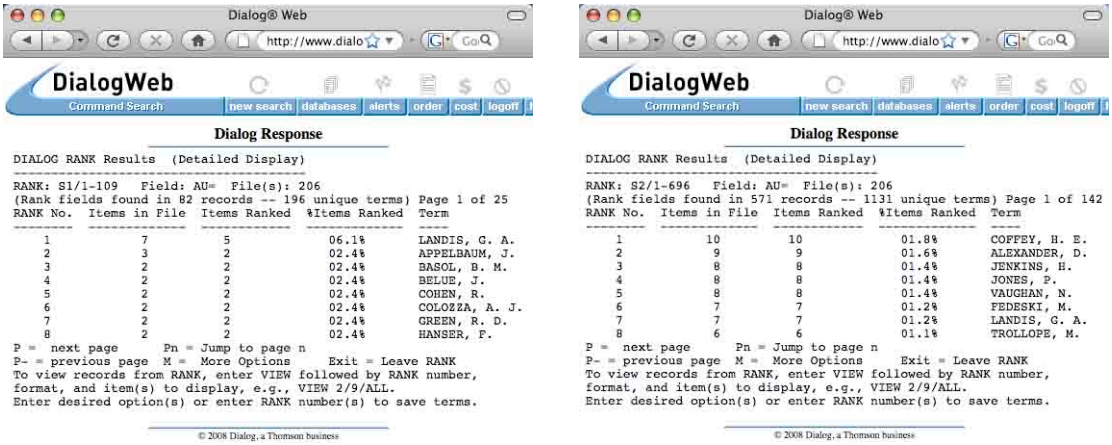

Abbildung 19: Rangordnung von Verfassern bei *Dialog*

In Abbildung 19 sind zwei Rangordnungen zweier Ergebnismengen dargestellt. Auf der linken Seite basiert die Rangordnung auf einem Suchergebnis zur initialen Anfrage mit dem Suchterm „photovoltaic“ im Basic Index der Datenbank „NTIS“ (National Technical Information Service). Auf der rechten Seite wurde in der gleichen Datenbank mit der initialen Anfrage nach „solar“

[198] Vgl. Homepage von STN International. Online: http://www.stn-international.de/stndatabases/databases/onlin_db.html [Abrufdatum: 30.07.2008, Datei: onlin_db.html]

gesucht. Durch den „rank“-Befehl wurden die Verfasserfelder aller gefundenen Datensätze ausgewertet, so dass die Verfasser mit dem höchsten Publikationsaufkommen zum Thema die Rangordnungen anführen. Da es sich bei den Suchtermen um verwandte Begriffe handelt, zeigen die beiden Rangordnungen im Vergleich bereits unter den ersten acht Positionen einen gemeinsamen Verfasser. Landis führt die Rangordnung zum Term „photovoltaic“ an und nimmt zum Term „solar“ den siebten Rang ein. Für den Nutzer, der seinen Informationsbedarf vage durch „Fotovoltaik“ und „Solar“ beschreiben würde, wäre die gezielte Suche nach den Werken von Landis anzuraten[199].

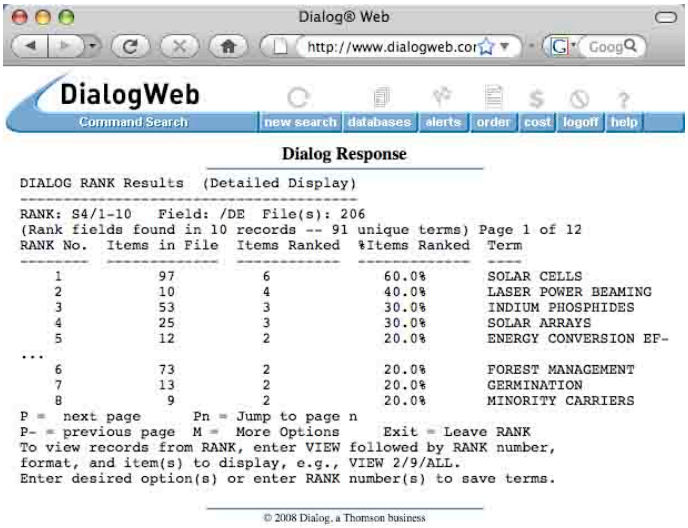

Abbildung 20: Rangordnung von Deskriptoren bei *Dialog*

Eine gezielte Suche nach dem Verfasser Landis und die Auswertung der Ergebnismenge nach Häufigkeit der zugeteilten Deskriptoren verschaffen Gewissheit über den fachlichen Schwerpunkt des Verfassers (vgl. Abbildung 20). Dieser Schritt ließe sich für die übrigen Verfasser ebenso durchführen und ist besonders dann sinnvoll, wenn umfangreiche Ergebnismengen mit mehreren hundert Treffern vorliegen, die ohne statistische Auswertung nicht beurteilt werden können. Die Analysemöglichkeiten werden durch die jeweilige Datenbank vorgegeben. *Dialog* bietet zu jeder Datenbank eine Beschreibung, in der neben der inhaltlichen Ausrichtung auch ein Musterdatensatz, sowie das

[199] Diese Aussage gilt rein hypothetisch auf das vorliegende Beispiel bezogen. Die tatsächlich produktivsten Verfasser zu den genannten Themen lassen sich durch die Dialog-Trainingsdatenbanken aufgrund von Unvollständigkeit und mangelnder Aktualität nicht ermitteln und sind im Kontext dieser Arbeit auch unerheblich.

Kategorienschema und die Indexierungsart der einzelnen Felder aufgeführt sind – der „rank“-Befehl kann zwar grundsätzlich auf numerische, wort- und phraseninvertierte Felder angewendet werden, dennoch sind je nach Datenbank einzelne Felder vom „rank“-Befehl ausgeschlossen.

Um zwei Rangordnungen zueinander ins Verhältnis zu setzen, bietet sich die Retrievalsprache „Messenger“ von *STN* an. Durch einen „analyze“-Befehl lassen sich bestimmte Feldinhalte aus einer vorliegenden Ergebnismenge extrahieren und nach Häufigkeit anordnen. Wird diese Extraktion für zwei Felder durchgeführt, lassen sich die Feldinhalte anschließend durch den „tabulate“-Befehl in Tabellenform überführen. Aus zwei extrahierten Listen der Verfasser und der Erscheinungsjahre lässt sich dann eine Zeitreihe gewinnen.

