Seminar für Computerlinguistik
Universität Zürich

Sommersemester 2001

Seminar „Semantikrepräsentation für Antwortextraktion“
Prof. Dr. M. Hess
lic. phil. S.Clematide

Information Retrieval – Eine Einführung in das Indexieren

X	D	A	N	A	L	Y	S	I	S	O	D	P	H	A	R	F
M	E	E	C	Y	C	N	E	U	Q	E	R	F	U	T	E	P
L	L	S	R	U	I	M	I	N	D	E	X	I	N	G	T	O
A	L	B	E	T	S	N	D	O	C	U	M	E	N	T	S	Y
N	O	I	T	A	C	I	F	I	S	S	A	L	C	V	U	R
G	R	Q	R	O	N	P	S	O	M	T	O	I	L	E	L	A
U	T	H	I	E	W	I	U	D	R	O	W	Y	E	K	C	L
A	N	K	E	B	O	O	L	E	T	M	B	Q	S	T	M	U
G	O	I	V	N	R	B	C	G	V	H	A	U	L	O	E	B
E	C	M	A	E	A	A	E	I	M	R	E	T	V	R	S	A
M	N	W	L	T	L	A	N	G	I	S	E	S	I	C	I	C
Z	U	S	A	L	I	N	T	U	A	X	D	S	A	O	O	O
U	B	D	Q	I	M	A	B	S	T	R	A	C	T	U	N	V
C	X	T	H	G	I	E	W	F	M	E	T	S	Y	S	R	E
H	C	R	A	E	S	T	O	P	W	O	R	D	A	L	G	I

abstract, analysis, boole, classification, cluster, data, documents, frequency, IDF, indexing, information, keyword, language, noise, precision, recall, retrieval, search, signal, similar, stopword, system, term, text, thesauri, tool, TREC, uncontrolled, vektor, vocabulary, weight

Alexandra Bünzli
Im Eichli 18
8162 Steinmaur

01 853 17 40
abuenzli@bluewin.ch

Abgabedatum: 20. September 2001

Lösung


X D A N A L Y S I S O D P H A R F
M E E C Y C N E U Q E R F U T E P
L L S R U I M I N D E X I N G T O
A L B E T S N D O C U M E N T S Y
N O I T A C I F I S S A L C V U R
G R Q R O N P S O M T O I L E L A
U T H I E W I U D R O W Y E K C L
A N K E B O O L E T M B Q S T M U
G O I V N R B C G V H A U L O E B
E C M A E A A E I M R E T V R S A
M N W L T L A N G I S E S I C I C
Z U S A L I N T U A X D S A O O O
U B D Q I M A B S T R A C T U N V
C X T H G I E W F M E T S Y S R E
H C R A E S T O P W O R D A L G I

Inhaltsverzeichnis

1 Einleitung

In der heutigen multimedialen Zeit werden wir jeden Tag mit einer riesigen Datenmenge konfrontiert. Die Informationsflut nimmt ständig zu. Noch vor ungefähr 200 Jahren wuchs das gesamte verfügbare Wissen relativ langsam. Ab 1800 verdoppelte sich die Zahl der wissenschaftlichen Arbeiten bereits alle 50 Jahre und heute hat sich das Wachstum beschleunigt. So stieg die Zahl der wissenschaftlichen Zeitschriften zwischen 1800 und 1966 von 100 auf 100'000 an, und die Grenzen des Wachstums sind bis heute nicht zu erkennen [Salton und McGill 1987, S. 2f.]. Auch in spezialisierten, thematisch abgegrenzten Gebieten besteht heutzutage eine Informationsüberlastung. Diese grosse Menge an Wissen ist verloren, wenn wir es nicht sammeln und für die Interessierten zugänglich machen. Damit man das Wissen nutzen kann, muss es so gespeichert werden, dass es bei Bedarf wieder gefunden werden kann. Dafür waren und sind hauptsächlich die Bibliotheken zuständig. Es ist daher nicht verwunderlich, dass der Ursprung des Information Retrieval im Bibliothekswesen liegt. Die vielen Bücher und Zeitschriftenartikel müssen aufgrund von bibliographischen Daten abgelegt werden. Jedes einzuordnende Objekt muss nach festgelegten Regeln inhaltlich erschlossen, mit entsprechenden Begriffen versehen, klassifiziert und in die bestehende Informationssammlung integriert werden. Als eine Folge davon gibt es auch feste Regeln, wie abgefragt werden muss und wie gesucht wird, das heisst, wie die Anfragen und Dokumente abzugleichen sind. Früher wurde alles von Hand ausgeführt, und im Laufe der Zeit erarbeitete man sich sehr gut funktionierende Verfahren. Durch den zunehmenden Einsatz von Computern entwickelten sich Methoden, die eine automatische Verarbeitung möglich machten. Solche Methoden bereitzustellen und zu verbessern ist die Aufgabe der Informatik. Dabei kann und soll sie auch auf die bestehenden Techniken der manuellen Methoden zurückgreifen, die sich jahrelang gut bewährt haben.

Ich werde in dieser Arbeit eine Einführung in die wichtigsten traditionellen Techniken der Indexierung geben. Das heisst, ich werde mich mit der Frage beschäfti-

gen, wie ein Information Retrieval System den Inhalt der einzuordnenden Objekte repräsentieren soll, damit bei der Suche nach Informationen optimale Ergebnisse erreicht werden. Zuerst stelle ich ganz kurz vor, was ein Information Retrieval System ist. Danach werde ich erläutern, was man unter Indexierung versteht und worauf man dabei achten muss. Es folgt ein Überblick über die traditionellen Methoden des Indexierens, wobei ich auch zwei bewährte Hilfsmittel, die Klassifikation und den Thesaurus, näher betrachten werde. Nur kurz möchte ich dann auf die manuelle Indexierung eingehen, der Schwerpunkt soll auf der maschinellen Verarbeitung liegen. So folgt eine Einführung in das automatische Indexieren, in der ich die einfachsten Methoden erklären werde. Nach einer kleinen Aufstellung einiger weiterführender Ansätze, möchte ich die Arbeit mit einem Blick in die Zukunft abschliessen, in der vor allem die Informationssuche im Internet eine grosse Rolle spielen wird.

2 Definition und Grundkonzepte von Information Retrieval Systemen

Leider gibt es für den Begriff des *Information Retrieval*[1] keine eindeutige Definition. Meist versucht man das Fachgebiet über die Eigenschaften des zu entwickelnden Systems zu definieren, indem man dessen Aufgaben beschreibt und es anderen Informationssystemen gegenüber abgrenzt[2]. Ganz allgemein könnte man sagen:

> Ein *Information Retrieval System* ist ein Informationssystem, das Dokumente[3] speichert, analysiert, wiederauffindbar macht, heraussucht und dem Benutzer verfügbar macht.

Ein solches System soll also eine Brücke bilden zwischen demjenigen, der eine Information kreiert, und demjenigen, der diese Information braucht. Idealerweise sollte eine Datensammlung immer aktuell und vollständig sein. Um die Aktualität zu erreichen, müssen ständig neue Dokumente hinzugefügt werden. Die Forderung nach Vollständigkeit hingegen verlangt, dass alle für die Sammlung wichtigen Dokumente aufgenommen werden. Die Informationseinheiten werden normalerweise datiert, um die Aktualität der Einträge festzuhalten. Dadurch verliert man die Information nicht, dass der Sachverhalt sich früher anders dargestellt hat. Wirklich veraltete Einheiten können jedoch entfernt und archiviert werden, um das System zu entlasten [Grossmann 1984b, S. 3]. Dass Aktualität und Vollständigkeit nicht völlig erreicht werden können, ist offensichtlich. Man versucht daher zwischen wichtigen und unwichtigen Informationseinheiten zu unterscheiden, was aber sehr schwierig

[1] Auch in der deutschen Literatur wird meistens die englische Bezeichnung verwendet. Eine deutsche Übersetzung wäre *Informationswiedergewinnungssystem*. In Zukunft werde ich Information Retrieval mit IR bzw. Information Retrieval System mit IRS abkürzen.

[2] So grenzt man es z. B. Datenbanksystemen gegenüber ab, die sich für die Speicherung von strukturierten Daten eignen, während sich IRS mit Daten beschäftigen, die sich nur schwer strukturieren lassen. Die Speicherung in fixen Datenfeldern wie bei den Datenbanksystemen ist daher nicht möglich.

[3] Ich werde im Rahmen dieser Arbeit den Begriff „Dokument“ für jegliche Art von Informationsquellen und -einheiten benutzen, also auch für Bilder, Tondokumente etc..

ist. Schlussendlich werden die Entscheidungen, welche Dokumente aufgenommen werden, ziemlich willkürlich gefällt [Salton und McGill 1987, S. 2].

Ein IRS muss fähig sein, mit verschiedensten Quellen von Information umzugehen. Es sollte einfach zu benutzen sein und gleichzeitig die Anforderungen unterschiedlicher Benutzer erfüllen können. Grosse Probleme bereiten vor allem das unsichere Wissen und die vagen Anfragen. Das unsichere Wissen ist ein Problem der begrenzten Repräsentation der Semantik der Dokumente. Textinhalte lassen sich nie perfekt durch einige wenige Begriffe erfassen. Das Problem wird insofern noch dadurch verschärft, dass im Prinzip die Darstellungsform des in einem IRS gespeicherten Wissens nicht beschränkt ist. Neben Texten können auch multimediale Objekte wie Bilder und Tondokumente, aber auch Regeln, Messwerte, Akten, semantische Netze etc. verarbeitet werden. Wie sollen solche Informationseinheiten inhaltlich erschlossen und repräsentiert werden? Mit dem gleichen Problem kämpft auch der Benutzer. Auch er muss sein Informationsbedürfnis auf eine angemessene Repräsentation abbilden. Ausserdem weiss er oftmals nicht genau, welche Datenelemente sein Informationsbedürfnis befriedigen und stellt vage Anfragen, bei denen die Antwort nicht a priori eindeutig definiert ist. Zusätzlich zu diesen Schwierigkeiten können natürlich auch die zu speichernden Daten selbst unsicher oder unvollständig sein [Fuhr 1998, S. 10]. Der Benutzer erhält keine Sicherheit, ob seine Antwort vollständig und korrekt ist.

Schlussendlich wird die Qualität eines IRS dadurch bestimmt, ob das System in angemessener Zeit alle relevanten[4], das heisst alle Dokumente, die das Informationsbedürfnis des Benutzers befriedigen, und *nur* die relevanten Informationseinheiten zu einer Anfrage findet [Grossmann 1984b, S. 10].

[4]Die Relevanz ist ein ziemlich abstraktes Mass, das subjektiv und schwer abzuschätzen ist.

