

Fachhochschule Stuttgart

Studiengang Informationswirtschaft

Wolframstrasse 32 – D-70191 Stuttgart
E-Mail: nohr@hbi-stuttgart.de



ARBEITSPAPIERE WISSENSMANAGEMENT
WORKING PAPERS KNOWLEDGE MANAGEMENT

Holger Nohr

**Automatische Dokumentindexierung –
Eine Basistechnologie für das
Wissensmanagement**

Arbeitspapiere Wissensmanagement
Nr. 2/2000

ISSN 1616-5349 (Internet)
ISSN 1616-5330 (Print)

Herausgeber:
Prof. Holger Nohr

Information

Reihe: Arbeitspapiere Wissensmanagement

Herausgeber: Prof. Holger Nohr
Fachhochschule Stuttgart
Studiengang Informationswirtschaft
Wolframstrasse 32
D-70191 Stuttgart
E-Mail: nohr@hbi-stuttgart.de
Homepage: <http://www.hbi-stuttgart.de/nohr>

Schriftleitung: Prof. Holger Nohr

ISSN: 1616-5349 (Internet); 1616-5330 (Print)

Ziele: Die Arbeitspapiere dieser Reihe sollen einen Überblick zu den Grundlagen des Wissensmanagements geben und sich mit speziellen Themenbereichen tiefergehend befassen. Ziel ist die verständliche Vermittlung theoretischer Grundlagen und deren Transfer in die Praxis.

Zielgruppen: Zielgruppen sind Forschende, Lehrende und Lernende im Fachgebiet Wissensmanagement sowie Praktiker in Unternehmen.

Quellen: Die Arbeitspapiere entstehen aus Forschungsarbeiten, Diplom-, Studien- und Projektarbeiten sowie Begleitmaterialien zur Lehr- und Vortragsveranstaltungen des Studiengangs Informationswirtschaft der Fachhochschule Stuttgart.

Hinweise: Falls Sie Arbeitspapiere in dieser Reihe veröffentlichen wollen, wenden Sie sich bitte an den Herausgeber.
Informationen über die Arbeitspapiere dieser Reihe finden Sie unter <http://www.hbi-stuttgart.de/nohr/Km/KmAP/KmAP.htm>

Inhaltsverzeichnis

1. EINLEITUNG.....	4
1.1 Ausgangslage.....	4
1.2 Indexierung.....	7
2. AUTOMATISCHE INDEXIERUNG	10
2.1 Grundsätze.....	10
2.2 Verfahrensübersicht.....	12
3. VERFAHREN DER AUTOMATISCHEN INDEXIERUNG	14
3.1 Statistische Verfahren.....	14
3.1.1 Beispiel einer einfachen statistischen Indexierung	17
3.2 Informationslinguistische Verfahren.....	22
3.2.1 Beispiel: Indexierungsverfahren im IZIS-ET.....	28
3.2.2 Beispiel: IDX – Ein wörterbuchbasiertes Verfahren	31
3.3 Pattern-Matching-Verfahren	35
3.4 Begriffsorientierte Verfahren	37
3.4.1 Beispiel: Das Verfahren AIR/X.....	38
3.4.2 Beispiel: GERHARD	42
4. KEYPHRASE EXTRACTION	44
4.1 Beispiel: NRC's Extractor.....	46
5. INDEXIERUNG UND RETRIEVALVERFAHREN	48
6. RESÜMEE.....	53
ANHANG.....	53
LITERATUR.....	54

1. Einleitung

1.1 Ausgangslage

Ein bedeutender Anteil der entscheidungsrelevanten Informationen liegt in Organisationen – in privatwirtschaftlichen Unternehmen, in Verbänden und auch in der öffentlichen Verwaltung – aufgezeichnet in unstrukturierten Dokumenten¹ vor, hauptsächlich in der Form von Texten. Schätzungen gehen von einem Anteil von rund 80% unstrukturierter Information in Unternehmen aus (Gerick 2000). Das Management dieser Form der Information stellt alle Organisationen vor zunehmend grösser werdende Probleme². Wenn Information – und daran besteht heute kaum mehr ein Zweifel – der entscheidende Wettbewerbsfaktor auf allen Märkten und in allen Branchen ist, bedarf es allerdings eines gezielten Managements gerade auch dieser Form der Information.

Wenn auch die Vorstellung von der papierlosen Organisation auf absehbare Zeit Utopie bleiben wird, so nimmt die Menge der elektronisch verfügbaren Textdokumente seit geraumer Zeit doch erheblich zu, bspw.:

- Elektronische Mails von Kunden, Lieferanten usw. erhalten Unternehmen spätestens mit ihrem Auftritt im Internet in grosser und rasch zunehmender Anzahl.
- Memos und andere Mitteilungen werden innerbetrieblich auf elektronischem Wege ausgetauscht.
- Wirtschafts- und Finanznachrichten gelangen über Online-Ticker in das Unternehmen (bspw. über die Angebote von Reuters, Bloomberg u.a.).
- Diverse Dokumente können aus dem Internet abgerufen werden. In grossem Umfang etwa technische Forschungsberichte (Preprints) oder Patente aber auch Produktbeschreibungen.
- Elektronische Zeitschriften und Newsletter stehen mit Volltextarchiven im Internet zur Verfügung.
- Dokumentenlieferdienste bieten elektronische Lieferungen an.
- Marktstudien, Produktspezifikationen, Technische Dokumentationen oder Projektberichte werden innerbetrieblich elektronisch erstellt bzw. in dieser Form von Anbietern bezogen.
- Wissensmanagement-Systeme (bspw. „Corporate Memories“, „Organizational Memories“, „Knowledge Repositories“ oder „Gruppendächtnisse“) enthalten Informationsobjekte, die eindeutig identifiziert und damit wiederauffindbar gemacht werden müssen (Klosterberg 1999, Nohr 1999b, Lehner 2000).

Diese und andere Dokumente enthalten eine Vielzahl wichtiger Informationen für alle Bereiche und Entscheidungen innerhalb eines Unternehmens.³ Die Masse der anfallenden Dokumente⁴ macht eine manuelle Auswertung oft jedoch unmöglich oder zumindest doch unzureichend und teuer. Es muss zudem Sorge dafür getragen werden, benötigte Informationen jederzeit, rasch und zielsicher auffinden zu können. Um gezielt nach Dokumenten mit bestimmten Inhalten suchen

¹ Diese Dokumente können neben Text auch andere Darstellungsformen, bspw. Graphiken, Bilder usw. enthalten. Für eine automatische Indexierung ist wenigstens z.Zt. noch das Vorhandensein von Text notwendig.

² Eine Übersicht über das Management unstrukturierter Informationen bieten Königer/Reithmayer (1998).

³ Auf die Bedeutung des Faktors Information bzw. Wissen muss an dieser Stelle nicht näher eingegangen werden, vgl. dazu die aktuelle Literatur unter dem Schlagwort Wissensmanagement.

⁴ Grosse Unternehmen erhalten heute z.T. mehr als 40.000 Emails in der Woche über den entsprechenden Kontakt auf ihrer Web-Site.

zu können, muss eine effektive Inhaltserkennung und –kennzeichnung erfolgen. Das bedeutet eine (automatische) Zuordnung der Dokumente zu bestimmten Themengebieten, um eine schnelle Auswahl zu ermöglichen. Zu diesem Zweck müssen die Dokumenten*inhalte* in einem Informationssystem repräsentiert, d.h. indexiert oder klassifiziert, werden.