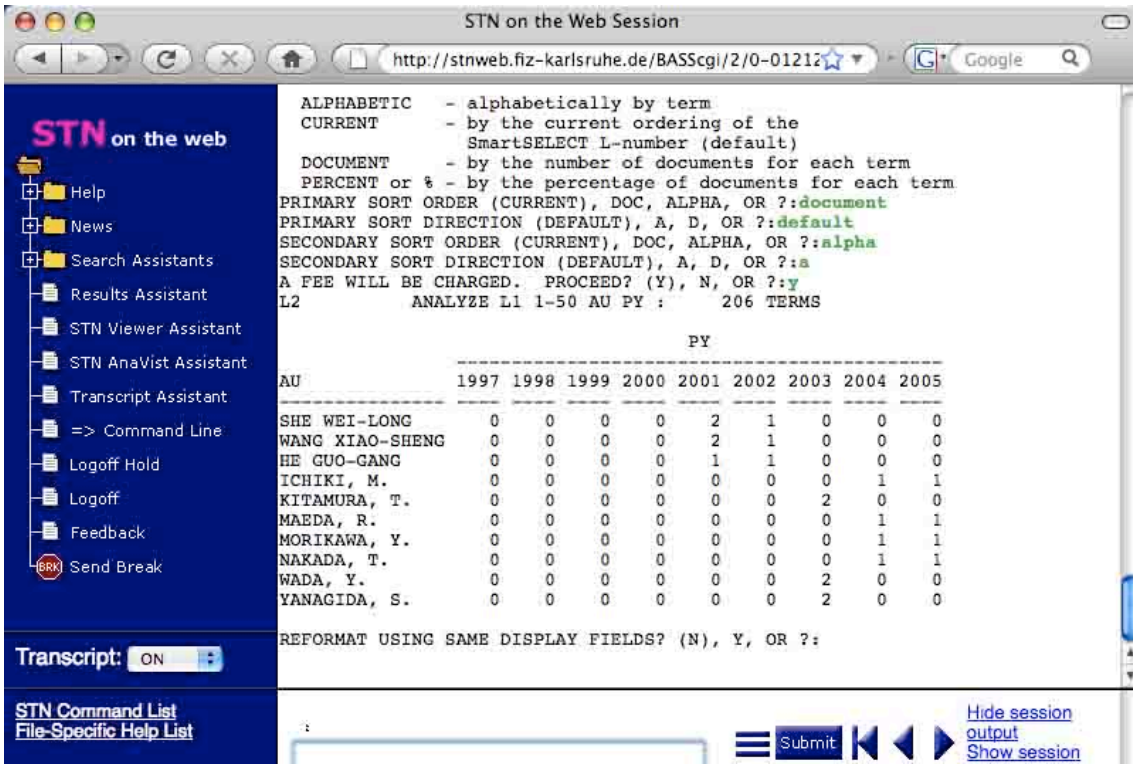

Abbildung 21: Zeitreihe bei *STN on the Web*

In einer Trainingsversion der bibliografischen Datenbank „INSPEC“ wurde eine initiale Anfrage mit dem Suchterm „photovoltaic“ durchgeführt. Per „analyze“-Befehl wurden die Feldinhalte der ersten 50 Datensätze der Felder „au“ (Verfasser) und „py“ (Erscheinungsjahr) extrahiert. Daran anschließend wurden die Listen durch den „tabulate“-Befehl und unter Angabe einer Reihe von Parametern (Ordnungskriterien, auf- oder absteigende Sortierung) in die in Abbildung 21 dargestellte Tabelle überführt. Die Tabelle stellt das

Publikationsgeschehen der zehn häufigsten Verfasser zum Thema „Fotovoltaik“ dar, die in „INSPEC“ über den Zeitraum von 1997 bis 2005 nachgewiesen sind.

Die informetrischen Analysefunktionen der Retrievalsprachen von *Dialog* und *STN* führen nicht unmittelbar zu neuen, relevanten Dokumenten. Sie können in einem iterierten Suchprozess als Zwischenschritte verstanden werden, die Informationen über die bis dahin erzielten Suchergebnisse liefern und für den weiteren Verlauf der Suche von strategischer Bedeutung sein können. Die Entwicklung neuer Suchstrategien kann sich auf Erkenntnisse stützen, die durch eine Rangordnung bzw. eine Zeitreihe gewonnen wurden.

6.4 *LexisNexis*

LexisNexis ist ein internationaler Anbieter von Fachinformationen aus den Bereichen Wirtschaft und Recht, der zum *Reed-Elsevier*-Konzern gehört. Im Gegensatz zu Anbietern wie *Dialog* und *STN International*, bietet *LexisNexis* in erster Linie Volltexte an, so dass der Zugriff durch Volltextretrieval realisiert wird. In einer Studie zu *LexisNexis* aus dem Jahr 2005 beziffern Stock und Stock das Angebot auf etwa 4,5 Milliarden Volltextdokumente aus mehr als 32.000 Quellen[200]. Neben wirtschaftlichen und juristischen Fachinformationen wie Firmen- und Länderprofile, Rechtsnormen, Urteile, Fachliteratur und Fachpresse, zählen auch Volltexte aus Zeitungen und Zeitschriften der Publikumspresse zum Bestand. Aus dem gesamten Dokumentenbestand generiert *LexisNexis* verschiedene Produkte, die auf verschiedene fachliche Zielgruppen (beispielsweise Juristen, Wirtschaftswissenschaftler oder Unternehmer), wie auch auf einzelne nationale Märkte ausgerichtet sind. Das folgende Anwendungsbeispiel wurde in dem Produkt „Nexis UK“ durchgeführt, dessen Schwerpunkt auf Wirtschaftsnachrichten für den britischen Markt liegt[201].

Die inhaltliche Erschließung der Volltexte erfolgt durch automatische Indexierung, da allein im Hinblick auf die Menge der Volltexte eine intellektuelle

---

[200] Stock, Mechtild; Stock, Wolfgang G.: Digitale Rechts- und Wirtschaftsinformationen bei LexisNexis. In: JurPC Web-Dok. 82(2005), Abs. 21-22

[201] Vgl. Homepage von Nexis UK. Online: http://www.lexisnexis.com/uk/business [Abrufdatum: 14.08.2008]. Zur Nutzung wurde eine temporäre Testkennung durch die LexisNexis Deutschland GmbH ausgestellt.

Erschließung nicht zu bewältigen wäre. Das als „SmartIndexing“ bezeichnete System kann Volltexte aus dem Englischen, Französischen und Deutschen auswerten und anhand der vorkommenden Terme aus einem kontrollierten Vokabular entsprechende Schlagwörter zuteilen. Die zu indexierenden Schlagwörter sind in die Kategorien „Company“, „Industry“, „Subject“, „Geography“ und „People“ unterteilt (die deutschen Konkordanzen lauten „Firma“, „Branche“, „Thema“, „Region“ und „Person“), wobei die verwendeten Schlagwortlisten durch *LexisNexis* selbst entwickelt werden und in keinem Zusammenhang zu bibliothekarischen Normdateien wie etwa der Schlagwortnormdatei oder der Personennamendatei stehen[202].

Bereits zur Formulierung der initialen Anfrage kann ein Browsing durch die Schlagwörter erfolgen, die anhand einer Hierarchie in ein semantisches Umfeld eingebettet sind:

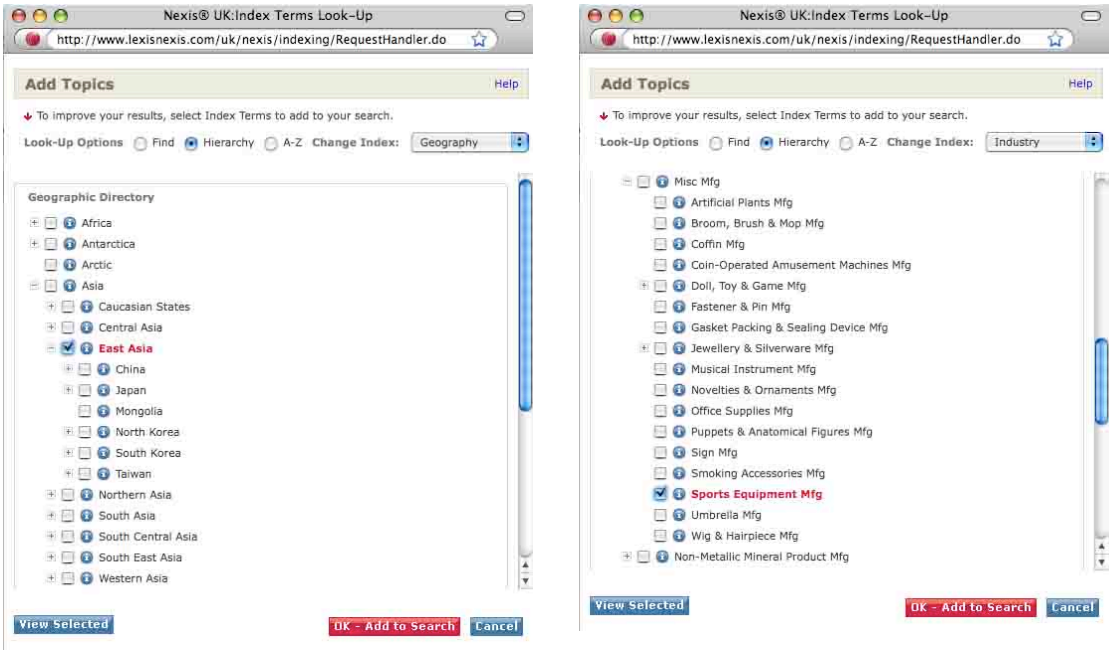

Abbildung 22: Schlagwörter bei *LexisNexis*

Per Klick auf die Schaltfläche „i“, die jeweils rechts neben der Checkbox angezeigt wird, können die Begriffsmerkmale des Schlagworts eingesehen werden.