Die Hauptaufgaben eines IR-Systems wurden von [Chowdhury 1999, S. 3] folgendermassen aufgelistet:

1. Identifizierung der Information (Quellen), die für die Zielgruppe relevant ist
2. Analyse des Inhalts dieser Dokumente
3. Erstellung einer Repräsentation des Inhalts dieser Dokumente, die geeignet ist, um sie mit den Anfragen der Benutzer abstimmen zu können
4. Analyse der Anfragen der Benutzer und Transformation der Anfragen in eine Form, die sich für die Abstimmung mit der erstellten Datenbank eignet
5. Suchen nach Übereinstimmung zwischen Anfrage-Repräsentation und den Datenbankeinträgen
6. Gewinnung der relevanten Information
7. Die Fähigkeit, nötige Anpassungen vorzunehmen, die auf dem Feedback der Benutzer basieren

IR-Systeme kann man in zwei Hauptkomponenten zerlegen. Abbildung 1 soll dies anschaulich machen. Die eine Komponente geht von den Informationsquellen (Information Sources) aus, analysiert sie und erstellt eine für die Suche geeignete Repräsentation dieser Dokumente (Analysis and Representation). Diese Repräsentationen wiederum werden abgespeichert und so organisiert, dass ein schneller Zugriff möglich ist (Organized Information). Das ist sozusagen die Vorarbeit, die geleistet wird. Die andere Komponente tritt erst bei einer konkreten Suchanfrage eines Benutzers (User) in Aktion. Diese Suchanfrage wird analysiert (Query Analysis) und ebenfalls in eine bestimmte Form gebracht (Analysed Queries), die für die Abgleichung mit den Repräsentationen der Dokumente geeignet ist. Nach der Abgleichung (Matching) werden diejenigen Dokumente ausgegeben (Retrieved Information), die das System aufgrund der Ähnlichkeit zwischen der Repräsentation der Dokumente

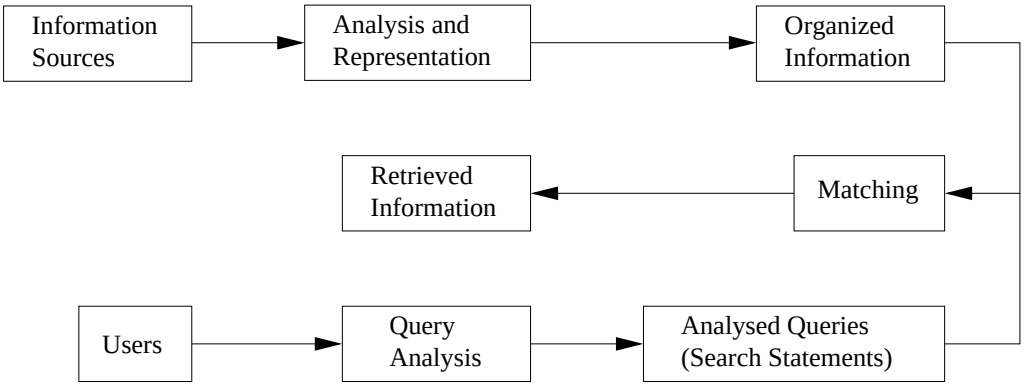

Abbildung 1: Allgemeines Modell eines IRS. Quelle: [Chowdhury 1999, S. 4]

und derjenigen der Suchanfrage gefunden hat. Ich werde mich im folgenden auf die Teilkomponente der Analyse und Repräsentation der Dokumente, die sogenannte Indexierung, beschränken und verschiedene Ansätze vorstellen.

3 Analyse und Repräsentation der Dokumente: Grundlagen der Indexierung

Ein grosses Forschungsfeld im IR befasst sich mit der geeigneten Repräsentation des Wissens, das in den Dokumenten enthalten ist. Es soll eine Form gefunden werden, die eine effiziente Suche erlaubt. Das Herstellen dieser Form nennt man *Indexieren*. Dabei müssen vor allem zwei Aufgaben erfüllt werden: Erstens soll eine solche Repräsentation verschiedene Formulierungen für das Gleiche auf eine einzige Repräsentation abbilden, damit bei einer Anfrage alle Dokumente zu diesem Thema gefunden werden, das heisst die *Abdeckung* (*Recall*)[5] erhöht wird. Zweitens müssen unklare Formulierungen (z.B. mehrdeutige Wörter) in eine eindeutige Form gebracht werden, damit die entsprechenden Dokumente nur bei einer Anfrage nach der zutreffenden Bedeutung geliefert werden. Dadurch wird die *Präzision* (*Precision*)[6] erhöht

[5] **Abdeckung** (im Deutschen wird ebenfalls *Ausbeute*, oft aber auch die englische Bezeichnung verwendet.): Ein Standardmass für die Evaluierung von IR-Systemen. Die Abdeckung bezeichnet den Anteil der gefundenen (GEF) relevanten (REL) Dokumente in Bezug auf alle in der Sammlung vorhandenen relevanten Dokumente (REL).

$$\text{Abdeckung: } r = \frac{|REL \cap GEF|}{|REL|}$$

Beispiel: In einer Sammlung befinden sich 80 relevante Dokumente zu einem bestimmten Thema. Das System X liefert 60 Dokumente zurück, wovon 40 relevant sind. Dann ist die Abdeckung von X $r(X) = 40/80 = 50\%$. In einer idealen Welt wäre die Abdeckung 100%. Weil dies jedoch einfach zu erreichen ist (indem alle Dokumente zurückgeliefert werden), versucht ein System sowohl Abdeckung als auch Präzision zu maximieren. Ein Problem bei der Berechnung dieser Formel stellt die Grösse (REL) dar. Wie findet man heraus, wieviele Dokumente zu einer Anfrage relevant sind? Man müsste alle Dokumente durchsehen, was bei grossen Sammlungen natürlich nicht möglich ist. Durch Stichproben oder Erweiterung der Suchanfrage versucht man die Menge aller relevanten Dokumente zu einer Anfrage abzuschätzen [Weiss 1997].

[6] **Präzision**: Ein Standardmass für die Evaluierung von IR-Systemen. Die Präzision gibt den Anteil der gefundenen (GEF) relevanten (REL) Dokumente an allen gefundenen Dokumenten (GEF) wieder.

$$\text{Präzision: } p = \frac{|REL \cap GEF|}{|GEF|}$$

Beispiel: In einer Sammlung befinden sich 80 relevante Dokumente zu einem bestimmten Thema. Das System X liefert 60 Dokumente zurück, wovon 40 relevant sind. Dann ist die Präzision von X $p(X) = 40/60 \approx 67\%$. In einer idealen Welt wäre die Präzision 100%. Weil dies jedoch einfach zu erreichen ist (indem nur ein relevantes Dokument zurückgeliefert wird), versucht ein System sowohl Präzision als auch Abdeckung zu maximieren [Weiss 1997].

[Fuhr 1998, S. 49].

Beim Indexieren werden dem Dokument Begriffe zugeordnet, die den Inhalt des Dokuments möglichst gut beschreiben. Diese Begriffe nennt man Deskriptoren oder auch (Index-)Terme[7]. Die Wahl dieser Begriffe zur Indexierung ist keine triviale Aufgabe. Sie sollen das Dokument gut beschreiben und gleichzeitig auch vom Benutzer mit grosser Wahrscheinlichkeit als Suchterme gewählt werden. Um eine hohe Abdeckung und eine hohe Präzision zu erreichen, muss mit einer grossen *Indexierungsgenauigkeit* gearbeitet werden. Das heisst, es muss darauf geachtet werden, dass *erschöpfend* (*exhaustive*) und mit einer grossen *Spezifität* (*term specifity*) indexiert wird. Erschöpfend ist eine Indexierung, wenn alle Themen, die in einem Dokument behandelt werden, durch Deskriptoren abgedeckt sind. Je erschöpfender indexiert wird, desto höher ist der Anteil relevanter Dokumente, die bei einer Anfrage zurückgeliefert werden, da alle Teilaspekte der Dokumente erfasst sind. Eine Erhöhung der bereits erwähnten Abdeckung ist die Folge. Die Spezifität hingegen bezieht sich auf das Indexierungsvokabular. Je präzisere und spezifischere Begriffe verwendet werden um zu indexieren, desto besser lassen sich die einzelnen Dokumente unterscheiden. Es werden weniger irrelevante Dokumente zu einer bestimmten Anfrage zurückgeliefert und somit wird die Präzision verbessert.

3.1 Arten der Indexierung

Man unterscheidet normalerweise zwischen *freiem* und *gebundenem Indexieren*. Beim *freien* Indexieren stammen die Deskriptoren entweder aus dem Titel oder Text des Dokuments[8], oder der Indexierer wählt selbst einen Term, der seiner Meinung

[7] Einige Autoren sprechen jedoch nur bei der Verwendung eines Thesaurus (vgl. Abschnitt 3.2.2) von Deskriptoren, während Begriffe, die aus dem Text der Dokumente stammen oder frei gewählt sind, einfach als Indexterme bezeichnet werden. Neben dem Indexieren gibt es auch noch andere Formen der inhaltlichen Erschliessung, wie zum Beispiel Kurzfassungen, Annotationen oder Sachtitel. Diese Formen ergänzen das Indexieren, für die automatische Bearbeitung sind jedoch Deskriptoren als Zugriffspfad zu den Dokumenten praktisch unumgänglich [Grossmann 1984b, S. 5].

[8] Diese Deskriptoren werden als Titel- oder Text-Stichwörter bezeichnet.

nach das Dokument gut beschreibt[9]. Man spricht dann von einem *unkontrollierten* Beschreibungsvokabular. Beim *gebundenen* Indexieren werden die Deskriptoren einem festen Ordnungsschema, einer *Dokumentationssprache* entnommen. Das Vokabular wird auf diese Weise einheitlich und *kontrolliert*. Man versucht damit, die Mehrdeutigkeiten der natürlichen Sprache aufzulösen und daraus resultierende Fehlinterpretationen zu verhindern.

Eine weitere Unterscheidung wird zwischen *singulären* und *kontextbezogenen* Deskriptoren gemacht. Singuläre Deskriptoren sind, wie der Name schon sagt, von einander unabhängige Einzelbegriffe, die das Dokument beschreiben. Erst bei einer Suchanfrage eines Benutzers werden die Begriffe miteinander in Verbindung gebracht. Es ist der Benutzer, der diese Bindung kreiert, die Beschreibungen der Dokumente bleiben davon unangetastet. Nur für die Dauer dieser Suchanfrage sind die Begriffe zueinander in Beziehung gesetzt, sogenannt koordiniert. Da es erst nach der Indexierung stattfindet, nennt man ein solches System *postkoordiniert*. Ein Beispiel: Ein Buch wird mit den einzelnen Termen „information“, „retrieval“ und „computerunterstützt“ indexiert. Bei der Suchanfrage [„information“ AND „retrieval“] wird dann nach Dokumenten gesucht, die sowohl „information“, wie auch „retrieval“ enthalten. Die gefundenen Dokumente erfüllen diese Beziehung, ihre Repräsentation im System bleibt jedoch unverändert und ist weiterhin durch Einzelbegriffe beschrieben. Kontextbezogene Deskriptoren hingegen werden schon bei der Indexierung zueinander in Beziehung gesetzt, sogenannt *präkoordiniert*, was eine bessere Beschreibung des Inhalts ergibt. So könnte das gleiche Buch wie oben zum Beispiel mit „computerunterstütztes information retrieval“ indexiert werden. Es gibt verschiedene Ansätze, wie die Präkoordination ausgeführt wird. Mehrwortbegriffe mit Substantiven, Adjektiven und Präpositionen können gebildet werden, es kann mit Relationen gearbeitet werden, aber auch Rollen wie „Aktion“, „Instrument“, „Subjekt“ etc. können

[9]Sogenannte Schlag- oder Sachwörter. Im Englischen macht man die Unterscheidung zwischen den textimmanenten Stichwörtern und den frei gewählten Schlag- und Sachwörtern nicht. Beide werden einfach *keywords* genannt.

eingesetzt werden [Salton und McGill 1987, S. 61].

Um die Qualität der Indexierung zu verbessern, werden die Deskriptoren oftmals noch gewichtet[10]. (Die automatische Gewichtung wird in Abschnitt 5.2.2 behandelt.) Das Gewicht soll ausdrücken, wie genau der vergebene Deskriptor das Dokument beschreibt. Dies ermöglicht dem System bei einer Anfrage die Liste der gefundenen Dokumente in eine Reihenfolge zu bringen, in der das relevanteste, das heisst das Dokument mit der grössten Summe der Gewichte, an erster Stelle steht. Für den Benutzer ist dies natürlich ein Vorteil, da er nach einer gewissen Anzahl durchgesehener Dokumente seine Suche beenden kann, ohne Bedenken haben zu müssen, dass das relevanteste erst an letzter Stelle stehen könnte. Zusätzlich kann das sortierte Resultat dem Benutzer Hinweise geben, wie er eine neue Frage formulieren soll. Bei einigen Systemen kann der Benutzer auf das Resultat Einfluss nehmen, indem es ihm erlaubt ist, seine Anfrageterme ebenfalls zu gewichten. Meistens kann er auch die Schwellwerte, bis zu welcher Ähnlichkeit Dokumente ausgegeben werden sollen, bestimmen.