Eine Lösung dieser inhaltsbezogenen Analyseaufgabe kann zudem helfen, den Informations- und Dokumentenfluss im Unternehmen zu steuern (Martin 1998). Eine entsprechende Dokumentenanalyse kann bspw. anhand erkannter Themenbezüge eingehende Post automatisch an den richtigen Empfänger, also etwa den Sachbearbeiter oder die zuständige Fachabteilung (bspw. Rechnungen in die Buchhaltung), im Unternehmen weiterleiten. Eine solche Aufgabe versucht bspw. das Softwaremodul GENIUS des Dokumenten-Managementsystem-Herstellers EASY über die Erkennung typischer Textsorten (Rechnung, Auftrag usw.) zu meistern. Automatische Analysen von Dokumenteninhalten werden zudem benötigt, wenn Mitarbeiter über einen persönlichen Informationsfilter relevante neue Dokumente aus Dokumenten- oder Wissensmanagement-Systemen automatisch zugestellt bekommen sollen (Foltz/Dumais 1992). Eine automatische und themenorientierte Variante des Push-Prinzips.

Neben der originär elektronischen Form, können heute Papierdokumente schnell und sicher in eine elektronische Form überführt werden. Dabei werden die Dokumente zumeist als Image im Dokumentenmanagement-System archiviert (z.B. auf Optical Disks) aber wenigstens temporär per OCR-Verfahren einer Indizierung zugeführt (Thiel 1994; Riggert 1998).

Damit ist heute eine nahezu komplette Dokumentenverwaltung in elektronischer Form möglich. Mit dieser Situation werden in Unternehmen Verfahren des Information Retrieval (Salton/McGill 1987, Forst 1999, Gerick 2000) relevant, die bislang auf die klassische Literaturdokumentation mit ihren bibliographischen Fachdatenbanken bzw. Volltextarchiven oder neuerdings auf die Informationsgewinnung aus dem Internet (Suchmaschinen) beschränkt waren. Durch die Einbindung von „Wissensdatenbanken“ in das Wissensmanagement der Unternehmen ist ein weiteres neues Anwendungsfeld für Verfahren des Information Retrieval entstanden und Information Retrieval zu einer Basistechnik für das betriebliche Wissensmanagement geworden.

Die oben angedeuteten Anforderungen an die Bearbeitung elektronischer Dokumente lassen zunehmend automatische Verfahren in den Blickpunkt einer breiteren Anwenderschaft in Unternehmen, Verbänden und öffentlichen Verwaltungen rücken.

In diesen Gebieten des Dokumenten-Managements sind in den nächsten Jahren die wichtigsten, den Markt treibende Innovationen zu erwarten. Hier bahnt sich eine vollständige Inversion der Rolle des Menschen beim Einsatz solcher Systeme an. (Kampffmeyer 1999, S. 12)

Auf dem Markt der Dokumenten-Managementsysteme zeichnen sich – diesem Trend folgend – strategische Unternehmenszusammenschlüsse ab. So kaufte, um nur ein Beispiel zu nennen, die EASY Software AG, einer der grossen Hersteller von DM-Systemen, Ende 1999 mit dem Softwarehaus ZERES GmbH einen Spezialisten für automatische Dokumentklassifizierungen und –indexierungen. Unter dem bereits erwähnten Namen EASY GENIUS wird das neue Produkt zur automatischen Erschließung vertrieben.

Bei der Anwendung automatischer Verfahren lassen sich grundsätzlich zwei Ansätze unterscheiden, die teilweise jedoch auf ähnlichen methodischen Grundlagen – der Extraktion von Schlüsselwörtern oder Schlüsselsätzen (Turney 1997; Sparck Jones 1999) – beruhen:

- die automatische Zusammenfassung von Textdokumenten (Automatic Text Summarization) und

■ die automatische Indexierung von Textdokumenten (Automatic Indexing)

Dieser Beitrag wird sich konzentrieren auf die automatische Indexierung als einem Teilgebiet des Information Retrieval. Im Abschnitt 4 – Keyphrase Extraction – befinden wir uns jedoch bereits auf einer Schnittstelle zur automatischen Zusammenfassung von Texten.

Gegenstand des Information Retrieval (IR) ist die Repräsentation, Speicherung und Organisation von Informationen und der Zugriff zu Informationen. (Salton/McGill 1987, S. 1)

Damit sind ganzheitliche Lösungen angesprochen, die aufeinander abgestimmte Verfahren der Indexierung (der Repräsentation), der Speicherorganisation und der Recherche (des Zugriffs) nach Informationen anstreben.

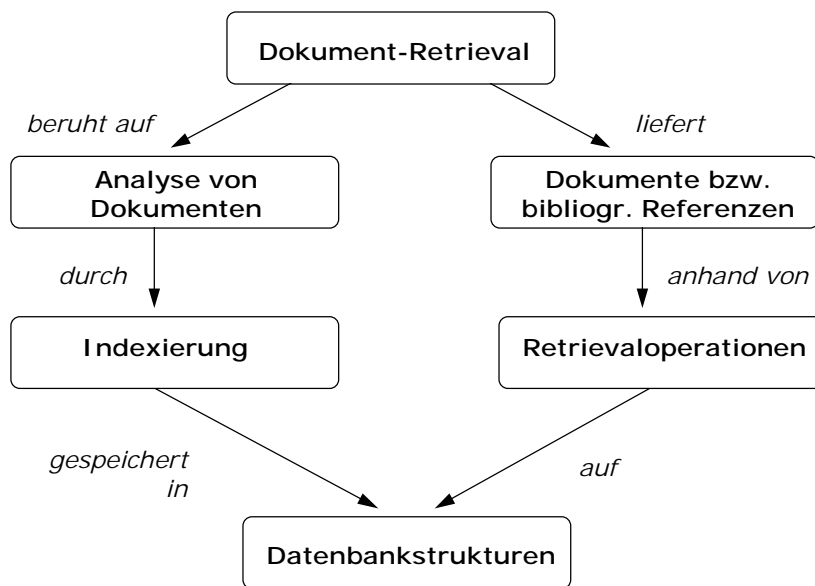


Abb. 1: Grundmodell des Dokumenten-Retrieval nach Fuhr (1997, S. 19)

Information Retrieval zeichnet sich aus durch vage Anfragen und unsicheres Wissen (Knorz 1995). Die Vagheit der Anfragesituation ergibt sich durch nicht eindeutig definierbare Antworten. Insbesondere sind die Beurteilungen erzielter Retrievalergebnisse abhängig von individuellen Relevanzbewertungen, die jeweils abhängig sind von der Tiefe des Informationsbedürfnisses und den Vorkenntnissen. Wenn bspw. Informationen über den *Maschinenbaumarkt in Estland* gesucht werden, ist eine Studie über den *Markt für Werkzeug- und Verarbeitungsmaschinen im Baltikum* eine relevante Antwort? Falls diese Antwort als relevant eingestuft wird, wäre es möglich durch eine entsprechende Reformulierung der Eingangsfrage weitere Dokumente zu erhalten, die dem ersten ähnlich sind. Information Retrieval ist also auch ein iterativer Prozess, der häufig erst im Dialog und in Abhängigkeit von der Bewertung der bisher erhaltenen Systemantworten zu einem gewünschten oder näherungsweise optimalen Resultat führt.

Resultate der Forschungsarbeit auf dem Gebiet des Information Retrieval gehen nur langsam in praktische Anwendungen ein. Erst mit Aufkommen der Internet-Suchmaschinen wurden und werden viele Erkenntnisse der Forschung auf breiter Basis eingesetzt und nun auch in kommerzielle Datenbank- und Retrieval-Systeme integriert (bspw. in Oracle, Fulcrum oder freeWAIS).

Der vorliegende Beitrag beschäftigt sich grundlegend mit dem Eingangsschritt des Information Retrieval, der Indexierung von Dokumenten. Da wir eine weitgehend elektronische Verfügbarkeit

der Dokumente annehmen und von manuell nicht beherrschbaren Dokumentenmengen ausgehen, werden automatische Verfahren der Indexierung behandelt.

1.2 Indexierung

Indexierung gilt klassisch als ein manuelles und intellektuelles Verfahren der inhaltlichen Dokumenterschliessung. So verstanden umfasst das Indexieren die Arbeitsschritte:

1. Das begriffliche Erfassen des Inhalts eines vorliegenden Dokuments (die Inhaltsanalyse),
2. Die Repräsentation dieses Inhalts durch die sprachlichen Elemente einer Indexierungssprache (bspw. durch Deskriptoren eines Thesaurus oder Codierungselementen eines Klassifikationssystems).