---

[202] Die Informationen über das System „SmartIndexing“ beruhen auf telefonischen Auskünften durch Herrn Roschanski. Vgl. Gesprächsprotokoll: Telefonat LexisNexis Deutschland GmbH [Datum: 06.08.2008, Gesprächspartner: Tim Roschanski]

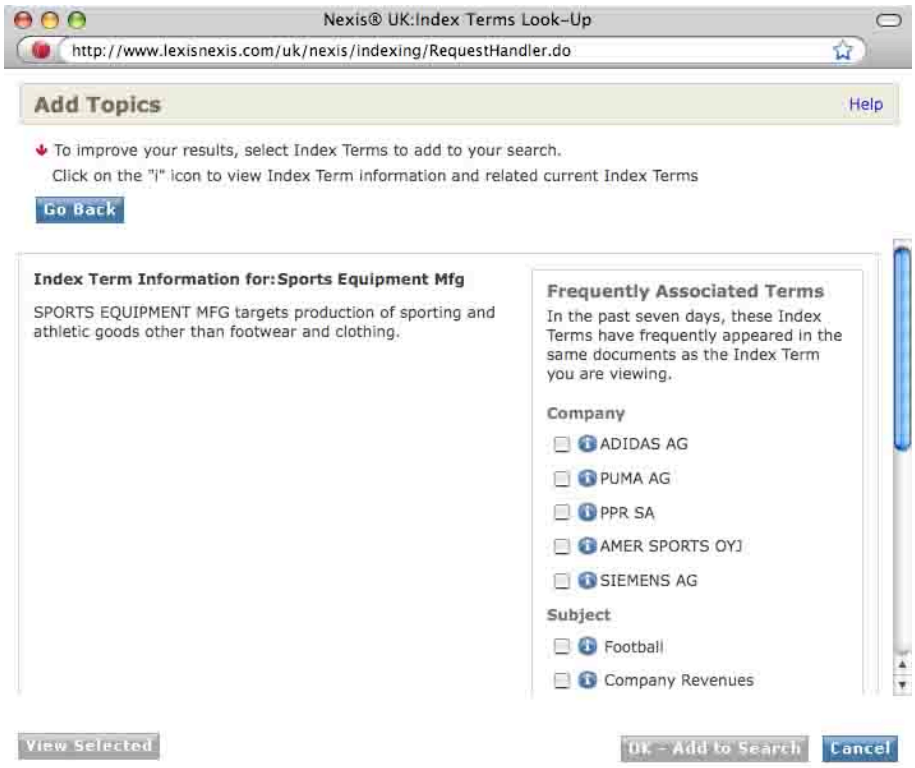

Abbildung 23: Schlagwortdetails bei *LexisNexis*

Die in Abbildung 23 dargestellte Detailansicht zum Schlagwort „Sports Equipment Mfg“ („Herstellung von Sportausrüstung“) liefert zunächst den wichtigen Hinweis, dass die Begriffe „clothing“ und „footwear“ („Sportkleindung“ und „Sportschuhe“) nicht zum Begriff der „Sportausrüstung“ gezählt werden. Hier wird eine Schwäche der Begriffsordnung offenbar, da „Sportausrüstung“, „Sportkleidung“ und „Sportschuhe“ in exakt dieser Reihenfolge durchaus eine schlüssige und transitive Hierarchie bilden. Neben den paradigmatischen, hierarchischen Begriffsordnungen, ermöglicht die Detailansicht außerdem die Nutzung einer weiteren, syntagmatischen Form der Relation, die „häufig assoziierten Schlagwörter“ (vgl. Abbildung 23). Darunter werden solche Schlagwörter verstanden, die häufig gemeinsam mit dem ausgewählten Schlagwort für ein Dokument vergeben wurden, wobei der Zeitraum auf die letzten sieben Tage vor dem Zeitpunkt der Suche beschränkt bleibt. Diese Form der Assoziationsrelation basiert auf statistischen Häufigkeitsmerkmalen, die, in Verbindung mit der zeitlichen Einschränkung, aktuelle Themenzusammenhänge reflektieren sollen.

Nach Auswahl der Schlagwörter, die den Informationsbedarf abdecken, kann schließlich die initiale Anfrage durchgeführt werden, die zu einer ersten

Ergebnismenge führt. Zur Anfragemodifikation bietet *LexisNexis* die Möglichkeit eines assoziativen Retrievals. Dabei sollen ähnliche Dokumente zu einem Musterdokument, auf der Grundlage gemeinsamer Indexterme gefunden werden.

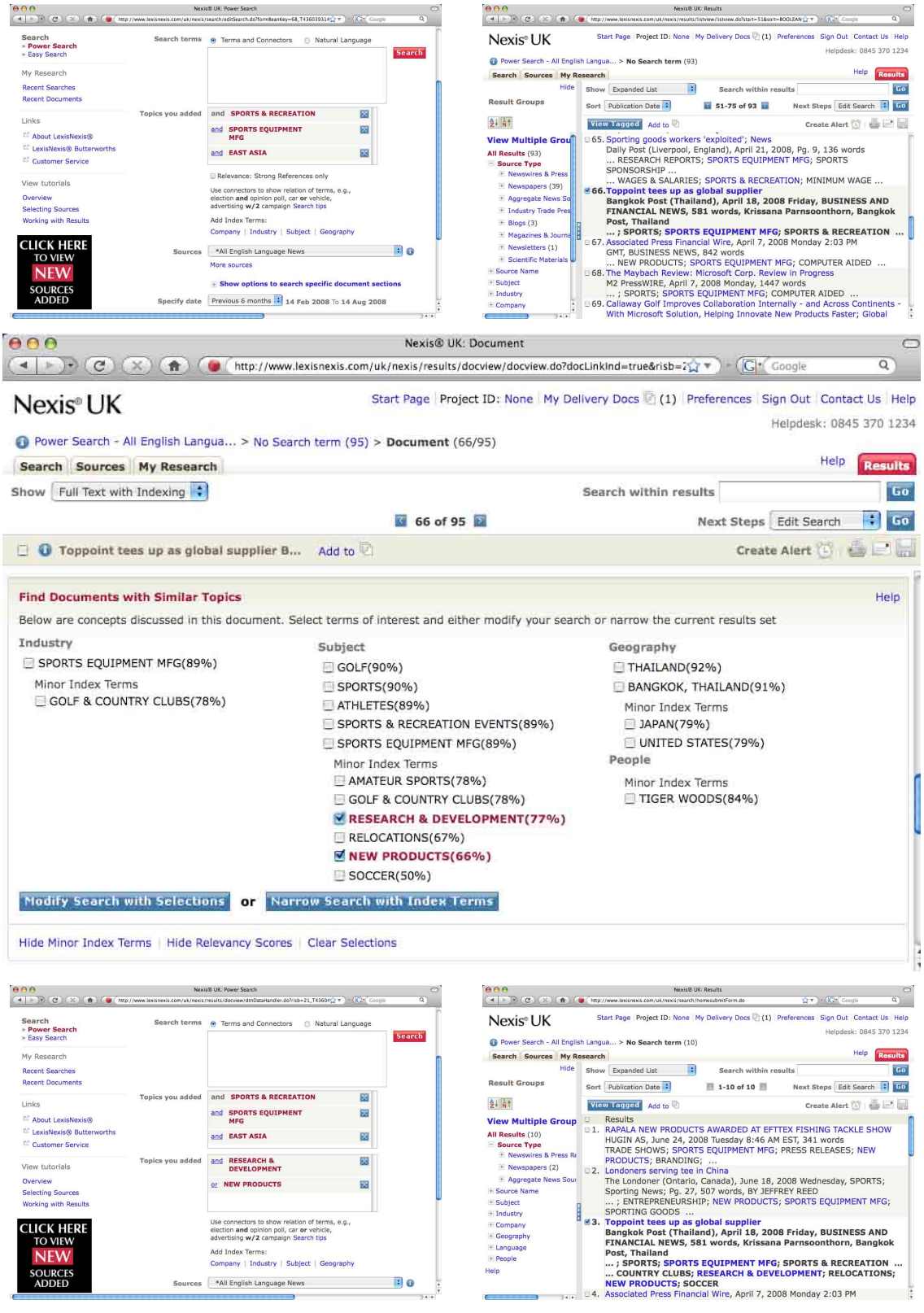

Abbildung 24: Retrieval durch Musterdokument bei *LexisNexis*

Abbildung 24 zeigt den Ablauf eines assoziativen Retrievals. Durch die oben links abgebildete initiale Anfrage ergibt sich die Ergebnismenge rechts daneben. In der Ergebnismenge wurde der Treffer 66 von insgesamt 93 als Musterdokument definiert und im Anzeigeformat „Fulltext with Indexing“ (im deutschsprachigen *LexisNexis* „Volltext / Schlagwortansicht“) aufgerufen. Im Anschluss an den Volltext des Dokuments erscheint das in Abbildung 24 in der Mitte dargestellte Feld „Find Documents with Similar Topics“. In diesem Feld werden alle von „SmartIndexing“ für das Dokument vergebene Schlagwörter aufgeführt.