[10] Es gibt verschiedene Arten, wie man Terme gewichten kann. Folgende Kriterien werden dabei beachtet:

- lokale Einflüsse
  - Häufigkeit des Terms im Dokument
  - Häufigkeit des Terms in Relation zum häufigsten Term im Dokument
  - bei Dokumenten mit Struktur: Gewichtung nach Ort des Vorkommens (z.B. wird ein Wort im Titelfeld als wichtiger eingestuft)
  - Gewichtung nach Position: Wichtige Terme eher am Anfang des Dokuments (z.B. Nachrichtenmeldungen haben immer den gleichen Aufbau: 1. wesentliche Neuigkeit, 2. Hintergründe, 3. Details und Kommentare)
- globale Einflüsse
  - Dokumentenhäufigkeit eines Terms (Terme, die in vielen Dokumenten vorkommen sind nicht geeignet zur Beschreibung und werden als Indexterme schwach gewertet)

In der Regel werden lokale und globale Kriterien bei der Gewichtung kombiniert, woraus dann die verwendete Gewichtung des Terms resultiert.

3.2 Dokumentationssprachen

Eine von den Textformulierungen relativ unabhängige Repräsentation soll eine Lösung von den konkreten sprachlichen Formulierungen ermöglichen. Durch den Gebrauch eines speziellen Vokabulars versucht man morphologische und syntaktische Probleme und Mehrdeutigkeiten der natürlichen Sprache zu vermeiden. Probleme beim Indexieren bereiten vor allem Derivations- und Flexionsformen, Homographen und Polyseme[11]. Diese Formen muss man auf eine eindeutige Form zurückführen. Ein weiteres Problem sind Komposita, wobei die Letzteren vor allem im Englischen Schwierigkeiten auslösen. Die Komposita werden dort nicht, wie es zum Beispiel im Deutschen oftmals geschieht, zu einem Wort verschmolzen. Es ist manchmal schwer zu entscheiden, welche Wörter zusammengehören[12]. Die Gesamtheit der Massnahmen, die der Kontrolle des verwendeten Vokabulars dienen, nennt man *terminologische Kontrolle*.

Bekannte klassische Dokumentationssprachen sind die Klassifikationen und die Thesauri. Diese beiden Ansätze werde ich nachfolgend kurz vorstellen. In jüngerer Zeit werden aber auch Methoden verwendet, die sich an der Künstlichen Intelligenz orientieren. Im Blickpunkt steht dabei vor allem KL-ONE, ein Formalismus für semantische Netzwerke, der in letzter Zeit für die Datenbankforschung zunehmend wichtiger geworden ist. Ich werde darauf jedoch nicht weiter eingehen[13].

[11] Homographen sind Worte, die man zwar gleich schreibt, aber unterschiedliche Bedeutungen haben. Ursprünglich waren es verschiedene Wortformen, kamen dann aber im Verlauf der Jahre zur Deckung. Z.B. Kiefer: (a) mittelhochdeutsch *kiver* – „Der Kiefer“, (b) althochdeutsch *kienforha* – „Die Kiefer“ (Kienföhre). Anders bei der Polysemie. Polysem ist ein Wort, wenn sich eine Wortform mit der Zeit in verschiedene Bedeutungen aufspaltet, so geschehen bei „Schloss“ (Gebäude, Schliessvorrichtung). Es ist also lediglich ein historisches Kriterium, das zur Unterscheidung von Polysemie und Homonymie herangezogen wird [Linke et al. 1996, S. 141].

[12] Ein schönes Beispiel aus [Hess 1999, S. 190]: „Airport long term car park courtesy vehicle pickup point“. Diese Nominalverkettung soll auf einer Hinweistafel am Flughafen Gatwick gefunden worden sein.

[13] Eine kurze Einführung kann man in [Fuhr 1998, S. 72-78] nachlesen.

3.2.1 Klassifikation

Klassifikationen strukturieren Themen oder Objekte eines Wissensgebietes nach einem mehr oder weniger formalen Schema. Die Objekte werden aufgrund ihrer Merkmale in Klassen und Unterklassen eingeteilt. Das Klassifikationsschema lässt in der Regel nur *eine* Sicht auf die Objekte zu und ordnet somit einem Objekt auch nur *eine* Klasse zu. Hierarchische Klassifikationssysteme werden als Baumstrukturen dargestellt. Eine Menge von Objekten oder Themen wird in Teilmengen aufgeteilt, die disjunkt zueinander sind. Alle Objekte müssen durch diese Teilmengen abgedeckt sein, die ebenfalls wieder in disjunkte Teilmengen zerlegt werden. Die Objekte in der gleichen Teilmenge gehören jeweils zur selben Klasse.

Die *internationale Dezimalklassifikation* ist wohl das bekannteste Beispiel eines Klassifikationssystems. Sie soll einen sachgebietsorientierten Zugang zu allen Wissensgebieten ermöglichen und geht auf Melvil Dewey (1851–1931) zurück, der 1876 „A classification and subject index for cataloguing and arranging the books and pamphlets of a library“ schrieb und damit das erste moderne Klassifikationsschema erfand. Seither wurden immer wieder neue verbesserte Versionen herausgegeben, und das Schema ist vor allem in den USA weit verbreitet. Aus der Dewey Dezimalklassifikation entwickelten die Belgier Paul Otlet und Henri Lafontaine die sogenannte *Universelle Dezimalklassifikation*. Wie der Name schon sagt, ist das Schema an das Dezimalsystem angepasst. Man kann sich das Schema als einen hierarchischen Baum vorstellen, der maximal zehn Verzweigungen pro Knoten aufweist. Ausgegangen wird von 10 Hauptabteilungen (siehe Tabelle 1), die dann jeweils wieder in (bis zu) 10 Unterabteilungen gegliedert sind usw..

Jede Klasse ist durch eine Zahl repräsentiert. Je tiefer man geht in der Hierarchie, desto länger werden die Zahlen, da in jeder Stufe eine Ziffer angehängt wird. Da es höchstens zehn Verzweigungen pro Knoten geben kann, verwendet man die ganzen Zahlen von null bis neun. Zur besseren Lesbarkeit werden die Ziffern in Dreiergruppen zusammengefasst. Ein Beispiel für einen Pfad durch die Klassifikationshierarchie

0	Allgemeines
1	Philosophie
2	Religion, Theologie
3	Sozialwissenschaften, Recht, Verwaltung
4	(Zur Zeit nicht belegt)
5	Mathematik, Naturwissenschaften
6	Angewandte Wissenschaften, Medizin, Technik
7	Kunst, Kunstgewerbe, Photographie, Musik, Spiel, Sport
8	Sprachwissenschaft, Philologie, Schöne Literatur, Literaturwissenschaft
9	Heimatkunde, Geographie, Biographien, Geschichte

Tabelle 1: Die zehn Hauptabteilungen der universellen Dezimalklassifikation. Quelle: [Rechenberg und Pomberger 1999, S. 917]

ist in Tabelle 2 zu sehen. Wird der Ziffer Fünf (Mathematik, Naturwissenschaften) eine Drei angefügt, ist das die Kennzeichnung für die Klasse „Physik“. Folgt auf die Fünf und die Drei eine Neun, geht man in der Klasse Physik noch eine Stufe tiefer und landet in der (Unter-)Klasse „Physikalischer Aufbau der Materie“ etc..

5	-	Mathematik, Naturwissenschaften
53	-	Physik
539	-	Physikalischer Aufbau der Materie
539.1	-	Kernphysik, Atomphysik, Molekülphysik
539.17	-	Kernreaktionen
539.172	-	Individuelle Kernreaktion
539.172.1	-	Kernreaktionen durch Atomkerne
539.172.13	-	Kernreaktionen durch Deuteronen

Tabelle 2: Beispiel eines Pfades durch die Klassifikationshierarchie. Quelle: [Rechenberg und Pomberger 1999, S. 917]

Zur Zeit enthält das ganze System über 130’000 Klassen, wobei jede Klassenzahl noch durch bestimmte Sonderzeichen eingeleitete Anhängezahlen erweitert werden kann, die die Klasse noch facettieren. Beispiele für solche Facettierungen sind in Tabelle 3 aufgeführt. Zum besseren Verständnis noch ein konkretes Beispiel für eine Facettierung: Die Kennzeichnung „53(038)“ steht für Physik-Wörterbuch, da 53 die

=	Sprache
(0...)	Form
(...)	Ort
(=...)	Rassen und Völker
„...“	Zeit
.00	Gesichtspunkt
-05	Person

Tabelle 3: Sonderzeichen zur Facettierung von Klassen

Klasse Physik bezeichnet und (038) für die Form „Wörterbuch“ steht.
Klassifikationen bergen gewisse Nachteile. Da sie aus festgelegten Klassen bestehen, sind sie wenig flexibel und schwer an neue Entwicklungen anzupassen. Beim Suchprozess muss sich der Benutzer schon früh zwischen Ästen der Klassifikation entscheiden, auch wenn ihm zu diesem Zeitpunkt die Gründe für diese Wahl noch nicht ersichtlich sind. Ausserdem können viele Dokumente fächerverbindend sein und zu mehreren Klassen gehören, was die Zuordnung zu nur einer Klasse unmöglich macht. Um die Systeme flexibler zu machen, können Polyhierarchien und Doppelstellen, sowie Polydimensionalität zugelassen werden. Polyhierarchien erlauben, dass eine Klasse zwei Superklassen, also zwei „Eltern“ -Klassen hat. Damit lassen sich Beziehungen zwischen Klassen besser darstellen. Polydimensionalität ist hilfreich, wenn es auf einer Stufe mehrere orthogonale Merkmale gibt, die in Untergruppen eingeteilt werden können. Orthogonal sind Merkmale, die Objekte einer Klasse nach verschiedenen Gesichtspunkten beschreiben, welche einander nicht ausschliessen. Für jedes orthogonale Merkmal wird eine eigene Richtung eingeführt, in der dann die Klassifikationshierarchie weitergeführt wird. Das heisst, ein Objekt könnte in mehreren dieser Richtungen eingeordnet werden, wird aber jeweils aus verschiedenen Sichten beurteilt. Abbildung 2 soll dies illustrieren. Auf der Stufe „Obstbaum“ gibt es zwei orthogonale Merkmale, sozusagen zwei Sichtweisen auf die nun zu klassifizierenden Objekte: Die Unterscheidung nach Fruchtart und die Unterscheidung nach Stammbildung. Für beide orthogonalen Merkmale wird nun je eine Richtung eingeführt.

In der einen Richtung stehen die Unterklassen von Obstbaum, die nach Stammbildung voneinander unterschieden werden, in der anderen Richtung die Unterklassen von Obstbaum, die nach Fruchtart voneinander unterschieden werden. Es geht also zweidimensional weiter in der Klassifikation. Will man nun eine bestimmte Sorte Apfelbaum klassifizieren, wird er sicher unter Kernobstbaum eingeordnet, aber auch in die Kategorie der halbstämmigen Obstbäume. Würde man Polydimensionalität

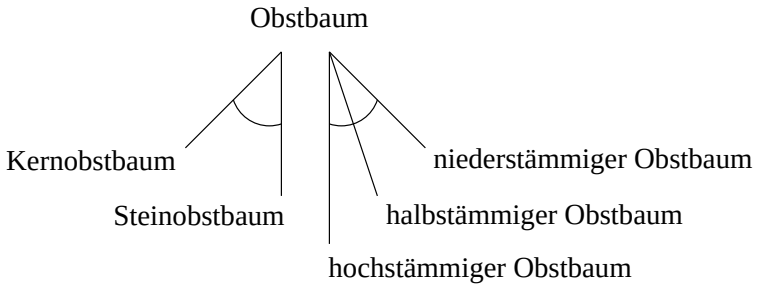

Abbildung 2: Polydimensionaltät. Quelle: [Fuhr 1998, S. 61]

nicht zulassen, könnte man das Beispiel wie in Abbildung 3 darstellen. Es muss eine neue Ebene für die orthogonalen Merkmale eingeführt werden.