Dieser intellektuelle Vorgang der Dokumenterschliessung bedingt eine Inhaltsanalyse, d.h. ein Verstehen des zu erschliessenden Dokuments (Nohr 1999a). Diese Form der inhaltlichen Dokumenterschliessung ist heute weit verbreitet und unumgänglich dort, wo Texte nicht-elektronisch vorliegen oder Dokumente ganz oder überwiegend aus nicht-sprachlichen Darstellungsformen (Bilder, Fotos usw.) bestehen.

Indexieren erfüllt den Zweck einer inhaltlichen Repräsentation von Dokumenten mit dem Ziel, diese im Zuge eines Information Retrieval unter entsprechenden Deskriptoren suchbar zu machen. Die Deskriptoren werden in der Regel einem normierten Vokabular (Indexierungs- oder Dokumentationssprache) entnommen. Dieses Vorgehen war historisch notwendig, da in der frühen Literaturdokumentation oder im Bibliothekswesen lediglich bibliographische Angaben eines Dokuments – evtl. mit Abstracts – gespeichert werden konnten, keinesfalls jedoch der Volltext. Der intellektuelle Ansatz verfolgt eine *begriffliche* Aufschliessung des Dokumenteninhalts, d.h. er verlässt die sprachliche (Oberflächen-)Ebene eines Textes und möchte stattdessen seine Bedeutungskomponente repräsentieren.

Maschineneinsatz, wachsende Speicherkapazitäten und eine stetige Zunahme elektronischer Texte führten zum Ansatz der automatischen Indexierungsverfahren. Der Beginn dieser Automatisierung wird mit den Arbeiten H.P. Luhn's (1958) verbunden. Im Gegensatz zu den intellektuellen Verfahren arbeiten alle automatischen Indexierungsmethoden mehr oder weniger mit der sprachlichen Oberfläche von Dokumenten – mit Termen (Zeichenketten) und nicht bzw. zunächst nicht mit ihrer Bedeutung.

Seither stehen intellektuelle und automatische Verfahren in Konkurrenz zueinander. Dieses Konkurrenzverhältnis scheint unauflösbar, da beiden Verfahren grundsätzlich unterschiedliche Ausgangspositionen und Grundannahmen zugrunde liegen (vgl. auch Nohr 2000):

Intellektuelle Verfahren

streben die korrekte und konsistente Repräsentation von Dokumenteninhalten (der Bedeutungsebene) an, indem sie behandelte Gegenstände durch normierte Benennungen einer Indexierungssprache (Deskriptoren) wiedergeben.

Automatische Verfahren

wollen vorliegende Dokumente in einer Weise aufbereiten, dass sie für anschliessende Retrievalfragen über Indexterme eine bestmögliche Wiederauffindbarkeit herstellen.

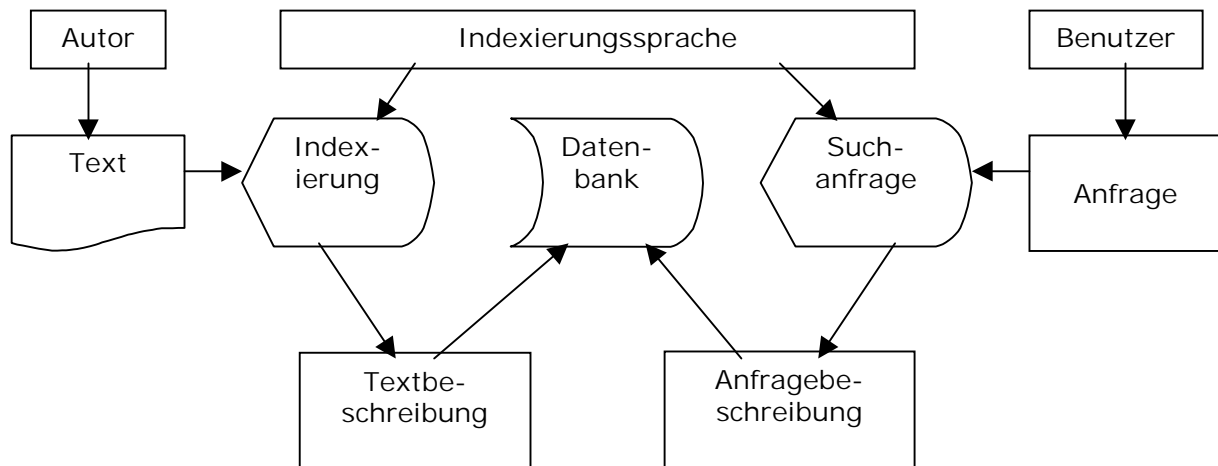


Abb. 2: Modell der Informationsschliessung

Kritiker der Automatisierung führen gewöhnlich die sprachliche Vielfalt und den Variantenreichtum sprachlicher Produktionen ins Feld. Diese seien durch maschinelle Verfahren unmöglich erkennbar oder gar „verstehbar“. Nicht jedes angesprochene Thema könne daher korrekt und konsistent repräsentiert werden. Automatischer Indexierung liege die irrige Annahme zugrunde, es genüge aus den Termen der Dokumentenvorlage eine Auswahl für die Generierung eines Index zu treffen. Eine Rückführung sprachlicher Ausdrücke auf ihren Bedeutungsgehalt unterbliebe ebenso wie die Lexikalisierung von Paraphrasen oder die Ausweisung von begrifflichen Relationen. Kurz, inhaltliche Dokumenterschliessung sei kein formalisierbarer Prozess und somit nicht geeignet für eine Automatisierung.

Anhänger der automatischen Verfahren verweisen dagegen auf die Unmöglichkeit der intellektuellen Behandlung heutiger Dokumentenfluten und versuchen darüber hinaus, durch empirische Untersuchungen (Retrievaltests) zu belegen, dass automatische Verfahren verglichen mit intellektuellen mindestens gleich gute Resultate bei der Wiederauffindung von Dokumenten erzielen. Dieser Beleg gelingt in einer grossen Zahl durchgeführter Tests. Da Retrievaltests jedoch auf einer Reihe subjektiver Parametrisierungen beruhen (bspw. der Festlegung eines Relevanzkriteriums für die aufgefundenen Dokumente), werden ihre Resultate vielfach angefochten (Sachse et al. 1998). Zudem lassen sich die Resultate von Retrievaltests nur schwer generalisieren. Bei der empirischen Untersuchung von Informationssystemen ist der Mensch als Informationsnachfrager ein – häufig kritischer – Bestandteil des Systems (Krause 1996).

Für die Bearbeitung grosser Dokumentenmengen bieten die Kritiker keine Lösungen an. Unseres Erachtens sind die Rahmenbedingungen in vielen Fällen zwingend für eine automatische Indexierung. Ohne Automatisierung lassen sich die Dokumentmengen häufig nicht bewältigen, d.h. intellektuelle Indexierung wäre nur unter Verzicht auf die inhaltliche Behandlung einer grösseren Teilmenge vorhandener Dokumente möglich. Zudem schlägt bei intellektueller Vergabe von Deskriptoren das Kostenargument zu Buche: Eine hohe Qualität intellektueller Indexierung wird nur durch den Einsatz entsprechend qualifizierten Fachpersonals erreicht. Daher fallen nach einer neuen Studie des IZ Sozialwissenschaften im Durchschnitt 22,- DM für die Indexierung eines Dokuments an (Krause/Mutschke 1999).

Eine Reihe von Nachteilen der intellektuellen Verfahren der Dokumenterschliessung hat kürzlich Hauer (2000, S. 203) zusammengefasst:

1. messbare Kosten bei Erstellung und Pflege von Terminologien (Nicht-Finden erzeugt keine messbaren Kosten);

2. schwerfälliges Handling von – noch immer oft – gedruckten Terminologien bei der Anwendung;
3. erhebliche Kosten bei der intellektuellen Indexierung;
4. menschliche Inkonsistenz bei der Indexierung mit der Folge Informationsverlust bei der Recherche;
5. ständiges Risiko der Veralterung eingesetzter Terminologien;
6. ständiges Risiko der Abspaltung von Anwendergruppen, da neue Themen und anderes Sprachverständnis (Sprache geht unter die Haut).