Die prozentualen Relevanzangaben hinter den Schlagwörtern dienen als Anhaltspunkt, wie präzise das jeweilige Schlagwort den Dokumentinhalt repräsentiert. Es handelt sich dabei um einen dokumentbezogenen, relativen Wert, der sich vor allem aus der Häufigkeit der Terme im Dokument (gemessen wird dabei die Häufigkeit aller Terme, die durch ein einzelnes Schlagwort repräsentiert werden) und deren Position errechnet. Als Schwellenwert für „besonders relevante“ Schlagwörter wurde ein Wert von 85 Prozent festgelegt[203]. Durch die explizite Kennzeichnung der „Minor Index Terms“ wird systemseitig bereits angedeutet, dass diese weniger geeignet sind, um möglichst ähnliche Dokumente zum Musterdokument zu finden. Stock und Stock bestätigen dies und heben hervor, dass ein Retrieval nach ähnlichen Dokumenten mit Schlagwörtern von weniger als 85 Prozent Relevanz, zu einer starken Verschlechterung der Precision führt[204].

Die ausgewählten Schlagwörter werden durch eine UND-Verknüpfung an die initiale Anfrage angehängt. Es besteht nun die Möglichkeit per Klick auf die Schaltfläche „Modify Search with Selections“ zum Anfragedialog zurückzukehren und die modifizierte Anfrage über sämtliche ausgewählten Quellen anzuwenden (wie in Abbildung 24 dargestellt), oder durch die Funktion „Narrow Search with Index Terms“ die modifizierte Anfrage nur auf die vorliegende Ergebnismenge anzuwenden.

---

[203] Vgl. Gesprächsprotokoll: Telefonat LexisNexis Deutschland GmbH

[204] Stock; Stock: Digitale Rechts- und Wirtschaftsinformationen bei LexisNexis, Abs. 73

6.5 *CiteSeer*

Die bibliografische Datenbank *CiteSeer* wurde 1997 bei einem amerikanischen Forschungsinstitut der Fa. *NEC* entwickelt, seit 2003 ist das College of Information Sciences and Technology der Pennsylvania State University für *CiteSeer* verantwortlich[205].

CiteSeer beschränkt sich auf den Nachweis wissenschaftlicher Literatur aus den Bereichen Informatik und Informationswissenschaft, deren formale Zitate für ein assoziatives Retrieval ausgenutzt werden. Ausgehend von der Idealvorstellung eines universellen, interdisziplinären Zitationsindex, wurde *CiteSeer* als Alternative zu den kommerziellen Angeboten wie dem *Science Citation Index* entwickelt. Dessen Schwäche, so die *CiteSeer*-Entwickler, bestünde vor allem in der Gefahr von Verzerrungen, da die im *Science Citation Index* berücksichtigten Publikationen allein der Auswahl des Institute for Scientific Information (ISI) unterliegen, das den *Science Citation Index* erstellt[206].

CiteSeer berücksichtigt ausschließlich Dokumente aus dem World Wide Web und bietet zudem – wo dies möglich ist – deren Volltexte kostenlos als PDF-Dateien an. Davon müssen solche Dokumente ausgenommen bleiben, deren Rechteinhaber ihre Verwertungsinteressen durch die Zugänglichmachung der Volltexte gefährdet sehen. Da die Verlage ihre Publikationen meist in geschützten Bereichen, dem sogenannten „Deep Web“ vorhalten, ist *CiteSeer* auf individuelle Vereinbarungen mit den Verlagen angewiesen, um wenigstens für eine Indexierung und Zitationsanalyse der Dokumente in diese geschützten Bereiche vorstoßen zu können. Der Gefahr der Verzerrung durch Unvollständigkeit ist *CiteSeer* damit grundsätzlich ebenso ausgeliefert.

Die Arbeit von *CiteSeer* ist vollständig automatisiert. Aus einem Zusammenspiel verschiedener Algorithmen ergibt sich ein „ACI“ (Automatic Citation Indexing)-System. Das ACI-System identifiziert wissenschaftliche Literatur im World Wide

---

[205] Vgl. Homepage von CiteSeer. Online: http://citeseerx.ist.psu.edu/about/site [Abrufdatum: 21.08.2008]

[206] Giles, C. Lee; Bollacker, Kurt D.; Lawrence, Steve: CiteSeer: an automatic citation indexing system. In: Witten, Ian H. (Hrsg.): Digital Libraries 98. The Third ACM Conference on Digital Libraries, 1998, S. 89-90

Web, extrahiert und analysiert die darin enthaltenen Referenzen, sowie deren Kontext und führt schließlich solche Referenzen zusammen, die sich auf die gleichen Dokumente beziehen[207].

Für die Gewinnung neuer Dokumente wird durch Webcrawler im World Wide Web gezielt nach Postscript- und PDF-Dateien mit den gewünschten Inhalten (Themen zur Informatik und Informationswissenschaft) gesucht und diese auf bibliografische Inhalte wie Literaturverzeichnisse hin geprüft. Werden formale Zitate erkannt, geht *CiteSeer* davon aus, dass es sich um relevante Dokumente handelt. Die Dateien werden auf einen *CiteSeer*-Server heruntergeladen und in Text konvertiert. Umfangreiche Informationen werden aus dem Text extrahiert und einer Indexierung zugeführt. Dazu zählen: URL und Header der Datei (aufgrund der darin vermuteten Metadaten wie Titel und Verfasser des Dokuments), gegebenenfalls ein Abstract, die enthaltenen Referenzen, der Kontext der Referenzen, sowie der Volltext des Dokuments[208].

Fehleranfällig ist diese automatisierte Vorgehensweise vor allem durch die uneinheitliche Strukturierung und Formatierung der Dokumente, die es schwierig macht, die notwendigen Arbeitsschritte in allgemein anwendbaren Algorithmen zu formulieren. Auch das formale Zitieren in wissenschaftlicher Literatur kann durchaus verschiedenen Mustern folgen, daher werden alle Referenzen zunächst einer Normalisierung unterzogen, in der Bindestriche und Fußnoten- oder Endnotenzeichen am Anfang entfernt, Abkürzungen durch ausgeschriebene Wörter ersetzt und eine Konvertierung in vollständige Kleinschreibung durchgeführt werden[209].

Die Identifizierung und Zusammenführung verschiedener Referenzen, die sich auf das jeweils selbe Dokument beziehen, erfolgt durch den Vergleich übereinstimmender Terme, entweder auf Zeichen-, Wort- oder Phrasenebene. Dazu wurden verschiedene algorithmische Lösungen erprobt, die auf der Grundlage terminologischer Übereinstimmung mit Schwellenwerten oder der

---

[207] Vgl. ebd., S. 90
[208] Vgl. ebd., S. 90-91
[209] Vgl. ebd., S. 91-92

Berechnung von Ähnlichkeitsmaßen arbeiten[210]. Die exakte Funktion des derzeit eingesetzten Verfahrens ist allerdings unbekannt.

Anhand der zusammengestellten Informationen lässt sich ein Zitationsindex erstellen, der als Basis für die Bestimmung von Dokumentähnlichkeiten nach den Prinzipien der bibliografischen Kopplung und der Kozitation herangezogen wird.