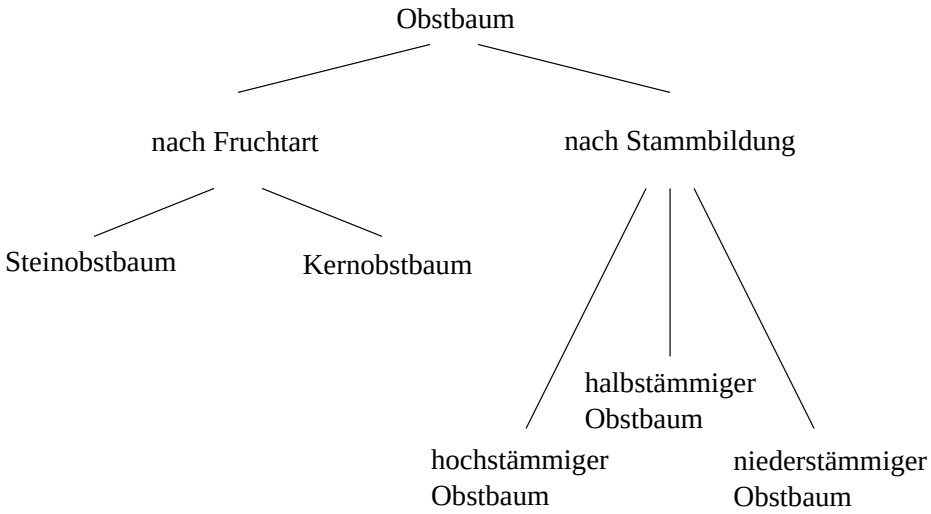

Abbildung 3: Aufgelöste Polydimensionaltät. Quelle: [Fuhr 1998, S. 61]

3.2.2 Thesauri

Ein Thesaurus ist eine geordnete Zusammenstellung von Begriffen eines Fachgebietes mit ihren (natürlichsprachlichen) Beziehungen. Es werden Synonyme[14], Homographen und Polyseme erfasst. Zusätzlich werden durch Beziehungen wie Ober- und Unterbegriffe und verwandte Begriffe auch hierarchische Strukturen erzeugt. Thesauri werden verwendet, um die Dokumente einer Sammlung möglichst eindeutig zu indexieren. Um dies zu erreichen, wird ein kontrolliertes Vokabular verwendet, das heisst, es werden bei der Indexierung nur Begriffe zugelassen, die in diesem Fachgebiet wohl definiert sind und den jeweiligen Sachverhalt genau beschreiben, wobei für einen Sachverhalt nur ein Begriff verwendet wird. Im IR spielen die Teil- und Quasisynonyme[15] eine grosse Rolle, da sie sehr häufig auftreten. Es ist nicht immer sinnvoll, alle diese Wörter als Deskriptoren zu verwenden. Meist fasst man diese Wörter zu sogenannten Äquivalenzklassen zusammen und führt nur einen Begriff als Stellvertreter. Synonyme, die nicht als Deskriptoren zugelassen sind, haben im Thesaurus lediglich einen Verweis, welcher Begriff an ihrer Stelle benutzt werden soll. Erst beim als Stellvertreter für die Äquivalenzklasse gewählten Deskriptor stehen die Informationen zu diesem Begriff oder besser zur gesamten Äquivalenzklasse. Dieser Eintrag enthält auch rückführende Verweise auf die Synonyme, für welche der Deskriptor als Stellvertreter benutzt wird.

[14]Synonym heisst bedeutungsgleich. Ein Beispiel wäre *anfangen – beginnen*. Synonymie gibt es nicht nur zwischen Worten, sondern vor allem auch zwischen satzartigen Gebilden: *Die Nadel ist nicht lang genug – Die Nadel ist zu kurz* oder *Es war schwer, die richtige Antwort zu finden. – Die richtige Antwort zu finden, war schwer. – Die richtige Antwort war schwer zu finden.* Auch Aktiv – Passiv Konstruktionen und Paraphrasen sind eine Art Synonymie. (Wie weit dieser Begriff gefasst wird, ist verschieden.) Die „echte“ Synonymie, also Wörter, die in ihrer Bedeutung exakt übereinstimmen und ohne Bedeutungsänderung einander ersetzen können, ist zwischen festen Wortformen jedoch selten [Linke et al. 1996, S. 142].

[15]Teilsynonyme sind Worte, die eigentlich nicht ganz die gleiche Bedeutung haben, aber trotzdem meistens im Wesentlichen für das Gleiche gebraucht werden. Häufig bilden sich solche Beziehungen zwischen Ober- und Unterbegriffen. Z.B. *Ei – Hühnerei*, *Holland – Niederlande*. Quasisynonyme sind Wörter, die sich zwar auf das gleiche Objekt beziehen, aber andere Bedeutungsaspekte hervorheben. Beispiele: *Pferd – Gaul*, *Frau – Weib*, *sterben – entschlafen – krepieren...* Diese Relationen werden auch einfach unter dem Begriff „Bedeutungsähnlichkeit“ zusammengefasst. Aus: [Linke et al. 1996, S. 142f.] und [Grossmann 1984a, S. 1f.]

Die Wahl der Synonyme legt fest, wie detailliert der Thesaurus das Fachgebiet erschliesst. Fasst man den Synonymbegriff weit und ordnet ihm auch verwandte Begriffe zu, werden weniger Details erfasst. Will man aber eine sehr genaue Zuordnung ermöglichen, dürfen wirklich nur Terme mit exakt der gleichen Bedeutung als Synonyme definiert werden, und breit verwendete Begriffe müssen in verschiedene Deskriptoreinträge aufgespalten werden, um die jeweiligen Bedeutungsnuancen zu erfassen.

Ein Thesauruseintrag enthält neben dem Deskriptor, den Verweisen zu den Synonymen (*use-for*-Relationen), den Ober- und Unterbegriffen und Verwandschaftsrelationen oftmals eine Kennziffer für den Deskriptor und eine sogenannte Scope-Note, das ist eine kurze Beschreibung der Bedeutung des Begriffs. Ausserdem wird meistens noch das Datum der Einführung oder der Streichung angefügt, um die Aktualität zu dokumentieren. Ein Beispiel für einen Thesauruseintrag findet man in Tabelle 4. In der rechten Spalte werden die verwendeten Abkürzungen aufgelöst.

Id :	3704	Kennziffer
Kw:	Agroindustry	Keyword: Deskriptor
Sn :	Involves The Integration Of Agricultural Production, Processing And Marketing Under A Single Comprehensive Management	Scope Note
Bt :	Industrial Sector	Broader Term: Oberbegriff
Nt :	Fishery Industry	Narrower Term: Unterbegriff
Nt :	Food Industry	
Nt :	Forestry Industry	
Nt :	Leather Industry	
Nt :	Oil Industry	
Rt :	Agriculture	Related Term: Verwandter Begriff
Rt :	Agriproduct Processing	
Uf :	Agricultural Industry	Use for: Benutze für
Uf :	Agroindustrial Complex	

Tabelle 4: Ein Eintrag aus dem OECD-Thesaurus, der viersprachig ist und die Bereiche ökonomische und soziale Entwicklung abdeckt (5. Edition 1998). Quelle: [Rechenberg und Pomberger 1999, S. 918]

Ein Thesaurus ist alphabetisch oder systematisch geordnet und mit einem Index versehen. In der Regel liegt der Thesaurus auch in gedruckter Form vor, um einfachen Zugriff darauf zu ermöglichen. So kann der Thesaurus in einem IRS von Indexierern und Nutzern dazu verwendet werden, geeignete Deskriptoren zu wählen. Durch die Relationen ist es ausserdem möglich, die Suche mit Oberbegriffen zu verallgemeinern, mit Unterbegriffen zu spezifizieren oder durch verwandte Terme auf ähnliche Dokumente zu stossen [Rechenberg und Pomberger 1999, S. 918f.].

3.3 Gebundenes Indexieren vs. freies Indexieren

Das gebundene Indexieren, also das Indexieren mit einer Dokumentationssprache, hat gegenüber dem freien Indexieren einige Vorteile. Es bildet mehrere Textformulierungen auf das Gleiche ab, wodurch der Recall erhöht werden kann. Da keine mehrdeutigen Begriffe in einer Dokumentationssprache zugelassen sind, kann ebenfalls eine höhere Präzision erreicht werden. Der Benutzer hat ausserdem den Vorteil, dass ihm die Formulierung seiner Anfrage weniger Schwierigkeiten bereiten sollte. Er muss sein Informationsbedürfnis lediglich auf die entsprechende Benennung der Dokumentationssprache abbilden. Die in Abschnitt 2 vorgestellten Probleme des unsicheren Wissens und der vagen Anfragen können auf diese Weise etwas entschärft werden [Fuhr 1998, S. 78].

Das gebundene Indexieren hat aber nicht nur Vorteile. Es muss zuerst ein Ordnungssystem erprobt werden, und dann muss dieses nicht nur dem Indexierer, sondern auch dem Benutzer zugänglich sein, damit er geeignete Deskriptoren für seine Anfrage wählen kann. Einer der grössten Nachteile von Dokumentationssprachen wie Klassifikationen und Thesauri ist jedoch ihre Pflege. Es besteht immer die Gefahr, dass man einzelne Dokumente nicht einem passenden Index zuordnen kann. Durch Entwicklungen des Fachgebietes, der Fachsprache und Verbesserungen in Bezug auf die Suchergebnisse ist eine ständige Anpassung notwendig, wobei aber die Konsistenz erhalten bleiben muss [Grossmann 1984b, S. 6]. Vor allem bei einer Ausweitung

auf neue Benutzergruppen oder bei der Abdeckung von neuen Sachgebieten kann es sogar nötig sein, das Ordnungssystem vollständig neu aufzubauen. Die Zusammenfassung von Begriffen zu Äquivalenzklassen in der Dokumentationssprache kann bei sehr spezifischen Anfragen gegenüber dem freien Indexieren zu einem Verlust der Präzision führen. Auch ist der Einsatz von kontrolliertem Vokabular wissenschaftlich sehr umstritten[16]. Für die Indexierer und Benutzer bedeutet es einen Mehraufwand mit der Dokumentationssprache umzugehen. Im Grunde genommen sollte die natürliche Sprache des Textes doch genau die richtigen Ausdrücke liefern, die das Dokument gut beschreiben [Salton und McGill 1987, S. 64]. Will man jedoch ohne Dokumentationssprachen auskommen, muss noch ein Weg gefunden werden, wie ohne Thesaurus synonyme und verwandte Begriffe für die Informationssuche zur Verfügung gestellt werden können. Auch stellt man heute fest, dass sich gewisse halb-formale Konzepte wie zum Beispiel Datums- und geographische Angaben mit freier Indexierung nicht abdecken lassen [Fuhr 1998, S. 78f.]. Die gebundene Indexierung erhält wieder Aufschwung. Vor allem Systeme, die beide Arten kombinieren, scheinen gute Ergebnisse zu liefern [Chowdhury 1999, S. 119ff.].

[16][Fuhr 1998, S. 78f.] führt dies auf die unzureichende experimentelle Basis für den Vergleich zwischen gebundenem und freiem Indexieren zurück. Experimente, die sich mit diesem Problem beschäftigen und die Meinung prägten, dass Dokumentationssprachen überflüssig sind, sind schon ziemlich alt und wurden mit kleinen Datenbanken durchgeführt, die sich nicht mit der Grösse der heutigen vergleichen lassen. Ausserdem sind heutzutage immer mehr maschinenlesbare Wörterbücher und grosse Thesauri verfügbar, die natürlich nötig sind für umfassende IRS mit gebundener Indexierung.

4 Manuelles Indexieren

Ich werde in diesem Abschnitt nur kurz auf die manuelle Indexierung eingehen, da der wichtigste Teil, die Wahl der Deskriptoren, eine intellektuelle Arbeit ist, und natürlich stark von der Art des Systems abhängt. Ich werde dafür die automatische Indexierung in Abschnitt 5 vertieft behandeln.