Dagegen stehen bspw. ebenso hohe Kosten für die kontinuierlich notwendige Pflege von Indexierungswörterbüchern bei vielen automatischen Verfahren oder auch die Erkenntnis, dass automatisch erzeugte Konsistenz bei der Indexierung noch keine korrekte Indexierung bedeutet. Ein Verfahren kann konsistent, aber eben auch konsistent fehlerhaft arbeiten. Indexierungskonsistenz ist damit nicht allein als Kriterium zulässig, wenn es um die Beurteilung eines Indexierungsverfahrens geht.

Ähnliche Diskussionen zwischen Befürwortern intellektueller bzw. automatischer Ansätze vollziehen sich auf dem Gebiet der Zusammenfassung von Texten (Summarization).

2. Automatische Indexierung

2.1 Grundsätze

Die automatische Indexierung ist hinsichtlich einer Reihe von Grundannahmen durchaus vergleichbar mit anderen Anwendungsfeldern der automatischen Sprachverarbeitung, wie bspw. der maschinellen Übersetzung (Zimmermann 1989; 1990; 1991) oder der sozialwissenschaftlichen Inhaltsanalyse (Mergenthaler 1996; Geis 1992). Die Prämisse einer jeglichen automatischen Sprachverarbeitung – und damit auch der automatischen Indexierung, gleich welchen Ansatzes – ist die Annahme einer mehr oder weniger starken Korrespondenz zwischen sprachlicher Repräsentation (der Sprachoberfläche) und der angestrebten Bedeutung (Korrespondenz- oder Abbildtheorie). Diese Annahme ist in hohem Masse kritisch zu bewerten, da eine solche Korrespondenz zwischen einer sprachlichen Ausdrucksweise und ihrer Bedeutung wenigstens auch auf Konvention beruht und nicht apriori gegeben ist. Konventionen aber werden heute zunehmend in Subsystemen aufgestellt (Fachgemeinschaften oder einzelnen Unternehmen, Communities of Practice usw.) In der modernen Sprachwissenschaft wird die Bedeutung eines Wortes durch den Gebrauch desselben im jeweiligen Kontext gesehen. (zur Diskussion der Korrespondenz- bzw. Abbildtheorie siehe Schmitz 1992, insbes. Kap. 2.3). So sieht Genzmer Wörter als Zeichen, die durch ihre Funktion definiert und damit Bedeutungsträger sind. Weiter führt er aus:

Es gibt aber keine andere Bedeutung als eine kontextabhängige.

(...)

Somit ist jedes Zeichen nur relativ definiert und nie absolut, immer jedoch durch seine Stellung im Kontext. Dadurch wird folglich so etwas wie eine eindeutige Lexikonbedeutung in Zweifel gezogen, denn es handelt sich dabei um nichts anderes als auf künstliche Weise isolierte und kontextfreie Bedeutungen. (Genzmer 1995, S. 14)

Automatische Verfahren der Inhaltserschließung beinhalten in einem ersten Schritt keine Inhaltsanalyse im Sinne eines Verstehens, wie sie bei intellektueller Indexierung notwendige Voraussetzung ist (Nohr 1999a). Die Analyse der Dokumente ist an statistischen bzw. linguistischen Kriterien orientiert oder folgt einer Mustererkennung. Diese Verfahren sind im Grunde Extraktionsmethoden, d.h. sie arbeiten ausschliesslich mit den sprachlichen Ausdrucksweisen, wie sie in der Dokumentvorlage gegeben sind. In den zum Zwecke des Wiederauffindens angelegten Indizes gehen nur Terme ein, die dem Dokument entnommen sind. Aus der Menge der Terme eines Dokuments wird eine Auswahl getroffen – es werden „gute“ Indexterme für die inhaltliche Repräsentation des Dokuments gesucht. Diese Auswahl kann anhand statistischer Methoden erfolgen (Abschnitt 3.1) oder durch den Abgleich mit elektronischen Wörterbüchern zustande kommen (Abschnitt 3.2). Eine weitere Methode die Auswahl zu treffen, ist die Mustererkennung (Abschnitt 3.3). Die extrahierten Terme werden meist einer mehr oder weniger weitgehenden Bearbeitung unterzogen, bspw. einer Rückführung auf die Grundform, den Wortstamm oder einer Dekomposition. Das Werkzeug dieser Bearbeitungen liefern die informationslinguistischen Verfahren. Eine Rückführung auf Bedeutungen wird nicht vorgenommen. So wird bei den genannten Verfahren bspw. nicht unterschieden, ob ein Textterm „Bank“ im Sinne eines Sitzmöbel oder eines Kreditinstituts gemeint ist. Die Repräsentation im Index bliebe in gleichen Fällen ununterscheidbar die selbe.

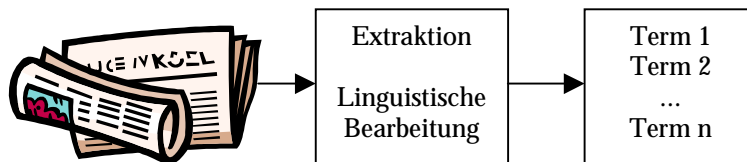


Abb. 3: Modell der Extraktionsverfahren

Verfahren, die eine begriffsorientierte Verarbeitung zum Ziel haben, wollen eine bedeutungsabhängige Repräsentation im Index erreichen, d.h. sie wollen nicht Terme sondern Bedeutungen ermitteln. Auch diese Verfahren leisten jedoch keine Inhaltsanalyse im eigentlichen Sinne. Vielmehr wird auch bei ihnen (bspw. aufgrund von Heuristiken oder Wahrscheinlichkeitswerten) vom Wortmaterial eines Textes (der Sprachoberfläche) geschlossen auf eine begriffliche Bedeutung (vgl. das Verfahren AIR, Abschnitt 3.4). Die Repräsentation erfolgt jedoch nicht durch extrahierte Terme. Vielmehr wird die erkannte Bedeutung abgebildet auf eine normierte Indexierungssprache (bspw. auf einen Thesaurus), deren Deskriptoren als Indexterme zugeteilt werden. Auch die Abbildung auf Klassen eines Klassifikationssystems ist möglich (vgl. das Verfahren GERHARD, Abschnitt 3.4), in diesem Falle werden Notationselemente zugewiesen.

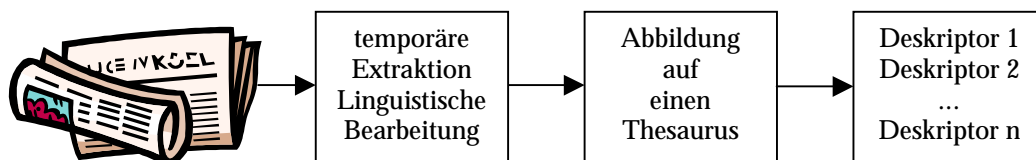


Abb. 4: Modell begriffsorientierter Verfahren

Umsetzbar in eine automatische Lösung sind Aufgaben die wohldefiniert und in ihren Abläufen und Bestandteilen gut beschreibbar sind. Ob der Vorgang des Indexierens (im Sinne einer begrifflichen Repräsentanz durch Deskriptoren) eine solche Aufgabe ist, bleibt wenigstens fraglich. Mater (1990) spricht vom Indexieren als einem „weitgehend irrationalen Vorgang“. Tatsächlich kann er Untersuchungsergebnisse zitieren, die belegen, dass mehrere Indexierer bei der Bearbeitung eines Textes weniger als 50% Übereinstimmung (Indexierungskonsistenz) erzielen. Diese Resultate lassen am begriffsorientierten Ansatz des automatischen Indexierens zweifeln, da eine Regelhaftigkeit des Indexierungsvorgangs, die Voraussetzung einer jeden Prozeßautomatisierung ist, tatsächlich nicht erkennbar zu sein scheint. Zugleich jedoch werden auch der intellektuellen Indexierung mangelhafte Resultate aufgrund von Konsistenzproblemen nachgewiesen.⁵ Wieder können hier nur empirische Retrieval-Studien anhand konkreter Systeme und definierter Anwendungsumgebungen einen Aufschluss über das geeignetere Indexierungsverfahren geben.