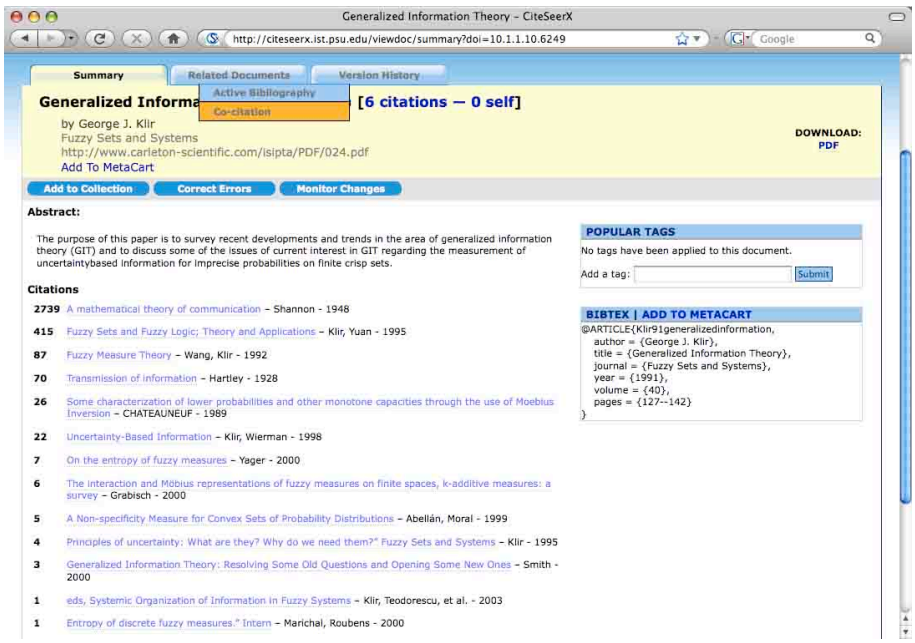

Abbildung 25: Assoziatives Retrieval bei *CiteSeer*
Quelle: N.N. 6c

Abbildung 25 zeigt im oberen, farbig hinterlegten Bereich eine Titelaufnahme zu einem Dokument, im unteren Bereich sind ein Abstract und die im Dokument enthaltenen Referenzen angeordnet. Die Anzahl, mit der eine Referenz in der *CiteSeer*-Kollektion insgesamt enthalten ist, ist jeweils vorangestellt, wobei die Anordnung nach absteigender Häufigkeit erfolgt. Das Dokument „Generalized Information Theory" selbst wird sechsmal als Referenz in der CiteSeer-Kollektion aufgeführt, darunter keine Selbstreferenzen[211], worauf der Zusatz „[6 citations – 0 self]" hinter dem Titel hinweist.

Ausgehend von „Generalized Information Theory" als Musterdokument kann nun ein assoziatives Retrieval per Klick auf „Related Documents" in Gang

---

[210] Vgl. ebd., S. 92

[211] Man spricht von einer Selbstreferenz, wenn der Verfasser auf seine eigenen, früheren Werke verweist. Vgl. Umstätter: Szientometrische Verfahren, S. 242

gebracht werden. Dem Nutzer werden zwei Optionen präsentiert: „Active Bibliography“ und „Co-Citation“.

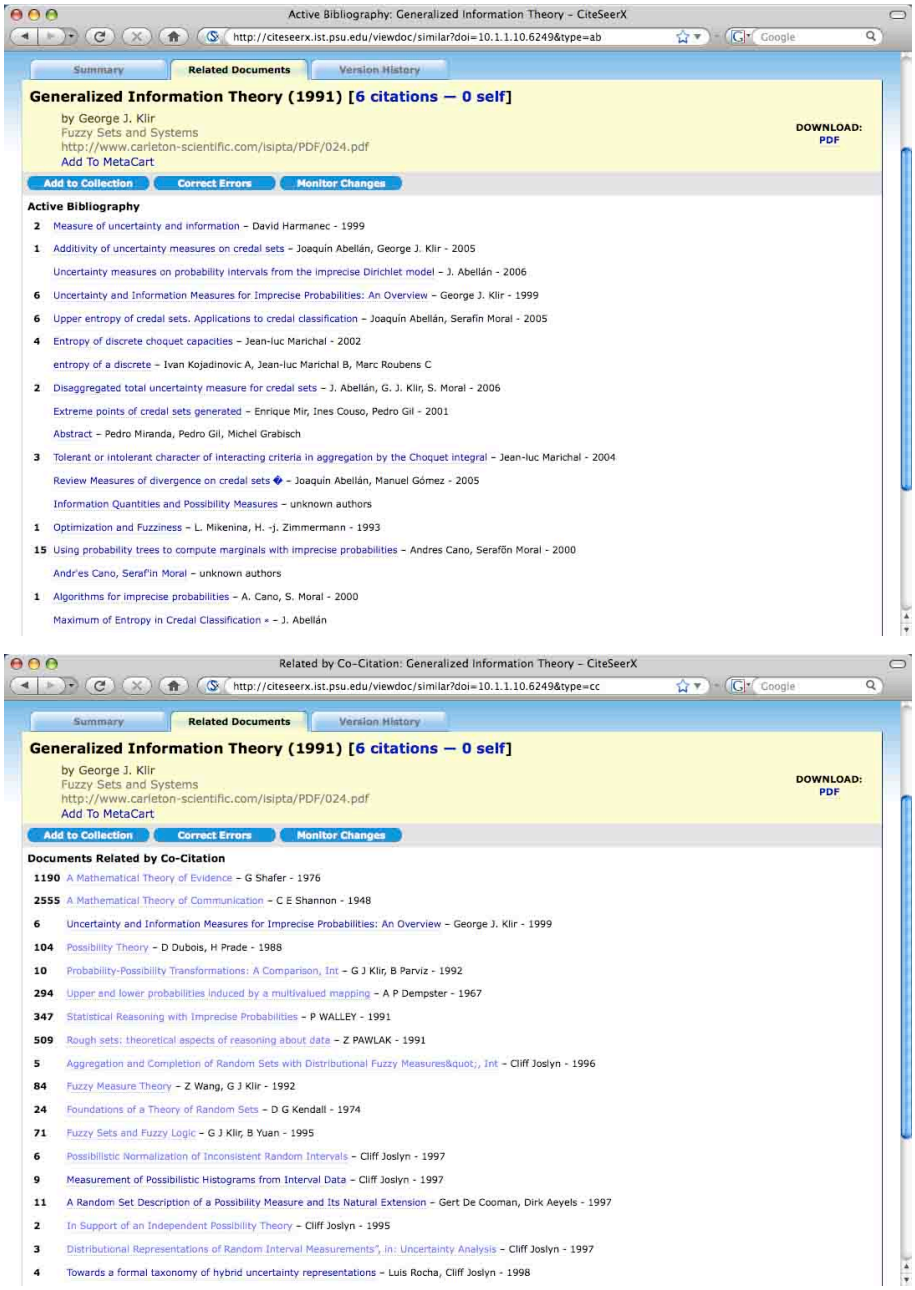

Abbildung 26: Bibliografische Kopplung und Kozitation bei *CiteSeer*
Quelle: N.N. 6d und N.N. 6e

Je nachdem wie erfolgreich das ACI-System Referenzen identifiziert und zusammenführt, können zu einem Musterdokument zahlreiche bibliografisch gekoppelte Dokumente gefunden werden (vgl. Abbildung 26 oben). Handelt es sich bei dem Musterdokument außerdem um einen zentralen, häufig zitierten Text, lassen sich ebenfalls zahlreiche, durch Kozitation assoziierte Dokumente finden (vgl. Abbildung 26 unten). Das Ranking der assoziierten Dokumente

erfolgt nach absteigender Intensität der bibliografischen Kopplung, bzw. der Kozitation, entsprechend den in Abschnitt 5.5.3 beschriebenen Prinzipien.