Bei der manuellen Indexierung durch Fachpersonen wird meistens mit einer Dokumentationssprache gearbeitet. Das Vokabular wird kontrolliert und es wird mit zusammengesetzten präkoordinierten Begriffen indexiert. Das Vorgehen ist streng vorgegeben und läuft nach genauen Regeln ab, um eine Konsistenz der Einträge zu erreichen. Den Indexierern stehen verschiedene Hilfsmittel wie Terminologielisten, Regelwerke, Thesauri, hierarchisch gegliederte Wörterbücher und Arbeitsblätter zur Verfügung. Auch heute ist die manuelle Indexierung noch weit verbreitet. Wird mit einer grossen Indexierungsgenauigkeit und konsistent gearbeitet, erhöht die terminologische Kontrolle und die kontextbezogene Präkoordination die in Abschnitt 3 erwähnten Masse Abdeckung und Präzision. Eine gute Indexierungsgenauigkeit ist aber schwer zu erreichen und sehr zeitraubend. Oftmals sind die vergebenen Deskriptoren nicht ausreichend oder zu wenig spezifisch. Inkonsistenzen sind nicht zu vermeiden: Verschiedene Indexierer liefern bei der Analyse der gleichen Dokumente unterschiedliche Ergebnisse, was Schwierigkeiten beim Abgleich von Suchanfrage und Dokumenten verursachen kann.

Heute ist es auch möglich, halbautomatisch oder vollautomatisch zu indexieren, wobei diese Verfahren vorwiegend auf statistischen Methoden basieren und mit Termhäufigkeiten arbeiten. Experimente zeigen, dass eine Qualität der Indexierung erreicht werden kann, die genauso gute Ergebnisse wie die manuelle Indexierung liefert [Salton und McGill 1987, S. 60-65].

5 Automatisches Indexieren

Beim automatischen Indexieren geht man davon aus, dass man geeignete Deskriptoren aus dem Text der Dokumente gewinnen kann. Als Quelle kann der ganze Text dienen, aber auch Abstracts[17] oder nur Titel können verwendet werden. Die Schwierigkeit beim automatischen Indexieren besteht darin, sinnvolle Deskriptoren zu finden, ohne den Text verstehen zu müssen. Von der automatischen Indexierung erhofft man sich mehr Konsistenz als beim manuellen Vorgehen und es wird eine Reduzierung von Zeitaufwand und Kosten angestrebt. Natürlich will man auch den Retrievalprozess selbst verbessern und effizienter machen.

5.1 Freitextsuche

Die Freitextsuche ist die einfachste Art automatisch zu indexieren[18]. Die Deskriptoren werden dabei aus dem Text der Dokumente gewonnen. Es wird keine Dokumentationssprache verwendet, sondern frei indexiert. Die Begriffe werden nicht zueinander in Beziehung gesetzt, sondern das Prinzip der Postkoordination wird angewendet. Die heutigen kommerziellen IR-Systemen benutzen meistens die Freitextsuche. Folgende Verarbeitungsschritte sind auszumachen:

1. Der Text wird in einzelne Wörter zerlegt, wobei Leer- und Interpunktionszeichen als Worttrenner aufgefasst werden.

2. Stoppwörter werden eliminiert. Stoppwörter sind nicht-bedeutungstragende Wörter wie Artikel, Konjunktionen etc.. Meist machen diese Wörter etwa die Hälfte des Textes aus!

[17] Es gibt unterschiedliche Definitionen für „Abstract“. Allgemein könnte man sagen, dass ein Abstract eine kurze Zusammenfassung ist, die die wichtigsten Inhalte eines Dokuments darstellt. Zusammenfassung also nicht im Sinne wie sie z.B. am Ende eines Artikels steht, sondern mehr in Form eines Kurzreferats.

[18] Ich folge im Abschnitt über die Freitextsuche hauptsächlich den Ausführungen von [Fuhr 1998, S. 50-59]

3. Bei einigen Systemen wird versucht, die Satzenden zu erkennen. Dies ermöglicht bei einer Anfrage die Angabe, ob die Suchterme im gleichen Satz (oder eben nicht im gleichen Satz) vorkommen müssen. Die Erkennung der Satzenden ist keineswegs eine triviale Aufgabe. Es gibt zahlreiche Probleme vor allem durch die Verwechslungsmöglichkeiten mit Abkürzungspunkten, Punkten in Zeitangaben und Ähnlichem [Hess 1999, S. 38ff.]. Darum kann diese Aufgabe nur approximativ gelöst werden.

Alle übrig bleibenden Wörter werden als Indexterme benutzt. Bei der Indexierung wird eigentlich keine neue unabhängige Repräsentation erstellt, sondern es werden Funktionen zur Verbesserung der Suche im Text bereitgestellt. In kleinen Sammlungen können die genannten Verarbeitungsschritte jeweils direkt vor der Suche durchgeführt werden, bei grösseren Sammlungen ist dies jedoch unmöglich. Meistens wird darum dieser reduzierte Text – oder besser die extrahierte Folge von Wörtern – in einem File (Index) gespeichert, worauf dann die Suche abläuft[19].
Wie geht man bei dieser Indexierungsart mit den verschiedenen Wortformen um? Vor allem die Flexions- und Derivationsformen muss man in den Griff bekommen. Diese Probleme versucht man mit verschiedenen Ansätzen zu lösen. Unterscheiden könnte man zwischen dem „Informatischen Ansatz“ und dem „Computerlinguistischen Ansatz“.

5.1.1 Informatischer Ansatz

Beim informatischen Ansatz wird rein auf der Zeichenkettenebene operiert. Texte sind also lediglich eine Folge von Wörtern. Ein Wort ist definiert als eine Zeichenfolge, die durch Leer- oder Interpunktionszeichen begrenzt ist. Es werden nun Operato-

[19]Bei der Speicherung wird häufig das Prinzip der invertierten Datei benützt. Im Indexfile wird bei jedem Deskriptor eine Referenz für jeden Text, den er beschreibt, abgespeichert. Bei einer Abfrage nach einem bestimmten Deskriptor wird auf das Indexfile zugegriffen und nach dem Term gesucht. Danach werden alle Texte, für die dort eine Referenz gespeichert ist, ausgegeben. Um eine schnelle Antwortzeit zu erreichen, ist es notwendig, dieses Indexfile möglichst zugriffsoptimal zu speichern. Die invertierten Dateien sind sehr geeignet für die Suche mit booleschen Operatoren.

ren für die Zeichenketten*suche* bereitgestellt. Es ist keine Indexierungsart, sondern verschiebt das Problem auf den Benutzer. Ihm werden Maskierungs-, Trunkations- und Kontextoperatoren bereitgestellt, um unabhängig von den verschiedenen Wortformen suchen zu können.

Maskierungs- und Trunkationsmethoden werden verwendet, um Derivations- und Flexionsformen zu vereinheitlichen. Bei einer Anfrage soll das einzelne Aufzählen der verschiedenen Formen vermieden werden und trotzdem sollen alle Ausformungen eines Wortes gefunden werden.

Die *Trunkation* operiert am Anfang und Ende eines Wortes. Es wird unterschieden zwischen Front- und End-Truncation. Die Front-Truncation dient dazu, verschiedene Präfixe (Vorsilben) bei der Suche zuzulassen, während die End-Truncation vor allem an den Wortstamm gehängte Flexionsendungen und andere Suffixe zusammenführt. In den folgenden Beispielen in Tabelle 5 wird für die beschränkte Trunkation (das heisst ein Operatorsymbol steht für genau ein Zeichen im Wort) das Symbol $, für die unbeschränkte Trunkation (das Operatorsymbol steht für eine beliebig lange Zeichenkette) das Symbol # verwendet. Auf der linken Seite steht jeweils die durch die Operatoren modifizierte Eingabe des Benutzers, rechts werden Beispiele für Wortformen aufgeführt, die mit dieser Eingabe gefunden würden.

schreib#:	schreiben, schreibt, schreibst, schreibe
schreib$$:	schreiben, schreibst
#schreiben:	schreiben, beschreiben, anschreiben, verschreiben
$$schreiben:	beschreiben, anschreiben

Tabelle 5: Beispiele für beschränkte und unbeschränkte Trunkation. Quelle: [Fuhr 1998, S. 51]

Die *Maskierung* liefert die gleichen Funktionen wie die Trunkation für die Mitte eines Wortes. Vor allem bei der Deklination und Konjugation von deutschen Wörtern ist die Maskierung wichtig, um die Veränderungen im Wortinnern zu erfassen. Tabelle 6 soll dies verdeutlichen.

schr$$b#: schreiben, schrieb / schrauben
h$$s#: Haus, Häuser / Hanse, Hass, hassen, hausen

Tabelle 6: Beispiele für die Maskierung. Quelle: [Fuhr 1998, S. 51]

Mit der zunehmenden Trunkierung und Maskierung wächst die Wahrscheinlichkeit, dass auch unerwünschte Wörter geliefert werden. Das obige Beispiel von h$$s# illustriert dieses Aspekt sehr schön. Indem man das Suchmuster so verallgemeinert, dass man neben der Singularform von „Haus“ auch die Pluralform „Häuser“ gewinnen kann, werden ebenfalls Wörter von vollkommen anderen Wortfamilien geliefert. Es entstehen grosse Verluste bei der Präzision. Um dieses Problem zu entschärfen, bieten die meisten Systeme die verschiedenen Wortformen, die ein Suchmuster erfüllen, zur Auswahl an. Der Benutzer wählt aus diesen Formen die gewünschten aus.
Kontextoperatoren werden vor allem dazu benutzt, Komposita zu suchen. Sie sind vor allem für das Englische wichtig, da die Komposita dort aus mehreren Worten zusammengesetzt sind. So kann zum Beispiel „information retrieval“ auch als „retrieval of information“ oder in der Form „information storage and retrieval“ in einem Text vorkommen, wobei ohne zusätzliche Operatoren diese Lösungen bei einer Anfrage nach der normalen Form „information retrieval“ jeweils nicht gefunden würden. Viele Systeme bieten daher solche Operatoren an:

- genauer Wortabstand ($ steht für genau ein Wort):
  retrieval $ information: retrieval of information, retrieval with information loss
- maximaler Wortabstand (# steht für kein oder ein Wort):
  information # # retrieval: information retrieval, information storage and retrieval
- Wortreihenfolge (, lässt die Wortreihenfolge unwichtig werden):
  information #, retrieval: informaton retrieval, retrieval of information

- gleicher Satz (. verlangt, dass die Wörter im gleichen Satz vorkommen): information # retrieval.: stimmt weder mit „...this information. Retrieval of data...“ noch mit „...storage of information. Its retrieval...“ überein.

Diese Art von Suche birgt grosse Nachteile. Sie setzt Kenntnisse des Benutzers über diese Abfrageform voraus, was man bei Laien und nur gelegentlichen Nutzern nicht verlangen kann. Darüber hinaus muss sich der Benutzer alle möglichen Formen eines ihn interessierenden Wortes vorstellen, was einen grossen Aufwand für ihn darstellt. Dies ist wahrscheinlich der grösste Nachteil dieses Ansatzes.

5.1.2 Computerlinguistischer Ansatz

Der computerlinguistische Ansatz versucht die gleichen Ziele zu erreichen wie der informatische Ansatz. Nun aber sollen mit Hilfe von Algorithmen Transformationen ausgeführt werden, die automatisch die Derivations- und Flexionsformen, sowie die möglichen Formen bei Komposita erkennen können. Ich werde in dieser Arbeit nicht im Detail darauf eingehen, sondern verweise auf die entsprechende Fachliteratur[20]. [Fuhr 1998, S. 52-59] spricht von folgenden Arten von computerlinguistischen Verfahren:

Graphematische Verfahren: Durch Analyse von Buchstabenfolgen sollen die Flexions- und Derivationsformen zusammengeführt werden. Im Englischen ist dies einfacher zu erreichen als im Deutschen, da das Englische eher schwach flektiert. Die Wörter werden auf ihre lexikalische Grundform oder den Wortstamm zurückgeführt, indem vorgegebene Ersetzungsregeln für die Endungen angewendet werden. Dieses Verfahren funktioniert im Englischen ziemlich gut.