Ein gravierendes Problem begriffsorientierter Ansätze ist der hohe Aufwand für die Implementierung. Diese Verfahren benötigen in der Regel schon für eingeschränkte Fachgebiete umfangreiche Erkennungswörterbücher und Thesauri sowie komplexe Mechanismen des Mappings zwischen diesen Werkzeugen. Für thematisch offene Indexierungsaufgaben gibt es für diesen Ansatz bislang keine befriedigenden Lösungen.

An der Sprachoberfläche orientierte Verfahren, die durch geeignete Massnahmen die Vielgestaltigkeit der Wortformen für ein Retrieval einschränken, scheinen auf Dauer einen wertvolleren Beitrag zur Verbesserung des Information Retrieval zu leisten.

⁵ Vgl. zur Konsistenzproblematik bei der Indexierung Panyr 1983.

2.2 Verfahrensübersicht

Automatische Verfahren der Indexierung lassen sich in die folgenden Kategorien einteilen⁶:

■ Einfache Stichwortextraktion / Volltextinvertierung

Gleichwohl dieses Verfahren automatisch arbeitet, wird es dennoch nicht eigentlich zu den automatischen Indexierungsverfahren gezählt, da wenigstens eine Auswahl- und/oder Bearbeitungsfunktion erwartet würde. Diese Technik wird hier daher nicht weiter behandelt.

■ Statistische Verfahren

Bspw. die Ansätze von H.P. Luhn, G. Salton sowie diversen Nachfolgern. Unter den Indexierungsverfahren sind statistische Ansätze historisch die ersten Ansätze gewesen. Heute sind vielfältigste Variationen im Einsatz, die jedoch noch immer auf die Grundannahmen Luhn's zurückgehen (Luhn 1958; Rijsbergen 1979; Salton/McGill 1987).

■ Informations- bzw. Computerlinguistische Verfahren

◆ Regelbasierte Verfahren

Bspw. in Okapi realisiert oder OSIRIS. Eine Verfahrensklasse, die sprachliche Regelmäßigkeiten in Algorithmen (bspw. Reduktionsalgorithmen) fasst und so eine generalisierte Textbearbeitung anstrebt (Kuhlen 1974; Schneider 1985; Ahlfeld 1995; Ronthaler/Sauer 1997).

◆ Wörterbuchbasierte Verfahren

Diese Verfahren beruhen auf umfangreichen Wörterbüchern. Die Wörterbücher dienen der Erkennung von Termen sowie ihrer anschließenden Bearbeitung. Alle möglichen Bearbeitungen (Reduktion, Dekomposition usw.) müssen in einem Wörterbuch angelegt sein. Beispiele: Die Indexierungssysteme PASSAT, IDX, EXTRAKT. (Gräbnitz 1987; Lustig 1986; Zimmermann 1996; Lepsky 1996a, b)

■ Pattern-Matching-Verfahren

Verfahren, die anhand erlernter oder implementierter Muster Terme oder Termgruppen erkennen. Beispiel: das Indexierungssystem FIPRAN (Volk et al. 1992).

■ Begriffsorientierte Verfahren

Verfahren, die aufgrund einer Textwortanalyse auf Dokumenteninhalte schließen und diese anschließend durch zugeteilte Indexierungsterme repräsentieren. Beispiele: die Indexierungssysteme AIR/X oder TCS. (Lustig 1986; Rau et al. 1989; Knorz 1994) oder das automatische Verfahren der Klassenzuteilung in GERHARD (Wätjen et al. 1998; Krüger 1999).

Andere Einteilungen sind denkbar, da die meisten realisierten Indexierungssysteme Elemente mehrerer Verfahren integrieren. So kann bspw. AIR/X auch als ein statistisches Verfahren angesehen werden, da die Deskriptoren u.a. auf statistischem Wege gewonnen werden (Reimer 1992, S. 176). Andererseits wenden nahezu alle Verfahren des automatischen Indexierens auch informationslinguistische Ansätze für die Wortstammreduktion an.

Eine gängige und sinnvolle Verfahrenskombination besteht aus informationslinguistischen Bearbeitungen im Vorfeld einer statistischen Indexierung.

⁶ An dieser Stelle ist eine Abgrenzung zur „automatischen Klassifikation“ zu treffen: Wir sprechen im Rahmen dieser Veröffentlichung von „automatischer Klassifikation“ im Sinne eines Clustering, d.h. wenn auch die Klassifikation selbst Ergebnis des automatischen Verfahrens ist. Werden hingegen Dokumente einem bestehenden Klassifikationssystem zugeordnet, rechnen wir diese Ansätze dem „automatischen Indexieren“ zu.

Im folgenden Abschnitt werden die einzelnen Verfahren eingehender betrachtet sowie einige Indexierungssysteme beispielhaft beschrieben.

3. Verfahren der Automatischen Indexierung

3.1 Statistische Verfahren

Die „Initialzündung“ statistischer Indexierungsansätze – der Beschäftigung mit automatischer Indexierung überhaupt – geht von H.P. Luhn aus, der in seinem berühmten Aufsatz *The Automatic Creation of Literature Abstracts* (1958) die folgende Prämisse formulierte:

It is here proposed that the frequency of word occurrence in an article furnishes a useful measurement of word significance.

Luhn ging es zunächst um einen Ansatz für automatisch generierte Abstracts, wenn er anschliessend über die ermittelten Worthäufigkeiten auf die Signifikanz von Sätzen abhebt. Seine Grundannahme, die Signifikanz von Wörtern für die Bedeutung eines Textes auf statistischem Wege ermitteln zu können, wurde in der Folgezeit jedoch in der Hauptsache als eine Ausgangsthese für eine Indexierung gewählt. Eine weitere wichtige Erkenntnis geht dabei auf den amerikanischen Philologen G.K. Zipf zurück, der statistische Gesetzmässigkeiten der Sprache studierte und im *Zipfschen Gesetz* eine konstante (C) Beziehung zwischen dem Rang (r) eines Wortes in einer Häufigkeitsliste und der Frequenz (f), mit der es in einem Text vorkommt postulierte:

$$r \times f = C$$

Die statistischen Indexierungsansätze gehen – im Unterschied zur Invertierung aller Terme eines Textes (Volltextinvertierung) – von folgenden zwei Grundpositionen aus:

- a) nicht alle Terme eines Dokuments sind als Indexterme geeignet – es muss daher eine geeignete Auswahl getroffen werden;
- b) nicht alle ausgewählten Indexterme besitzen hinsichtlich der inhaltlichen Bedeutung die gleiche Wertigkeit – es muss daher eine Gewichtung der Indexterme vorgenommen werden.

Diese Differenzierungen zwischen einzelnen Termen werden anhand von statistischen Häufigkeiten ihres Auftretens ermittelt (Termfrequenzen). Statistische Indexierungsverfahren sind Oberflächenverfahren, d.h. sie versuchen nicht die tieferliegende Bedeutung eines Wortes zu ermitteln oder gar zu „verstehen“. Statistische Masszahlen werden als semantische Indikatoren verwendet. D.h., statistische Masszahlen der Frequenz des Auftretens eines Terms innerhalb eines Dokumentes bzw. in einer Dokumentensammlung werden als Anhaltspunkt für eine geringere oder höhere Bedeutung hinsichtlich des Inhalts angesehen. Hier wird mit rein statistischen Mitteln der Versuch unternommen die Frage zu beantworten: „Wann ist ein Term ein guter Indexterm?“

Die Termfrequenz (TF) im Dokument lässt sich mit der folgenden Formel berechnen:

$$TF_{td} = \text{FREQ}_{td} / \text{GESAMT}_{td}$$

FREQ_{td} = Frequenz eines Terms im Dokument
 GESAMT_{td} = Gesamtzahl der Terme im Dokument

Die Anwendung der Formel soll an einem Beispiel demonstriert werden. Dazu wird eine Agenturmeldung von Reuters herangezogen⁷. Wir betrachten dazu den Term „Marktplatz“, er tritt im Dokument 4mal auf. Das Gesamtdokument enthält 87 Terme (Sonderzeichen wurden als Leerzeichen gewertet.).