Der in Abschnitt 5.5.3.2 erfolgte Hinweis auf die unterschiedlichen Muster, die sich durch beide Beziehungsarten ergeben, scheint sich für das in Abbildung 26 dargestellte Beispiel zu bestätigen. Die durch bibliografische Kopplung gewonnen Dokumente sind (von Ausnahmen abgesehen) aktuelleren Datums als die Dokumente, die durch Kozitation gefunden wurden. Die Kozitation führte außerdem zu Dokumenten mit einer wesentlich höheren Zitierhäufigkeit. Das Phänomen lässt sich durch das Prinzip der Kozitation erklären: da nach Dokumenten gesucht wird, die mit dem Musterdokument gemeinsam zitiert werden, besteht eine höhere Wahrscheinlichkeit, Dokumente mit einer insgesamt hohen Zitierhäufigkeit zu finden.

Der große Vorteil der Sprachunabhängigkeit der zitationsanalytischen Verfahren bleibt aufgrund des rein auf das Englische ausgerichteten ACI-Systems ungenutzt. Zwar finden sich eine gewisse Anzahl nicht-englischsprachiger Dokumente in der *CiteSeer*-Kollektion, allerdings sind weder die Indexierung noch das Retrieval auf solche Dokumente eingestellt, bereits die Verarbeitung von Diakritika ist nicht möglich.

7. Zusammenfassung und Ausblick

Zu Beginn dieser Arbeit wurde die Anfragemodifikation als wesentlicher Aspekt für den freien Informationsfluss zwischen Retrievalsystem und Nutzer benannt. Darauf aufbauend wurden Suchanfragen und Anfragemodifikationen in ihrem Verhältnis zueinander betrachtet und eine Charakterisierung des Fakten- und des Information Retrievals gegeben. Die Verfahren zur Anfragemodifikation wurden, dem Schema der intellektuellen, automatischen und interaktiven Verfahrensklassen folgend, dargestellt und durch Anwendungsbeispiele illustriert.

Um eine Annäherung an den Themenkomplex aus der Perspektive des Nutzers vorzunehmen, muss vor allem dessen Suchverhalten beachtet werden. Die Information Retrieval-Forschung war in der Vergangenheit jedoch wesentlich stärker an rein technischen Fragestellungen interessiert, so dass die Entwicklung automatischer und interaktiver Modifikationsverfahren häufig gleichgesetzt wurde mit der Entwicklung neuer Ranking- und Clustering-Algorithmen oder Term- und Dokumentgewichtungsverfahren. Zweifellos ist die technische Fortentwicklung durch die Informatik ein zentraler Bestandteil in der Entwicklung des Information Retrieval. Dabei jedoch an der Vorstellung eines Suchverhaltens festzuhalten, das sich als linearer Anfragedialog beschreiben lässt, ist kein taugliches Konzept für die Gestaltung von Such- und Modifikationsprozessen.

Bates hebt in ihren Arbeiten aus den 1980er und frühen 1990er Jahren die Notwendigkeit der Berücksichtigung des Nutzerverhaltens bei der Entwicklung neuer Retrievalsysteme hervor[212]. In einer aktuelleren Arbeit nimmt Ruthven einen ähnlichen Standpunkt ein, indem er das interaktive Information Retrieval ausdrücklich unterteilt in zwei Aufgabenfelder, nämlich einerseits die Erforschung der Informationssuche und des Suchverhaltens und andererseits die Forschungs- und Entwicklungsarbeit an neuen Methoden der Interaktion, wozu letztlich auch die technische Realisierung von konkreten Retrievalsystemen und damit auch von Verfahren zur Anfragemodifikation

[212] Bates: The design of browsing and berrypicking techniques for the online search interface, S. 407-424 und Bates: Where should the person stop and the information search interface start?, S. 575-591

gezählt werden können[213]. In dem im vorangegangenen Abschnitt vorgestellten, experimentellen System *Daffodil* liegt ein Arbeitsergebnis vor, das aus einem solchen, interdisziplinären Verständnis heraus entwickelt wurde.

Die Entwicklung des World Wide Web und vor allem das Webretrieval, das in einer sehr viel dynamischeren Umgebung mit einem sehr heterogenen Nutzerkreis unter meist kommerziell geprägten Rahmenbedingungen stattfindet, nimmt ebenfalls Einfluss auf das klassische Information Retrieval. Durch die Nutzung von Websuchmaschinen haben sich die Nutzererwartungen und -vorstellungen wiederum verändert, der Trend weist auf sehr kurze Suchanfragen, Ignorierung erweiterter Suchmodi zur Formulierung komplexerer Anfragen und oberflächliche Auswertung der Ergebnismengen hin[214]. Solches Suchverhalten kann kurzfristig kaum zu optimalen Ergebnissen führen, stattdessen dürfte sich die Wahrnehmung der Informationssuche als iterativer Prozess festigen. Verfahren zur Anfragemodifikation sollten damit auch in einer, vermutlich stärker durch das Webretrieval beeinflussten Zukunft unentbehrlich sein, allein um die Schwächen menschlichen Suchverhaltens auszugleichen.

Damit die Modifikationsverfahren als Hilfsangebote wahrgenommen und genutzt werden, müssen sie in Abstimmung auf die Nutzererwartungen und das Suchverhalten kontinuierlich weiterentwickelt werden. Die Fülle der in Abschnitt 5 dargestellten, methodischen Ansätze des Information Retrieval bietet dazu eine Grundlage, deren Weiterentwicklung in alle Richtungen vorangetrieben werden sollte. Dazu zählt neben der Informatik, die bei der Realisierung konkreter Systeme gefordert ist, auch die Computerlinguistik im Bereich der morphologischen und syntaktischen Analysen, die empirische Informationswissenschaft im Bereich der Informetrie und schließlich das Bibliotheks- und Dokumentationswesen bei der Erstellung und Pflege von Normdateien und semantisch relationierten Dokumentationssprachen.

---

[213] Ruthven: Interactive information retrieval, S. 44
[214] Vgl. ebd., S. 49-50

8. Literaturverzeichnis

8.1 Literatur

Bates, Marcia J.: Information search tactics. In: Journal of the American Society for Information Science 30(1979)4, S. 205-214

Bates, Marcia J.: Search techniques. In: William, Martha E. (Hrsg.): Annual Review of Information Science and Technology 15(1981), S. 139-169

Bates, Marcia J.: The design of browsing and berrypicking techniques for the online search interface. In: Online Review 13(1989)5, S. 407-424

Bates, Marcia J.: Where should the person stop and the information search interface start? In: Information Processing & Management 26(1990)5, S. 575-591

Bayer, Oliver; Höhfeld, Stefanie; Josbächer, Frauke [u.a.]: Evaluation of an ontology-based knowledge-management-system. A case study of Convera RetrievalWare 8.0. In: Information Services & Use 25(2005)3/4, S. 181-195. Online: http://www.phil-fak.uni-duesseldorf.de/infowiss/admin/public_dateien/files/30/1138616739isu471_pdf.pdf [Abrufdatum: 23.07.2008, Datei: 1138616739isu471_pdf.pdf]

Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier: Modern Information Retrieval. Harlow [u.a.] : Addison-Wesley, 1999.

Bertram, Jutta: Einführung in die inhaltliche Erschließung. Grundlagen – Methoden – Instrumente. Würzburg : Ergon-Verlag, 2005.

Dumais, Susan: Latent semantic analysis. In: Cronin, Blaise (Hrsg.): Annual Review of Information Science and Technology 38(2004), S. 189-230

Efthimiades, Efthimis: Query expansion. In: William, Martha E. (Hrsg.): Annual Review of Information Science and Technology 31(1996), S. 121-187

Engelbert, Heinz: Der Informationsbedarf in der Wissenschaft. Leipzig : Bibliographisches Institut, 1976.

Feldman, Susan: Find what I mean, not what I say. Meaning-based search tools. In: Online 24(2000)3, S. 49-56

Ferber, Reginald: Information Retrieval. Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web. 1. Aufl. Heidelberg : dpunkt-Verlag, 2003.

Frants, Valery; Brush, Craig: The need for information and some aspects of information retrieval system construction. In: Journal of the American Society for Information Science 39(1988)2, S. 86-91

Frants, Valery; Shapiro, Jacob; Voiskunskii, Vladimir: Automated Information Retrieval. Theory and Methods. San Diego, Calif. [u.a.] : Academic Press, 1997.