Lexikalische Verfahren: Ein Wörterbuch führt die Flexions- und Derivationsformen mit den zugehörigen Grundformen auf. Dieses Verfahren wird vor allem bei stärker

[20]Für nähere Informationen empfehle ich die Unterlagen von G.Schneider zu den Kursen „Morphologieanalyse und Lexikonaufbau“, sowie „Formale Grammatiken und Syntaxanalyse“ der computerlinguistischen Abteilung der Universität Zürich. Sie sind auf dem Netz unter `http://www.ifi.unizh.ch/cl/study/unterlagen.html` zu finden.

flektierten Sprachen nötig, wo graphematische Verfahren nicht ausreichen. Das Lexikon enthält ebenfalls Komposita und deren Einzelteile, sowie die verschiedenen Bedeutungen von mehrdeutigen Wörtern.
Syntaktische Verfahren: Syntaktische Verfahren dienen hauptsächlich zur Identifikation von mehrgliedrigen Ausdrücken (Komposita).

5.1.3 Nachteile der Freitextsuche

Eine Hauptschwierigkeit, mit der Systeme, die Freitextsuche verwenden, zu kämpfen haben, liegt darin, dass die Wortwahl massgeblich die Ergebnisse der Suche bestimmt. Dieses Problem lässt sich mit keinem der oben genannten Ansätzen lösen. Ausserdem ist es eine ziemlich primitive Methode, die nicht versucht, den Inhalt zu beschreiben, sondern nur die Suche nach Wörtern im Text erleichtert. Für ein effizientes Retrieval ist das nicht ausreichend. Es wird auch keine Lösung für das Problem der Homographen und Polyseme bereitgestellt. Viele automatische Systeme benutzen daher noch einen Thesaurus[21], um diese Wörter in den Griff zu bekommen.

5.2 Statistische Methoden

Es ist nicht sinnvoll jedes Wort, das nicht durch die Stoppwortliste eliminiert wird, als Deskriptor zuzulassen. Dadurch wird in keiner Weise eine adäquate Beschreibung des Inhalts dieses Dokuments geliefert. Themen, die nur in einem Nebensatz erwähnt werden, erhalten die gleiche Bedeutung wie das Hauptthema. Um eine bessere Beschreibung des Dokuments zu erreichen, können zusätzlich zu den oben genannten Schritten statistische Methoden angewendet werden. Man geht davon aus, dass die Häufigkeit, mit der die Worte in einem Dokument oder in der ganzen Dokumentensammlung auftreten, verwendet werden kann, um die Aussagekraft eines Wortes zu messen[22]. Eine einfache Methode wäre Folgendes: Man berechnet für jeden Term die

[21] Diese Thesauri sind meistens von Hand erstellt, da die automatische Generierung von Thesauri noch keine befriedigenden Ergebnisse liefert.
[22] Das ist die sogenannte Hypothese von Luhn [Salton und McGill 1987, S. 65].

Häufigkeit seines Vorkommens im Text. Terme, deren Termhäufigkeit unter einem festgelegten Schwellwert liegen, werden als Indexterme nicht zugelassen und nicht in die Liste eingefügt[23]. Diese Methode ist jedoch zu wenig ausgereift und birgt gewisse Nachteile. Eine hohe Termhäufigkeit heisst nicht unbedingt, dass der Term das Dokument gut beschreibt: In einer Sammlung von Dokumenten über das Gebiet der Informatik wird das Wort „Computer“ ziemlich häufig vorkommen, ohne dass es bei der Unterscheidung zwischen verschiedenen Dokumenten helfen würde. Verwendete man es als Indexterm, würde beinahe die ganze Sammlung bei einer Anfrage geliefert werden. Lässt man aber generell sehr häufige Wörter weg, leidet die Abdeckung darunter. Man kann also nicht mit *absoluten* Termhäufigkeiten operieren, sondern muss mit *relativen* Häufigkeiten arbeiten, die die Wortwahl der anderen Dokumente in der jeweiligen Sammlung miteinbeziehen. Ausgereiftere Ansätze arbeiten daher mit Gewichtungsfunktionen für die einzelnen Terme, die diese Aspekte berücksichtigen. Für jeden Term wird ein Gewicht berechnet, wobei dieses Gewicht die Wichtigkeit des Terms für die Beschreibung des Dokuments ausdrücken soll. Bevor ich jedoch einige Methoden zur Berechnung von Termgewichten vorstelle, führe ich zum besseren Verständnis der späteren Erläuterungen noch die bevorzugte Speicherungsart für Systeme, die mit gewichteten Deskriptoren arbeiten, ein.

5.2.1 Vektordarstellung

Für das Abspeichern der gewichteten Deskriptoren eignet sich besonders die *Vektordarstellung*. Dabei wird jedem Dokument i ein Dokumentvektor $\vec{d_i}$ zugeordnet:

$$\vec{d_i} = (G_i^1, G_i^2, \ldots, G_i^k, \ldots, G_i^m)$$

In der ganzen Sammlung werden m Deskriptoren zur Beschreibung verwendet. Jedes

[23]Sehr selten auftretende Wörter sind oft zu spezifisch und nicht geeignet als Deskriptoren, obwohl sie eine gute Unterscheidungskraft zwischen Dokumenten hätten. Aber Vorsicht: Wo zieht man die Grenze? Gefahr von grossen Verlusten bei der Präzision!

G_i^k stellt das Gewicht dar, das dem Dokument i für den Deskriptor an der k-ten Stelle gegeben wird. Ist eine Komponente G_i^k gleich Null, so bedeutet dies, dass der Deskriptor für dieses Dokument nicht relevant ist, es also nicht mit diesem Term indexiert wird.

Man stellt sich einen m-dimensionalen Raum vor, wobei für jeden Deskriptor eine Richtung gegeben ist. Ein Dokumentvektor lässt sich nun in diesem Raum als Vektor darstellen, dessen Komponenten den dokumentspezifischen Gewichten der einzelnen Deskriptoren entsprechen. Ein Beispiel aus [Burger 1984, S. 6f.] soll dies verdeutlichen:

Nehmen wir an, wir haben als Dokumente vier Kritiken über Restaurants. Nach der Indexierung haben wir drei Deskriptoren (gut, preiswert, billig) gefunden, mit denen die vier Dokumente indexiert werden. Durch die Gewichtung der Deskriptoren der vier Dokumente (für die Berechnung von Gewichten siehe Abschnitt 5.2.2) erhalten wir die in Tabelle 7 abgebildete Gewichte-Matrix.

	gut	preiswert	billig
Restaurant 1	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{2}{3}$
Restaurant 2	$\frac{4}{3}$	$\frac{1}{4}$	0
Restaurant 3	0	$\frac{2}{4}$	$\frac{2}{3}$
Restaurant 4	$\frac{2}{3}$	$\frac{2}{4}$	$\frac{1}{3}$

Tabelle 7: Matrix mit den gewichteten Deskriptoren. Quelle: [Burger 1984, S. 7]

Aus dieser Matrix sind die Vektoren für die einzelnen Dokumente ersichtlich. Sie bestehen aus den einzelnen Zeilen der Matrix. Da wir drei Deskriptoren haben, kann man die Vektoren in einem dreidimensionaler Raum darstellen (Abbildung 4). Für Restaurant 1 sind die einzelnen Komponenten (Gewichte) der Vektoren zur Illustration als gepunktete Linie eingezeichnet.

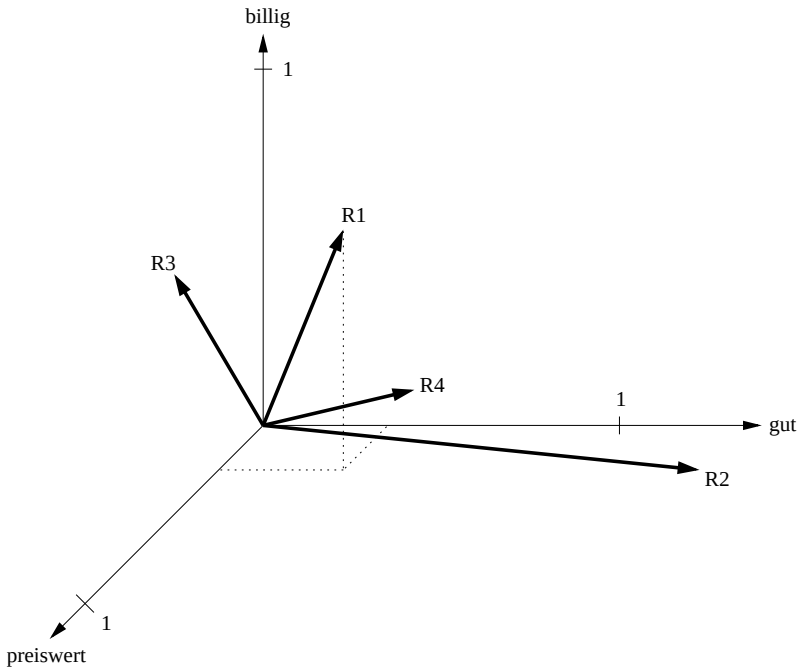

Abbildung 4: Dokumentvektoren in einem Vektorraum. Quelle: [Burger 1984, S. 7]

Die Anfrage wird dann ebenfalls durch einen Vektor dargestellt[24] und mit den Dokumentvektoren verglichen. Dokumente mit einem genügend grossen Ähnlichkeitswert (vgl. Ähnlichkeitsmasse Abschnitt 5.2.2.3) werden ausgegeben.

Um die Effizienz beim Suchprozess zu steigern, fasst man oftmals eine ganze Gruppe verwandter Dokumente zu einer Untermenge, einem sogenannten *Cluster* zusammen. Der Vergleich mit der Abfrage wird dann nur mit einem Stellvertreter für die einzelnen Cluster durchgeführt, wodurch nicht jedes Dokument verglichen werden muss, sondern gleich ganze Gruppen als irrelevant ausgeschieden werden können. Die Suche konzentriert sich dann auf die als genügend relevant eingestuften Cluster. Auch in den Cluster selbst kann man wieder „Sub-“ Cluster bilden, um eine effizientere Suche zu ermöglichen[25].

[24]Oftmals kann der Benutzer die Terme seiner Anfrage selbst gewichten oder die Anfrage in natürlicher Sprache formulieren. Die Anfrage wird dann mit denselben Methoden wie die Dokumente analysiert und indexiert.

[25]Diese Methode eignet sich durch weitere darauf angewandte Funktionen auch zur automati-

5.2.2 Berechnung von Termgewichten

Wie gewichtet man nun die Terme automatisch? Ich werde in den folgenden Abschnitten[26] drei bewährte Methoden vorstellen: Die inverse Dokumentenhäufigkeit, die Berechnung mit den Grössen Signal und Rauschen und die Diskriminationsmethode. Um die Formeln einheitlich darzustellen, werde ich zuerst einige Bezeichnungen einführen:

n: Anzahl Dokumente in der Sammlung

f_i^k: Häufigkeit des Terms k im Dokument i

$F^k = \sum_{i=0}^{n} f_i^k$: Gesamthäufigkeit des Terms k

D^k: Anzahl Dokumente, in denen Term k vorkommt *(document frequency)*

5.2.2.1 Inverse Dokumentenhäufigkeit

Das Prinzip der *inversen Dokumentenhäufigkeit (inverse document frequency)* wird sehr häufig verwendet. Man geht davon aus, dass ein Term k wichtig ist, wenn seine Gesamthäufigkeit (F^k) in der Sammlung klein, seine Häufigkeit in einzelnen Dokumenten i (f_i^k) jedoch gross ist. So hat er eine grosse Unterscheidungskraft zwischen den Dokumenten und ist für das einzelne Dokument aufgrund seines dortigen häufigen Vorkommens ein guter Beschreibungsterm. Wie bekommt man nun einen solchen Wert? Weder die Dokumentenhäufigkeit D^k noch die Gesamthäufigkeit F^k allein sind aussagekräftig genug. Einen brauchbaren Wert liefert die Formel für die *inverse Dokumentenhäufigkeit* IDF:

schen Generierung von Thesauri. Man versucht mit Hilfe der Gewichte-Matrix Synonyme zu finden. Dabei wird angenommen, dass in sehr ähnlichen Dokumenten synonyme Begriffe bis zu einem gewissen Grad die gleiche Wichtigkeit haben. Die automatische Erzeugung von Thesauri liefert jedoch noch keine wirklich befriedigenden Resultate [Burger 1984, S. 10f.].