⁷ Die Beispieltex te I und II sind im Anhang dieses Beitrages abgedruckt.

$$TF_{td} = 4 / 87 = 0,05$$

Um die Spannbreite der ermittelten Gewichte kleiner zu halten, wird häufig auch mit logarithmischen Werten gerechnet (nach Stock 2000, S. 161):

$$TF_{td} = (\text{ld } [FREQ_{td} + 1]) / \text{ld } GESAMT_{td}$$

ld = logarithmus dualis
(Logarithmus auf der Basis 2)

Reimer (1992, S. 175) beschreibt den Termfrequenzansatz wie folgt:

Nach diesem Ansatz ist ein Indexterm also je aussagefähiger für den Inhalt eines Dokuments, je häufiger er in einem Dokument auftritt und je seltener er überhaupt vorkommt.

Daher muss neben der bereits ermittelten Termfrequenz im Dokument auch die Frequenz eines Terms in der gesamten Dokumentkollektion berechnet werden:

$$TF_{tk} = FREQ_{tk} / GESAMT_{tk}$$

$FREQ_{tk}$ = Frequenz eines Terms i.d. Kollektion
 $GESAMT_{tk}$ = Gesamtzahl der Terme i.d. Kollektion

Unser Beispiel: Wir nehmen an, der Term „Marktplatz“ tritt in der gesamten Dokumentenmenge 350mal auf und die Gesamtzahl der Terme in der Kollektion beträgt 100.000.

$$TF_{tk} = 350 / 100.000 = 0,0035$$

Die Signifikanz (S) eines Terms kann nun u.a. wie folgt ermittelt werden:

$$S = TF_{td} - TF_{tk}$$

$$S = 0,05 - 0,0035 = 0,0465$$

Dem Termfrequenzansatz liegen also folgende Annahmen zugrunde:

1. Häufig auftretende Wörter haben für die Bedeutung *eines Dokuments* eine höhere Signifikanz als Wörter mit einem geringem Vorkommen, sind aus dieser Sichtweise bessere Indexterme.
2. Seltener auftretende Wörter haben innerhalb einer *Dokumentenkollektion* einen höheren Diskriminanzeffekt als häufig vorkommende Wörter, sind damit aus dieser Sichtweise bessere Indexterme.

Um in der Praxis der Indexierung beide Faktoren in ein direktes Verhältnis zu setzen, wird meist die *inverse Dokumenthäufigkeit (IDF)* herangezogen. Dabei wird die Frequenz eines Terms (t) in einem Dokument (d) ermittelt (s.o.) und in Beziehung gesetzt zu der Anzahl der Dokumente in denen (t) auftritt:

$$IDF(t) = FREQ_{td} / DOKFREQ_t$$

Auch hier sei wieder ein Beispiel gegeben: Im gegebenen Beispieldokument tritt „Marktplatz“ 4mal auf. Nehmen wir an, in einer 1000 Dokumente umfassenden Kollektion enthalten 50 Dokumente den Term „Marktplatz“.

$$IDF(t) = 4 / 50 = 0,08$$

Um die Spannweite der Gewichtungswerte nicht zu gross werden zu lassen, kann bei der Berechnung der inversen Dokumenthäufigkeit mit logarithmischen Werten gearbeitet werden. So bspw. in der Formel von Sparck Jones (1972), in der N für die Gesamtzahl der Datensätze steht und n für die Anzahl der Datensätze, in denen t auftritt:

$$\text{IDF}(t) = (\log_2 N / n) + 1$$

„Gute Indexterme“ (entscheidungsstarke Indexterme) weisen eine hohe Frequenz bei gleichzeitig niedriger Dokumentfrequenz auf. Je höher der Wert für die inverse Dokumenthäufigkeit ist, desto entscheidungsstärker ist ein Indexterm. Die Entscheidungsstärke beschreibt die Fähigkeit eines Indexterms, relevante Dokumente aus einer Kollektion zu selektieren und irrelevante zurückzuweisen. Entscheidungsstärkste Deskriptoren sind die Terme in einem mittleren Frequenzbereich (B). Hoch- (A) und niedrigfrequente (C) Terme erfüllen das Kriterium der Entscheidungsstärke nicht, sie werden über die zu definierenden Schwellenwerte (D) und (E) ausgeschlossen.

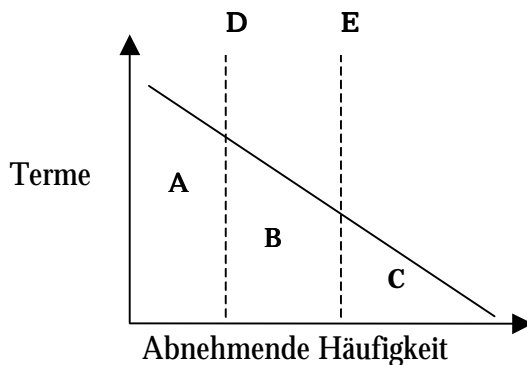


Abb. 5: Termhäufigkeitsverteilung

Diese Berechnungsmethode wurde im Laufe der Zeit vielfältig modifiziert bzw. verfeinert, besteht im Kern jedoch bis heute fort (Salton/McGill 1987). In der Fachliteratur werden diverse Varianten behandelt. Als Verfeinerung dieses Ansatzes können die eingehenden Terme bereits einer Gewichtung unterzogen werden, bspw. indem ihr Auftrittort im Dokument berücksichtigt wird. So können bspw. Terme aus dem Titel oder aus Kapitelüberschriften höher gewichtet werden als Terme aus dem Textkörper. Solcherart Eingangsstufungen nehmen in der Praxis auch einige Suchmaschinen des Internet vor.

Als Terme werden dabei meist die Grundformen angesehen, d.h. vor einer statistischen Berechnung müssen die auftretenden Formen eines Wortes auf ihre jeweilige Grundform zurückgeführt werden. Im o.a. Beispiel würde also auch die Wortform „Marktplätze“ in die Berechnung der Grundform „Marktplatz“ eingehen. Damit kommen auch bei primär statistischen Verfahren linguistische Ansätze zur Geltung (vgl. Abschnitt 3.2).

Die Verwendung der Termgewichtung als Indexierungsverfahren hat ein gewichtetes Retrieval mit anschliessendem Ranking der ausgegebenen Dokumente nahezu zwangsläufig zur Folge.

Voraussetzungen für die Anwendung statistischer Verfahren sind:

- a) mindestens Referate
D.h. es wird eine für ein statistisches Analyseverfahren ausreichende Textbasis für jedes Dokument benötigt
- b) ein homogener Diskursbereich

D.h. ein Themenbereich, in dem auftretende Terme ein grundsätzlich ähnliches Gewicht für den fachlichen Diskurs besitzen. Bei einer Anwendung in sich thematisch überschneidenden Diskursbereichen bieten die Ergebnisse statistischer Verfahren tendenziell keinen Mehrwert.

c) eine grosse Dokumentensammlung.