Fuhr, Norbert: Information Retrieval. Skriptum zur Vorlesung im SS 06. Duisburg : Universität Duisburg-Essen, Campus Duisburg, Fachbereich Ingenieurswissenschaften, Abteilung Informatik und angewandte Kognitionswissenschaft, Fachgebiet Informationssysteme. Online: http://www.is.informatik.uni-duisburg.de/courses/ir_ss06/folien/irskall.pdf [Abrufdatum: 23.07.2008, Datei: irskall.pdf]

Fuhr, Norbert: Theorie des Information Retrieval I: Modelle. In: Kuhlen, Rainer; Seeger, Thomas; Strauch Dietmar (Hrsg.): Grundlagen der praktischen Information und Dokumentation. Band 1: Einführung in die Informationswissenschaft und -praxis. 5., völlig neu gefasste Ausg. München : Saur, 2004, S. 207-214

Garfield, Eugene: Announcing the SCI compact disc edition: CD-ROM gigabyte storage technology, novel software, and bibliographic coupling make desktop research and discovery a reality. In: Current Comments 11(1988)22, S. 160-166. Online: http://www.garfield.library.upenn.edu/essays/v11p160y1988.pdf [Abrufdatum: 24.07.2008, Datei: v11p160y1988.pdf]

Giles, C. Lee; Bollacker, Kurt D.; Lawrence, Steve: CiteSeer: an automatic citation indexing system. In: Witten, Ian H. (Hrsg.): Digital Libraries 98. The Third ACM Conference on Digital Libraries. New York : ACM, 1998,

S. 89-98. Online: http://clgiles.ist.psu.edu/papers/DL-1998-citeseer.pdf [Abrufdatum: 21.08.2008, Datei: DL-1998-citeseer.pdf]

Gödert, Winfried; Lepsky, Klaus: Semantische Umfeldsuche im Information Retrieval in Online-Katalogen. Köln : Fachhochschule Köln, Fachbereich Bibliotheks- und Informationswesen, 1997. (Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft ; 7). Online: http://www.fbi.fh-koeln.de/institut/papers/kabi/volltexte/band007.pdf [Abrufdatum: 23.07.2008, Datei: band007.pdf]

Gödert, Winfried: Navigation und Konzepte für ein interaktives Retrieval im OPAC. Oder: von der Informationserschliessung zur Wissenserkundung. In: AKMB-news 10(2004)1, S. 27-30. Online: http://archiv.ub.uni-heidelberg.de/artdok/volltexte/2007/312/pdf/2004_Goedert.pdf [Abrufdatum: 23.07.2008, Datei: 2004_Goedert.pdf]

Griffiths, Alan; Luckhurst, Claire H.; Willett, Peter: Using interdocument similarity information in document retrieval systems. In: Journal of the American Society for Information Science 37(1986)1, S. 3-11

Harman, Donna: Relevance feedback and other query modification techniques. In: Frakes, William; Baeza-Yates, Ricardo (Hrsg.): Information Retrieval. Data Structures & Algorithms. Upper Saddle River, NJ : Prentice Hall, 1992, S. 241-263

Harman, Donna: Relevance feedback revisited. In: Belkin, Nicolas; Ingwersen, Peter; Pejtersen, Annelise M. (Hrsg.): Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York : ACM, 1992, S. 1-10

Harter, Stephen: Online Information Retrieval. Concepts, Principles, and Techniques. Orlando [u.a.] : Academic Press, 1986.

Kessler, Meyer M.: Bibliographic coupling between scientific papers. In: American Documentation 14(1963)1, S. 10-25

Klas, Claus-Peter; Kriewel, Sascha; Fuhr, Norbert [u.a.]: Daffodil – Nutzerorientiertes Zugangssystem für heterogene Digitale Bibliotheken.

In: Ockenfeld, Marlies (Hrsg.): Leitbild Informationskompetenz. Positionen – Praxis – Perspektiven im europäischen Wissensmarkt. 27. Online-Tagung der DGI, 57. Jahrestagung der DGI. Frankfurt am Main : DGI, 2005, S. 163-175

Klas, Claus-Peter: Daffodil. Strategische Unterstützung bei der Informationssuche in Digitalen Bibliotheken. Duisburg : Universität Duisburg-Essen, Fachbereich Ingenieurswissenschaften, Dissertation, 2007. Online: http://www.is.informatik.uni-duisburg.de/bib/pdf/ir/Klas:07b.pdf [Abrufdatum: 18.08.2008, Datei: Klas:07b.pdf]

Kolke, Ernst-Gerd vom: Online-Datenbanken. Systematische Einführung in die Nutzung elektronischer Fachinformation. 2. Aufl. München [u.a.] : Oldenbourg, 1996.

Kowalski, Gerald J.; Maybury, Mark T.: Information Storage and Retrieval Systems. Theory and Implementation. 2nd ed. Boston, Mass. [u.a.] : Kluwer Academic Publ., 2000.

Kriewel, Sascha; Fuhr, Norbert: Adaptive search suggestions for digital libraries. In: Goh, Dion Hoe-Lian (Hrsg.): Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers, ICADL 2007. Berlin [u.a.] : Springer, 2007, S. 220-229. Online: http://www.is.informatik.uni-duisburg.de/bib/pdf/ir/Kriewel_Fuhr:07.pdf [Abrufdatum: 24.07.2008, Datei: Kriewel_Fuhr:07.pdf]

Kuhlen, Rainer: Hypertext. Ein nicht-lineares Medium zwischen Buch und Wissensbank. Berlin [u.a.] : Springer, 1991.

Kürsten, Jens: Systematisierung und Evaluierung von Clustering-Verfahren im Information Retrieval. Chemnitz : Technische Universität Chemnitz, Fakultät für Informatik, Professur Medieninformatik, Diplomarbeit, 2006. Online: http://archiv.tu-chemnitz.de/pub/2006/0203/data/Clustering_Verfahren_im_Information_Retrieval.pdf [Abrufdatum: 24.07.2008, Datei: Clustering_Verfahren_im_Information_Retrieval.pdf]

Lancaster, Frederick W.: Information Retrieval Systems. Characteristics, Testing and Evaluation. 2nd ed. Chichester [u.a.] : Wiley, 1979.

Lepsky, Klaus; Vorhauer, John: Lingo – ein open source System für die Automatische Indexierung des Deutschen. In: ABI Technik 26(2006)1, S. 18-29. Online: http://www.iws.fh-koeln.de/institut/personen/lepsky/lingo-uebersichtsartikel.pdf [Abrufdatum: 24.07.2008, Datei: lingo-uebersichtsartikel.pdf]

Lewandowski, Dirk: Web Information Retrieval. Technologien zur Informationssuche im Internet. Frankfurt am Main : DGI, 2005. (DGI-Schrift Informationswissenschaft ; 7). Online: http://www.durchdenken.de/lewandowski/web-ir/download/Web-IR-Buch.pdf [Abrufdatum: 24.07.2008, Datei: Web-IR-Buch.pdf]

Lezius, Wolfgang; Rapp, Reinhard; Wettler, Manfred: A freely available morphological analyzer, disambiguator and context sensitive lemmatizer for german. In: Proceedings of the 17th International Conference on Computational Linguistics – Volume 2. Montreal [u.a.], Canada : Government of Canada, Université de Montréal, 1998, S. 743-748. Online: http://www.wolfganglezius.de/lib/exe/fetch.php?id=public%3Acl%3Amorphy&cache=cache&media=public:cl:coling.pdf [Abrufdatum: 23.07.2008, Datei: coling.pdf]

Mandala, Rila; Tokunaga; Takenobu; Hozumi, Tanaka: The use of WordNet in information retrieval. In: Harabagiu, Sanda (Hrsg.): COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, 1998, S. 31-37. Online: http://tanaka-www.cs.titech.ac.jp/publication/archive/168.pdf [Abrufdatum: 24.07.2008, Datei: 168.pdf]

Miller, George; Beckwith, Richard; Fellbaum, Christiane [u.a.]: Introduction to WordNet: an on-line lexical database. In: International Journal of Lexicography 3(1990)4, S. 235-244

N.N.: Datenbasis. In: Grundlagen der praktischen Information und Dokumentation. Band 2: Glossar. 5., völlig neu gefasste Ausg. München : Saur, 2004, S. 22

Nohr, Holger: Grundlagen der automatischen Indexierung. Ein Lehrbuch. 3., überarb. Aufl. Berlin : Logos-Verlag, 2005.