[26] In den Abschnitten über die Berechnung der Termgewichte folge ich in grossen Teilen den Ausführungen von [Grossmann 1984a, S. 2-9] und [Salton und McGill 1987, S. 65-76].

$$IDF^k = \log_2(\frac{n}{D^k}) + 1 = \log_2 n - \log_2 D^k + 1$$

Kommt nun ein Term k in jedem Dokument vor, das heisst $D^k = n$, so ist die inverse Dokumentenhäufigkeit $IDF^k = 1$. Würde er nur in jedem zehnten Dokument vorkommen, würde folgendes gelten:

$$10D^k = n \Longrightarrow IDF^k = \log_2(10) + 1 \approx 4.32$$

Der Schwerpunkt der inversen Dokumentenhäufigkeit liegt auf den Termen, die eine niedrige Dokumentenhäufigkeit aufweisen. Jetzt muss noch in die schlussendliche Gewichtung miteinbezogen werden, dass das Gewicht eines Terms k in Bezug zu einem bestimmten Dokument i grösser werden soll, wenn der Term in diesem Dokument häufig vorkommt, also f_i^k gross ist. Gleichzeitig sollte das Gewicht bei einem Anstieg der Dokumentenhäufigkeit D^k kleiner werden. Ein solches Gewicht G_i^k kann erreicht werden, indem die inverse Dokumentenhäufigkeit mit der Häufigkeit des Terms k im Dokument i multipliziert wird:

$$G_i^k = f_i^k IDF^k$$

Diese Verfahren ist relativ einfach und schnell zu implementieren und bringt gute Resultate.

5.2.2.2 Signal und Rauschen

Eine weitere Methode zur Berechnung von Termgewichten benützt als Grundlage die von Shannon 1948 entwickelte Informationstheorie[27]. Je grösser die Wahrscheinlich-

[27]Sein Werk „The Mathematical Theory of Communication“ ist unter `http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html` zu finden.

keit für das Auftreten eines Terms ist, desto kleiner ist die gewonnene Information, wenn er auftritt. Häufige Terme kann man leicht vorhersagen und nach ihrem Auftreten weiss man nicht viel mehr als vorher. Je kleiner jedoch die Auftrittswahrscheinlichkeit eines Terms ist, desto mehr Information bringt er mit seinem Auftauchen. Seltene Terme erwartet man nicht, kann sie kaum vorhersagen, aber wenn sie dann auftreten, verändern sie den Wissensstand erheblich.
Die Wahrscheinlichkeit p des Auftretens für den Term k an beliebiger Stelle im Dokument i bestimmt sich aus:

$$p_i^k = \frac{f_i^k}{w} \quad (w = \text{Anzahl Wörter im Dokument } i)$$

Den Informationsgehalt I, den das Auftreten des Terms k bringt, definiert man als:

$$I^k = -\log_2 p_i^k$$

Den mittleren Informationsgehalt $I_{average}$ eines Terms in einem Dokument i, das durch m Terme beschrieben wird, kann man durch folgende Formel errechnen:

$$I_{average} = -\sum_{k=1}^{m} p_i^k \log_2 p_i^k$$

Analog zu dieser Definition des mittleren Informationsgehaltes $I_{average}$ aus der Informationstheorie definiert man eine Grösse, die die Verteilung eines Terms in der ganzen Sammlung misst. Diese Grösse nennt man *Rauschen* (*noise*)[28]. Das Rauschen N eines Terms k wird folgendermassen definiert:

$$N^k = -\sum_{i=1}^{n} \frac{f_i^k}{F^k} \log_2 \frac{f_i^k}{F^k} \quad \text{, wobei} \quad \sum_{i=1}^{n} \frac{f_i^k}{F^k} = 1$$

[28]Wird zum Teil auch *Ballast* genannt.

Ist ein Term k völlig gleichmässig in der Sammlung verteilt, zum Beispiel wenn er genau t-mal in jedem Dokument vorkommt, ist das Rauschen maximal (N^k_{max}):

$$\begin{aligned} f^k_i &= t \quad \text{für alle Dokumente } i \\ F^k &= nt \\ N^k_{max} &= -\sum_{i=1}^{n} \frac{t}{nt} \log_2 \frac{t}{nt} \\ &= -\frac{n}{n} \log_2 \frac{1}{n} \\ &= \log_2 n \end{aligned}$$

Kommt ein Term k nur in einem Dokument i einer Sammlung t-mal vor, ist sein Rauschen minimal (N^k_{min}), also gleich Null:

$$\begin{aligned} f^k_j &= \begin{cases} t & \text{für } j = i \\ 0 & \text{sonst} \end{cases} \\ F^k &= t \\ N^k_{min} &= -\frac{t}{t} \log_2 \frac{t}{t} \\ &= -\log_2 1 \\ &= 0 \end{aligned}$$

Dieser Term besitzt dann eine gute Unterscheidungskraft. Allgemeine, nichtspezifische Begriffe tendieren zu einer gleichmässigeren Verteilung in den Dokumenten als spezifische. Je grösser das Rauschen eines Terms ist, desto gleichmässiger ist er verteilt und desto ungeeigneter ist er als Deskriptor. Man definiert daher eine

inverse Funktion, das sogenannte *Signal* S (*signal*) für einen Term k[29]:

$$S^k = \log_2 F^k - N^k$$

Das Signal ist ein Mass für die Konzentration eines Terms in einem Dokument. Je grösser das Rauschen, desto kleiner wird das Signal. Steigt die Gesamthäufigkeit, ohne dass das Rauschen grösser wird, bedeutet das, dass der Term in wenigen Dokumenten häufig vorkommt, und das Signal wird grösser. Wenn der Term in jedem Dokument genau einmal vorkommt (das Rauschen ist dann maximal), nimmt das Signal den Wert Null an, da $F^k = n$ ist.
Oftmals betrachtet man auch das Verhältnis von Signal und Rauschen S^k/N^k, da das Signal bei einer grossen Gesamthäufigkeit F^k des Terms k auch bei starkem Rauschen (also auch wenn er in vielen Dokumenten vorkommt) noch relativ gross ist. Der Wert lässt sich dann nicht von Fällen unterscheiden, in denen praktisch kein Rauschen auftritt, jedoch die Gesamthäufigkeit eher klein ist. Ein Beispiel aus [Grossmann 1984a, S. 6] soll dies verdeutlichen. Wieder soll der Term k in jedem Dokument t-mal vorkommen:

$$S^k = \log_2 nt - \log_2 n = \log_2 t$$

Kommt der Term k nur in einem Dokument t-mal vor, erhält man:

$$S^k = \log_2 t - 0 = \log_2 t$$

Wie man sieht, erhält das Signal S^k in beiden Formeln den gleichen Wert. Die grosse Gesamthäufigkeit in der ersten Formel ermöglicht trotz maximalem Rauschen einen gleich grossen Signalwert wie in der zweiten Formel, in der das Rauschen null

[29] Das Signal wird auch *Informationswert* genannt. Achtung: Informationswert ist nicht dasselbe wie Informationsgehalt.

ist. Zur Unterscheidung dieser Fälle empfiehlt sich daher, das Verhältnis S^k/N^k miteinzubeziehen.
Wie kann man nun mit dem Signalwert Terme gewichten? Man definiert eine ähnliche Funktion wie bei der inversen Dokumentenhäufigkeit:

$$G_i^k = f_i^k S^k$$

Die Verteilung der Terme in einer Sammlung kann man mit den Grössen Signal und Rauschen sehr genau erfassen. Trotzdem zeigen Experimente, dass sie keine optimalen Ergebnisse beim Retrieval liefern. Der Hauptgrund liegt darin, dass Terme mit niedriger Häufigkeit bei dieser Gewichtung überschätzt werden [Salton und McGill 1987, S. 75/79].

5.2.2.3 Diskriminanzwert

Die *Diskriminationsmethode* arbeitet nicht auf der Basis der relativen Häufigkeit, sondern mit einem Diskriminanzwert, der aufzeigen soll, wie gut ein Begriff ein Dokument von einem andern unterscheiden kann. Dafür brauchen wir einen Ähnlichkeitsfaktor $s(D_i, D_j)$ (*similarity*), der misst, wie gross die Ähnlichkeit zwischen zwei Dokumenten D_i und D_j ist. Bestimmt wird die Ähnlichkeit anhand der entsprechenden Dokumentvektoren $\vec{d_i}$ und $\vec{d_j}$. Ein Ähnlichkeitsmass s sollte folgende Eigenschaften haben:

1. s nimmt Werte von 0 bis 1 an, d.h. $0 \leq s(D_i, D_j) \leq 1$
2. s ist symmetrisch, d.h. $s(D_i, D_j) = s(D_j, D_i)$
3. $s(D_i, D_i) = 1$

Wenn keinerlei Übereinstimmung herrscht, ergibt das Mass den Wert 0, bei zwei identischen Dokumenten den Wert 1 und bei teilweiser Übereinstimmung Werte

dazwischen. Ein typisches Beispiel für ein solches Ähnlichkeitsmass[30] ist das Cosinus-Mass. Man geht dabei davon aus, dass sich zwei Dokumente umso ähnlicher sind, je kleiner der Winkel zwischen ihren Dokumentvektoren ist. Allgemein gilt, dass der Cosinus des Winkels φ zwischen zwei Vektoren $\vec{u} = (u_1, u_2, \ldots, u_m)$ und $\vec{v} = (v_1, v_2, \ldots, v_m)$ im m-dimensionalen Raum gegeben ist durch:

$$cos(\varphi) = \frac{\vec{u} \cdot \vec{v}}{|\vec{u}||\vec{v}|} = \frac{\sum_{k=1}^{m} (u_k v_k)}{\sqrt{\sum_{k=1}^{m} (u_k)^2} \sqrt{\sum_{k=1}^{m} (v_k)^2}}$$

Aus dieser allgemeinen Formel ergibt sich nun direkt die Definition des Cosinus-Masses ($s_{cosinus}$):

$$s_{cosinus}(D_i, D_j) = \frac{\vec{d_i} \cdot \vec{d_j}}{|\vec{d_i}||\vec{d_j}|} = \frac{\sum_{k=1}^{m} (G_i^k G_j^k)}{\sqrt{\sum_{k=1}^{m} (G_i^k)^2 \sum_{k=1}^{m} (G_j^k)^2}}$$

Mit Hilfe eines Ähnlichkeitsmasses sind wir nun in der Lage, die durchschnittliche Ähnlichkeit $s_{average}$ aller Dokumente zu berechnen. Diese Durchschnittsähnlichkeit stellt eine Art Dichte des Dokumentenraumes dar. Je näher, also je ähnlicher sich die Dokumente sind, desto grösser ist ihre Dichte.

$$s_{average} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} s(D_i, D_j)$$

Die Berechnung der Durchschnittsähnlichkeit ist auf diese Weise jedoch sehr aufwendig, da jedes Dokument mit jedem verglichen werden müsste. Um effizienter vorzugehen, wird meist ein künstliches mittleres Dokument $\overline{D} = (d_1, d_2, d_3, \ldots, d_n)$

[30] Es gibt verschiedene Ähnlichkeitsmasse, andere Möglichkeiten wären: Das Dice'sche Mass und das Jaccard'sche Mass. Wie diese Masse mathematisch definiert sind, kann man in [Grossmann 1984a, S. 6ff.] nachlesen.

als Vektor definiert. Ein solches durchschnittliches Dokument nennt man *Zentroid* und es enthält alle Begriffe mit ihrer durchschnittlichen Häufigkeit d_k:

$$d_k = \frac{1}{n} F^k$$

Durch Vergleiche der Dokumente mit dem Zentroid kann nun die Dichte Q bestimmt werden, die gleichbedeutend mit der Durchschnittsähnlichkeit $s_{average}$ ist, aber einfach mit Hilfe des Zentroids errechnet wird. Es werden die Summen der Ähnlichkeiten der Dokumente mit dem Zentroid berechnet, geteilt durch die Anzahl Dokumente, um den Durchschnitt zu erhalten:

$$Q = \frac{1}{n} \sum_{i=1}^{n} s(\overline{D}, D_i)$$

Lässt man nun einen bestimmten Term k weg, verändert sich die Dichte Q. Ist k ein häufiger Term mit einer gleichmässigen Verteilung in der Sammlung, verringert sich nach der Entfernung die Dichte, die Repräsentationen der Dokumente werden sich also unähnlicher. Der Term ist folglich nicht als Deskriptor erwünscht, da er, wenn er benutzt wird, nicht hilft, die Dokumente voneinander zu unterscheiden, sondern die Dichte erhöht und die Beschreibungen der Dokumente einander ähnlicher macht. Wird jedoch ein weniger häufiger Term entfernt, der für einige Dokumente ein hohes Gewicht, für andere ein tiefes Gewicht hat, wird die Dichte erhöht. Wird dieser Term als Deskriptor benutzt, reduziert sich die Dichte des Dokumentenraumes. Der Term ist also sehr nützlich, um die Dokumente voneinander zu unterscheiden. Für jeden Term k kann der *Diskriminanzwert DV* (*discrimination value*) berechnet werden. Mit Q_k bezeichnen wir die Kompaktheit des Dokumentenraumes nach der Elimination des Terms k:

$$DV_k = Q_k - Q$$

Die Terme lassen sich aufgrund des Diskriminanzwertes in eine Rangfolge bringen. Sie können in drei grobe Kategorien eingeteilt werden:

- *Gute Deskriptoren* mit einem positiven DV_k. Werden sie verwendet, reduziert sich die Dichte des Dokumentenraumes und die Repräsentationen der Dokumente werden sich unähnlicher.
- *Indifferente Deskriptoren* mit einem nahe bei Null liegendem DV_k. Werden sie verwendet, ändert sich die Dichte nicht.
- *Schlechte Deskriptoren* mit einem negativen DV_k. Werden sie verwendet, wird die Dichte des Dokumentenraumes erhöht und die Repräsentationen der Dokumente einander ähnlicher.