Rein statistische Verfahren weisen ohne Zweifel einige Probleme auf. So erkennen diese Verfahren keine Homographen („Bank“ / „Bank“) und können folglich auch keine Unterscheidung hinsichtlich der verschiedenen Bedeutungen vornehmen. Problematischer erscheint noch die Nichtbehandlung von Mehrwortbegriffen. „Total Quality Management“, „Kosten- und Leistungsrechnung“ oder – aus dem Beispieldokument I – „elektronischer Marktplatz“ werden nicht als begriffliche Einheit erkannt. Sie werden in ihre Einzelworte aufgelöst und separat behandelt. Dabei wird für „Kosten-“, nicht die notwendige Wortbindestrichergängung zu „Kostenrechnung“ durchgeführt. Komposita bilden ein weiteres Problem. „Qualitätsinformationssystem“ kann durch rein statistische Verfahren nicht auch unter „Qualität“ und „Informationssystem“ indexiert werden. Abhilfe können hier die unter 3.2 behandelten informationslinguistischen Verfahren schaffen.

3.1.1 Beispiel einer einfachen statistischen Indexierung

Gegeben sind fünf Texte:

Text 1:

Computer werden im Information Retrieval eingesetzt. Es existieren Verfahren auf Computern für ein automatisches Retrieval. Moderne Computer ermöglichen ein effizientes Retrieval nach spezifischer Information.

Text 2:

Nutzer von Systemen zum Information Retrieval wurden befragt. Viele Nutzer waren mit der Funktionalität des Retrieval zufrieden. Die vorhandenen Systeme zum Information Retrieval genügen den Anforderungen der Nutzer. Es existieren eine Reihe von Systemen auf Computern.

Text 3:

Die Entwicklung neuer Systeme für das Information Retrieval wird von vielen Nutzern begrüßt. Die Entwicklung zielt auf neue Methoden des Retrievals mit Computern ab. Systeme zum effizienten Retrieval nach Information befinden sich derzeit in der Entwicklung.

Text 4:

Das Information Retrieval wird in Datenbanken durchgeführt. Verschiedene Datenbanken haben eine Oberfläche für den Nutzer, die ein zielgerichtetes Retrieval in Informationsräumen ermöglicht. Verschiedene Systeme für ein Retrieval in Datenbanken stehen derzeit dem Nutzer zur Verfügung.

Text 5:

Die Entwicklung von Systemen zum Retrieval in Informationsräumen ist für viele Nutzer von Datenbanken interessant. In Informationsräumen kann man navigieren und somit das Information Retrieval unterstützen. Der Informationsraum wird dreidimensional auf Computern visualisiert.

Schritt 1:

Zunächst sind Vorbedingungen für die Auswahl der weiterzuverarbeitenden Terme zu definieren:

Über eine Stoppwortliste werden die folgenden Terme aus der weiteren Bearbeitung ausgeschlossen:

Verfahren, Anforderung, Reihe, Methode, Verfügung, Funktionalität, Oberfläche

Der weiteren Bearbeitung liegen informationslinguistische Auswahl- und Bearbeitungsregeln zugrunde:

a) es werden nur Substantive extrahiert, b) extrahierte Wörter werden reduziert auf Nominativ und Singular

Schritt 2:

Bestimmung der Termfrequenz (FREQ) und der Dokumentfrequenz (DOKFREQ):

Indexterm	FREQ1	FREQ2	FREQ3	FREQ4	FREQ5	DOKFREQ
Computer	3	1	1	-	1	4
Information	2	2	2	1	1	5
Retrieval	3	3	3	3	2	5
Nutzer	-	3	1	2	1	4
System	-	3	2	1	1	4
Entwicklung	-	-	3	-	1	2
Datenbank	-	-	-	3	1	2
Informationsraum	-	-	-	1	3	2

Tab. 1: Term- und Dokumenthäufigkeit

Schritt 3:

Berechnung der Termgewichtung eines jeden Indexterms mit dem Termfrequenzansatz:

$$\text{Termgewichtung} = \text{FREQ}_{dt} / \text{DOKFREQ}_t$$

Dabei wird eine Gewichtung ermittelt, die eine Relation herstellt aus der Häufigkeit eines Terms t im Dokument d (FREQ_{dt}) und umgekehrt proportional der Gesamtzahl der Dokumente (DOKFREQ_t), in denen der betreffende Term auftritt. Die Berechnung führt zu folgendem Ergebnis:

Indexterm	Text1	Text2	Text3	Text4	Text5
Computer	0.75	0.25	0.25	0	0.25
Information	0.4	0.4	0.4	0.2	0.2
Retrieval	0.6	0.6	0.6	0.6	0.4
Nutzer	0	0.75	0.25	0.5	0.25
System	0	0.75	0.5	0.25	0.25
Entwicklung	0	0	1.5	0	0.5
Datenbank	0	0	0	1.5	1.5
Informationsraum	0	0	0	0.5	1.5

Tab. 2: Termgewichtung

Schritt 4:

Für die Bestimmung „guter Indexterme“ kann nun ein unterer Schwellenwert eingeführt werden, der ein weiteres Steuerungsinstrument der Auswahl darstellt. Hier wird der Schwellenwert von 0.5 festgelegt. Ein oberer Schwellenwert braucht hier nicht definiert zu werden, da hochfrequente Wörter im ersten Schritt über eine Stopwortliste ausgeschlossen wurden.

Auf der Grundlage der ermittelten Termgewichtungen und unter Berücksichtigung des Schwellenwertes kann nun ein invertierter Index erstellt werden:

Indexterm	Texte				
Computer	Text 1				
Retrieval	Text 1	Text 2	Text 3	Text 4	
Nutzer		Text 2		Text 4	
System		Text 2	Text 3		
Entwicklung			Text 3		Text 5
Datenbank				Text 4	Text 5
Informationsraum				Text 4	Text 5

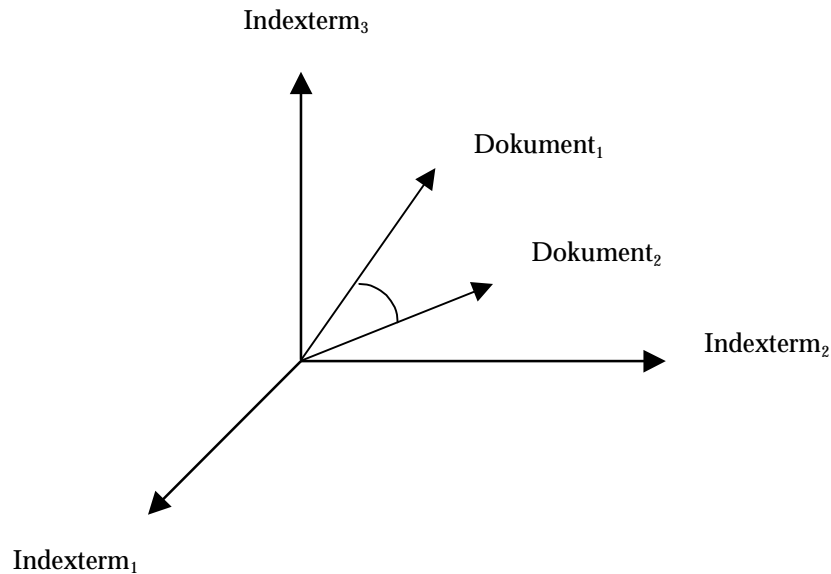
Tab. 3: Invertierter Index

Das Ergebnis dieser Tabelle kann als eine automatische Indexierung der Texte für ein Retrieval-System bereits genutzt werden. D.h. die automatische Indexierung ist an dieser Stelle zu einem ersten verwertbaren Resultat gelangt und wird vielfach an dieser Stelle abgeschlossen.

Schritt 5:

Ein weiterführendes Modell zu einer automatischen Klassifizierung entsteht, wenn auf Basis der bereits ermittelten Termfrequenzen (Tabelle 2) Dokumentähnlichkeiten berechnet werden. Auf diese Weise wird ein *Vektorraummodell* aufgebaut. Das Vektorraummodell ist heute die am häufigsten eingesetzte Methode des Information Retrieval und in zahlreichen IR-Systemen im Einsatz (Fuhr 1997).