Nohr, Holger: Theorie des Information Retrieval II: Automatische Indexierung. In: Kuhlen, Rainer; Seeger, Thomas; Strauch, Dietmar (Hrsg.): Grundlagen der praktischen Information und Dokumentation. Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis. 5., völlig neu gefasste Ausg. München : Saur, 2004, S. 215-225

Plassmann, Engelbert; Rösch, Hermann; Seefeldt, Jürgen [u.a.]: Bibliotheken und Informationsgesellschaft in Deutschland. Eine Einführung. Wiesbaden : Harrassowitz, 2006.

Rijsbergen, Cornelis J. van: Information Retrieval. 2nd ed. London : Butterworths, 1979. Online: http://www.dcs.gla.ac.uk/Keith/Preface.html [Abrufdatum: 24.07.2008, Datei: Preface.html]

Robertson, Stephen; Sparck-Jones, Karen: Relevance weighting of search terms. In: Journal of the American Society for Information Science 27(1976)3, S. 129-146

Robertson, Stephen: The probability ranking principle in IR. In: Sparck-Jones, Karen; Willett, Peter (Hrsg.): Readings in Information Retrieval. San Francisco, Calif. : Morgan Kaufmann, 1997, S. 281-286

Ruthven, Ian; Lalmas, Mounia: A survey on the use of relevance feedback for information access systems. In: The Knowledge Egineering Review 18(2003)2, S. 95-145. Online: http://inex.is.informatik.uni-duisburg.de:2004/pdf/ker_ruthven_lalmas.pdf [Abrufdatum: 24.07.2008, Datei: ker_ruthven_lalmas.pdf]

Ruthven, Ian: Interactive information retrieval. In: Cronin, Blaise (Hrsg.): Annual Review of Information Science and Technology 42(2008), S. 43-91

Salton, Gerard; Buckley, Chris: Improving retrieval performance by relevance feedback. In: Journal of the American Society for Information Science 41(1990)4, S. 288-297

Salton, Gerard; McGill, Micheal: Information Retrieval. Grundlegendes für Informationswissenschaftler. Hamburg [u.a.] : McGraw-Hill, 1987.

Schaefer, André; Jordan, Matthias; Klas, Claus-Peter [u.a.]: Active support for query formulation in virtual digital libraries: A case study with Daffodil. In: Rauber, Andreas; Christodoulakis, Stavros; Tjoa, A Min (Hrsg.): Research and Advanced Technologies for Digital Libraries. 9th European Conference, ECDL 2005. Berlin [u.a.] : Springer, 2008, S. 414-425. Online: http://www.is.informatik.uni-duisburg.de/bib/pdf/ir/Schaefer_etal:05.pdf [Abrufdatum: 24.07.2008, Datei: Schaefer_etal:05.pdf]

Schenkel, Ralf; Weikum, Gerhard: Vorlesung „Informationssysteme“. Sommersemester 2004. Saarbrücken : Max-Planck-Institut für Informatik, Department 5: Databases and Information Systems. Online: http://www.mpi-inf.mpg.de/departments/d5/teaching/ss04/is04/skripte/kap2.pdf [Abrufdatum: 01.09.2008, Datei: kap2.pdf]

Small, Henry: Co-citation in the scientific literature: a new measure of the relationship between two documents. In: Journal of the American Society for Information Science 24(1973)4, S. 265-269

Stock, Mechtild; Stock, Wolfgang G.: Digitale Rechts- und Wirtschaftsinformationen bei LexisNexis. In: JurPC Web-Dok. 82(2005), Abs. 1-105. Online: http://www.jurpc.de/aufsatz/20050082.htm [Abrufdatum: 05.08.2008, Datei: 20050082.htm]

Stock, Mechtild; Stock, Wolfgang G.: One-Stop-Shops internationaler Fachinformation. Wie gut sind Dialog / DataStar? In: Password 18(2003)4, S. 22-29. Online: http://wwwalt.phil-fak.uni-duesseldorf.de/infowiss/admin/public_dateien/files/1/1078481212password_4.pdf [Abrufdatum: 29.07.2008, Datei: 1078481212password_4.pdf]

Stock, Wolfgang G.; Stock, Mechtild: Wissensrepräsentation. Informationen auswerten und bereitstellen. München : Oldenbourg, 2008.

Stock, Wolfgang G.: Information Retrieval. Informationen suchen und finden. München : Oldenbourg, 2007.

Stock, Wolfgang G.: Natürlichsprachige Suche – More like this! In: Password 13(1998)11, S. 21-28. Online: http://www.phil-fak.uni-duesseldorf.de/infowiss/admin/public_dateien/files/1/1126182092114_pdf.pdf [Abrufdatum: 29.07.2008, Datei: 1126182092114_pdf.pdf]

Stock, Wolfgang G.: Textwortmethode. In: Password 15(2000)7/8, S. 26-35. Online: http://www.phil-fak.uni-duesseldorf.de/infowiss/admin/public_dateien/files/1/1078740450password_7.pdf [Abrufdatum: 29.07.2008, Datei: 1078740450password_7.pdf]

Trunk, Daniela: Semantische Netze in Informationssystemen. Verbesserung der Suche durch Interaktion und Visualisierung. Köln : Fachhochschule Köln, Fakultät für Informations- und Kommunikationswissenschaften. Institut für Informationswissenschaft, 2005. (Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft ; 51). Online: http://www.fbi.fh-koeln.de/institut/papers/kabi/volltexte/Band051.pdf [Abrufdatum: 24.07.2008, Datei: Band051.pdf]

Umstätter, Walther: Szientometrische Verfahren. In: Kuhlen, Rainer; Seeger, Thomas; Strauch, Dietmar (Hrsg.): Grundlagen der praktischen Information und Dokumentation. Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis. 5., völlig neu gefasste Ausg. München : Saur, 2004, S. 237-243

8.2 Weitere Quellen

Gesprächsprotokoll: Telefonat LexisNexis Deutschland GmbH. [Datum: 06.08.2008, Gesprächspartner: Tim Roschanski] Das Protokoll liegt beim Verfasser vor.

Homepage der Dialog Corporation. Online: http://www.thomsondialog.com [Abrufdatum: 29.07.2008]

Homepage der Fa. Textec. Online: http://www.textec.de [Stand: Januar 2008, Abrufdatum: 15.07.2008, Dateien: lexiquo.html, lexilib.html]

Hompage des Bundesumweltamtes. Online: http://www.umweltbundesamt.de/uba-info/dokufabib/thes.htm [Stand: Juni 2008, Abrufdatum: 24.07.2008, Datei: thes.htm]

Homepage von CiteSeer. Online: http://citeseerx.ist.psu.edu [Abrufdatum: 21.08.2008]

Homepage von Daffodil. Online: http://www.daffodil.de [Abrufdatum: 24.07.2008]

N.N. 5a: http://wordnet.princeton.edu/perl/webwn?o2=&o0=1&o7=&o5=&o1=1&o6=&o4=&o3=&s=right&i=27&h=00000000000000010001000010000000000000000000#c [Abrufdatum: 24.07.2008]

N.N. 6a: http://www.lexiquo.net/lexiquo_result.html?DEU&Keine&pro&Mi%DFtrauensvotum&google&screenlang=deu [Abrufdatum: 15.07.2008]

N.N. 6b: http://www.lexiquo.net/lexiquo_result.html?DEU&Keine&undefined&Schloss&google&screenlang=deu [Abrufdatum: 15.07.2008]

N.N. 6c: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.10.6249 [Abrufdatum: 23.08.2008]

N.N. 6d: http://citeseerx.ist.psu.edu/viewdoc/similar?doi=10.1.1.10.6249&type=ab [Abrufdatum: 24.08.2008]

N.N. 6e: http://citeseerx.ist.psu.edu/viewdoc/similar?doi=10.1.1.10.6249&type=cc [Abrufdatum: 24.08.2008]

Eidesstattliche Erklärung

Hiermit versichere ich, die Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt zu haben.

Leverkusen, den 22.09.2008 _____