Analog zu den Gewichtungsformeln der inversen Dokumentenhäufigkeit und des Signals, kann auch hier eine Gleichung formuliert werden:

$$G_i^k = f_i^k DV_k$$

5.2.3 Indexierungsablauf

Wie können nun diese statistischen Methoden in ein System eingebettet werden? Analog zu den schon in Abschnitt 5.1 genannten Schritten der Freitextsuche wird zuerst der Text in einzelne Wörter zerlegt. Die Stoppwörter werden ebenfalls eliminiert. Als nächstes müssen die übriggebliebenen Wörter normalisiert werden, das heisst, die verschiedenen Flexions- und Derivationsformen eines Wortes müssen auf den Wortstamm reduziert werden. So sollen die Worte „analysis“, „analyse“, „analyser“, „analysed“, „analysing“ etc. alle auf den Wortstamm „analy“ zurückgeführt werden. Dies erreicht man durch graphematische und lexikalische Verfahren des computerlinguistischen Ansatzes. Für jeden Wortstamm wird dann für jedes Dokument das Gewicht mit einer Gewichtungsfunktion (zum Beispiel mit der inversen Dokumentenhäufigkeit, dem Signal oder dem Diskriminanzwert) berechnet. Nachdem

die Gewichte zugeteilt sind, können diejenigen Terme, die ein zu kleines Gewicht aufweisen, eliminiert werden.
Die Elimination von Termen ist jedoch nicht ganz ungefährlich. Das Entfernen von hochfrequenten Begriffen kann zu Verlusten bei der Abdeckung führen, während die Entfernung von niedrigfrequenten Begriffen Präzision und Abdeckung sinken lassen kann, da die Indexierungsgenauigkeit verringert wird. [Salton und McGill 1987, S. 81] schlagen vor, vor allem niedrigfrequente Begriffe daher nicht einfach zu löschen, sondern zu verbessern. Der Gebrauch eines Thesaurus könnte hier zum Beispiel Abhilfe schaffen. Für weitere Informationen sei auf [Salton und McGill 1987, S. 81-96] verwiesen.

5.3 Weiterführende Methoden

Neben diesen drei traditionellen Ansätzen gibt es neuere Methoden, die ich nur kurz erwähnen möchte. Viele neue Ideen stammen aus der Computerlinguistik. Man versucht, wichtige sprachliche Konzepte zu erkennen und so die Bestimmung geeigneter Deskriptoren zu vereinfachen. Damit will man das grundsätzliche Problem der traditionellen Methoden lösen, die keinerlei Versuche unternehmen, die Texte zu verstehen. Auch die Anfrage des Benutzers soll in natürlicher Sprache formuliert werden können, um ihm die Möglichkeit zu geben, sein Informationsbedürfnis besser beschreiben zu können. Die Anfrage soll dann natürlich ebenfalls mit computerlinguistischen Methoden analysiert werden.
Ein Grundproblem des Information Retrieval ist die Darstellung des Inhalts der jeweiligen Informationseinheiten. Vor allem für multimediale Objekte hat man noch keine befriedigende Lösung gefunden. Ein Ansatz besteht darin, Prädikatenlogik anzuwenden. Durch die erweiterten Beschreibungsmöglichkeiten will man die Probleme der begrenzten Repräsentation der Semantik besser lösen[31].
Ein anderer Ansatz hat seine Wurzeln in der Wahrscheinlichkeits- und Entschei-

[31] [Fuhr 1998] stellt zwei Ansätze vor: Die *terminological logic* und *Datalog*.

dungstheorie, das sogenannte probabilistische Modell. Probabilistische Modelle versuchen, das Problem der Unsicherheit der Textrepräsentation besser in den Griff zu bekommen. Es wird ein Mass für die Bedeutsamkeit eines Begriffs verwendet, das der Wahrscheinlichkeit entspricht, mit der dieser Begriff in einer bestimmten Umgebung vorkommt, oder mit der dieser Begriff einem Dokument als Deskriptor zugeordnet werden sollte. Dieses Mass lässt sich als Deskriptorgewicht benutzen, so dass sich die Dokumente in eine Rangfolge bringen lassen, die der Wahrscheinlichkeit der Relevanz bezüglich bestimmter Suchanfragen entspricht [Weiss 1997][32].
Neuere Ansätze versuchen nur Passagen oder gar nur die Antwort auf die Anfrage und nicht ganze Dokumente zurückzuliefern[33]. So muss der Benutzer nicht ein ganzes Dokument durchsehen, bis er die ihn interessierende Stelle findet, die dann unter Umständen nicht einmal relevant für ihn ist.

[32]Für weitere Informationen empfehle ich [Fuhr 1998, S. 99-137].
[33]Begriffe wie Passagenretrieval und Information Extraction kommen dabei ins Spiel.

6 Ausblick

Die vorgestellten klassischen Techniken des Indexierens sind auch heute noch aktuell. Besonders die Klassifikationen und Thesauri sind oft verwendete nützliche Werkzeuge. Die stetig wachsende Datenmenge lässt den Ruf nach effizienter automatischer Indexierung jedoch immer lauter werden, um den zeitlichen und finanziellen Aufwand zu minimieren.

In den letzten zehn Jahren hat sich das Gebiet des Information Retrieval sehr verändert. Entwicklungen in der Technik haben die Leistungsfähigkeit der Hardware stark erhöht und immer mehr Texte sind elektronisch erfasst. Dadurch ergeben sich neue Anwendungsbereiche wie die Dokumentation von technischen Handbüchern und Büroinformationssysteme. Aber vor allem die Einführung des *World Wide Web* brachte grosse Veränderungen. Das Forschungsfeld ist nunmehr nicht nur für Bibliothekare und Informationswissenschaftler von Interesse. Jeder kann seine eigenen Ideen und Dokumente aufs Netz stellen, ohne grosse Kosten und Anstrengung. Das Web wächst enorm schnell und ist ständig in Bewegung. Neues kommt hinzu, anderes verschwindet wieder. Leider sind die Informationseinheiten im Netz inhaltlich, wie auch in Bezug auf Format, Struktur und Länge völlig ungeordnet. Das bleibt nicht ohne Folgen. Brauchbare Informationen im Netz zu finden, ist oftmals eine sehr langwierige und ermüdende Arbeit. Ausserdem weiss man nie genau, ob die gefundenen Informationen richtig und vollständig sind. Die Aufgabe, diese Suche effizienter zu gestalten, lenkte die Aufmerksamkeit vieler Forscher auf das Gebiet des IR. Neue, auf das Web abgestimmte Techniken wurden und werden noch benötigt. Nicht nur Text, sondern vor allem auch Bilder und andere multimediale Objekte müssen verarbeitet werden. Das Hauptaugenmerk für die Zukunft ist auf drei Aufgaben gerichtet: Auf die Verbesserung der Qualität des Retrieval-Ergebnisses, auf die Maximierung der Antwortgeschwindigkeit und auf die Optimierung der Schnittstellen zum Benutzer, der mit seiner Interaktion die Qualität der Ergebnisse stark beeinflusst.

Das heutige Information Retrieval ist sehr komplex und vielfältig. Es geht nicht mehr nur um die Realisierung des eigentlichen Retrieval-Prozesses, sondern vor allem auch die Wissensverarbeitung beim Menschen muss berücksichtigt werden. Der Benutzer und die grafische Darstellung rücken immer mehr in den Vordergrund. Es ist ein Forschungsgebiet geworden, dass viele Disziplinen in sich vereinigt. Nicht nur die Informationswissenschaft, sondern auch die Mathematik, die Logik, die Linguistik, die Psychologie und viele andere Fachrichtungen sind wichtige Komponenten geworden. Die Zukunft des Information Retrieval scheint in der Entwicklung von Techniken zu liegen, die durch das Zusammenarbeiten der verschiedenen Disziplinen und durch Kombination ihrer Methoden entstehen.

Index

Literatur

[Baeza-Yates und Ribeiro-Neto] Baeza-Yates, Ricardo und B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, New York.

[Burger 1984] Burger, Cyrill (1984). *Moderne Konzepte - SMART*. Institut für Informatik der Universität Zürich. Fortbildungsseminare in Wirtschaftsinformatik, Seminar vom 23. Oktober Hotel Zürich, Zürich.

[Chowdhury 1999] Chowdhury, G.G. (1999). *Introduction to Modern Information Retrieval*. Library Association Publishing, London.

[Fuhr 1998] Fuhr, Norbert (1998). *Information Retrieval*. Skriptum zur Vorlesung. URL: `http://ls6-www.informatik.uni-dortmund.de/ir/teaching/courses/ir/`.

[Grossmann 1984a] Grossmann, Jürg (1984a). *Automatisches Indexieren. Theorie und Methoden*. Institut für Informatik der Universität Zürich. Fortbildungsseminare in Wirtschaftsinformatik, Seminar vom 23. Oktober Hotel Zürich, Zürich.

[Grossmann 1984b] Grossmann, Jürg (1984b). *Information Retrieval (IR). Grunbegriffe und Einführung*. Institut für Informatik der Universität Zürich. Fortbildungsseminare in Wirtschaftsinformatik, Seminar vom 23. Oktober Hotel Zürich, Zürich.

[Hess 1999] Hess, Michael (1999). *Einführung in die Computerlinguistik I*. Skript zur Vorlesung. URL: `http://www.ifi.unizh.ch/cl/study/unterlagen.html`.

[Linke et al. 1996] Linke, Angelika, M. Nussbaumer und P. R. Portmann (1996). *Studienbuch Linguistik*. Reihe Germanistische Linguistik; 121 Kollegbuch. Max Niemeyer Verlag, Tübingen, 3. unveränderte Aufl.

[Rechenberg und Pomberger 1999] Rechenberg, Peter und G. Pomberger, Hrsg. (1999). *Informatik-Handbuch*. Hanser, München; Wien, 2. aktualisierte und erw. Auflage Aufl.

[Salton und McGill 1987] Salton, Gerard und M. J. McGill (1987). *Information Retrieval - Grundlegendes für Informationswissenschaftler*. McGraw-Hill, Hamburg; New York.

[Sparck-Jones und Willet 1997] Sparck-Jones, Karen und P. Willet (1997). *Readings in Information Retrieval*. Morgan Kaufmann Publishers, San Francisco California.

[Weiss 1997] Weiss, Scott (1997). *Glossary for Information Retrieval*. URL: `http://www.cs.jhu.edu/~weiss/glossary.html`.