Ein automatisches Verfahren könnte auch ohne die früher errechneten Termgewichte aus Tabelle 2 an dieser Stelle einsetzen und stattdessen einfache Binärvektoren verwenden. Dabei erzeugt ein auftretender Term eine Eins, anderenfalls eine Null. Der Vektor unseres Beispieltextes vier sähe dann folgendermassen aus: (0,1,1,1,1,0,1,1). Im Vektorraummodell wird der Dokumentenraum durch n Indexterme aufgespannt. Vektorräume sind daher vieldimensionale Räume in denen jedes Dokument aufgrund der gewichteten bzw. binären Indexierung als Dokumentvektor repräsentiert wird. Suchfragen werden wie Dokumente behandelt und ebenfalls durch einen Vektor im Raum repräsentiert. Zwischen Dokumentvektoren und Anfragevektoren kann somit ein Ähnlichkeitsabgleich vorgenommen werden.

**Abb. 6:** Vektorraummodell

Eine paarweise Dokument-Dokument-Ähnlichkeit ergibt sich aus dem Skalarprodukt $\sum x_i y_i$. D.h. die Termgewichtungen (Tab. 2) werden jeweils paarweise miteinander multipliziert und aufsummiert. Dieser Berechnung der Ähnlichkeit der Indexterme zweier Dokumente D_i und D_j liegt folgende Formel zugrunde:

$$\text{Ähn}(D_i, D_j) = 1/n \sum_{k=1}^n g_{ik} g_{jk}$$

n = Anzahl der Indexterme
 g_{ik} = Gewicht des Indexterms k im Dokument D_i

Das Resultat dieser Berechnung, für die Texte zusammengefasst, ergibt folgende Matrix:

	Text 1	Text 2	Text 3	Text 4	Text5
Text 1	-	0.71	0.71	0.44	0.51
Text 2	0.71	-	1.15	1.01	0.76
Text 3	0.71	1.15	-	0.68	1.31
Text 4	0.44	1.01	0.68	-	1.96
Text 5	0.51	0.76	1.31	1.96	-

Tab. 4: Dokument-Dokument-Ähnlichkeitsmatrix

Das Verfahren wird folgendermassen interpretiert (Riggert 1998, S. 69):

- Niedrigfrequente Terme verringern die Ähnlichkeit zwischen den Dokumenten, da sie seltener auftreten.
- Hochfrequente Terme steigern die Ähnlichkeit, da sie in vielen Dokumenten auftreten.
- Terme mittlerer Frequenz gliedern eine Kollektion tendenziell in inhaltlich verwandte Cluster.

Die sorgfältige Wahl geeigneter Schwellenwerte spielt bei diesen Verfahren eine zentrale Rolle. Eine Reihe anderer Ähnlichkeitsmasse (z.B. das Cosinus-Mass) beschreibt Gebhardt (1981, S. 165 ff).

Das Vektorraummodell erlaubt nun mittels eines automatischen Clusterverfahrens⁸ eine automatische Klassifizierung der Dokumente. Clusteranalytische Verfahren werden eingesetzt, um hochkomplexe Strukturen (Klassen, Hierarchien von Clustern) in hochdimensionalen Räumen aufzufinden, wenn diese mit einem Distanzmaß ausgestattet werden können. Dem Verfahren liegt die Annahme zugrunde, dass eine Struktur von Klassen innerhalb dieses Merkmalsraumes existiert. Cluster sind Mengen ähnlicher Dokumente, die bereits bei der Anlegung der Kollektion ermittelt und gemeinsam gespeichert werden. Dabei wird angenommen, dass die Ähnlichkeit zwischen Dokumenten, über die Relevanz von Dokumenten bezüglich einer Suchanfrage entscheidet (Teuber 1996). Die Ähnlichkeit von Dokumenten wird aufgrund Merkmalen berechnet, als Merkmale werden die Textterme herangezogen.

Aus der Ähnlichkeitsmatrix können Cluster durch eine Reihe verschiedener Algorithmen berechnet werden. Ein einfacher Cluster-Algorithmus sähe bspw. folgendermassen aus:

1. Wahl eines geeigneten Schwellenwertes für die Ähnlichkeit.
2. Clusterbildung durch Dokumente, die den Schwellenwert unterschreiten.
3. Alle Dokumentenpaare, deren Ähnlichkeit den Schwellenwert überschreiten, werden durch eine Kante verbunden.
4. Die maximal zusammenhängenden Komponenten des Ähnlichkeitsgraphen bilden die gesuchten Cluster.
5. Abschliessend die Bestimmung eines Zentroiden als Repräsentanten des Clusters.

Dokumentrecherchen sowie die Einordnung neuer Dokumente geschehen über einen Vergleich mit den Clusterzentroiden. Im Falle der Hinzufügung neuer Dokumente muss der Clusterzentroid jeweils neu berechnet werden.

Wird im Beispiel aus der Tabelle 4 ein Schwellenwert von 1.00 zugrunde gelegt, lässt sich die folgende clustergraphische Darstellung ableiten:

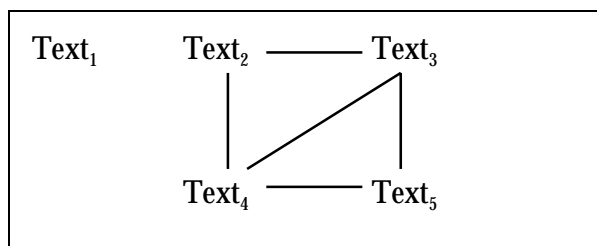


Abb. 7: Clustergraphische Darstellung

⁸ Eine ausführliche Beschreibung des Clusterverfahrens für die Dokumentklassifikation gibt Salton (1978).

3.2 Informationslinguistische Verfahren

Die beschriebenen statistischen Indexierungsverfahren berücksichtigen, sofern sie allein Anwendung finden, keine Phänomene auf der Sprachebene. So würden bspw. folgende Terme bei der Termfrequenzermittlung jeweils als eigenständiger Term angesehen und behandelt werden:

Beispiel 1	Beispiel 2	Beispiel 3
Zahlung Zahlungen Zahlungstermin bargeldlose Zahlung	Haus Häuser Hausmakler Bürohaus	Niederschlag niederschlagen Niederschlagung niedergeschlagen

In diesen Beispielen sind eine Reihe (weitaus nicht alle!) verschiedener sprachlicher Probleme angedeutet. Diese Beispiele lassen sich betrachten vor dem Hintergrund,

- informationslinguistische Lösungen in der Vorbereitung für eine statistischen Indexierung einzusetzen (bspw. in Vorbereitung auf den Termfrequenzansatz) oder
- informationslinguistische Lösungen allein zur Produktion automatischer Indexate anzuwenden.

Statistische Verfahren sollten wenigstens Teillösungen der informationslinguistischen Ansätze berücksichtigen (bspw. eine Grundformenreduktion), da in der Folge die Termfrequenzermittlung zu besseren Ergebnissen kommen kann.

Eine Diskussion der o.g. Beispiele im Detail:

Beispiel 1:

„Zahlung“ und „Zahlungen“ stellen dasselbe Wort dar, jedoch verschiedene *Wortformen*. „Zahlungen“ ist eine Flexionsform der Grundform „Zahlung“, mit dem regelmässigen Flexionsuffix „en“. Wortformen der selben Grundform dürfen in der Termfrequenzermittlung nicht getrennt betrachtet werden. Bei informationslinguistischer Indexierung muss die Wortform durch eine Reduktion auf die Grundform oder gar die Stammform zurückgeführt werden. „Zahlungstermin“ ist ein Kompositum, dessen einzelne Bestandteile für die Indexierung berücksichtigt werden müssten (Kompositumszerlegung). Schliesslich ist die „bargeldlose Zahlung“ ein Mehrwortbegriff, der auch in seiner Gänze als Indexterm berücksichtigt werden müsste.

Beispiel 2:

Im Falle von „Haus“ und „Häuser“ liegen wiederum Grundform und pluralbedingte Wortform vor. Die Anforderungen einer Zusammenführung sind auch hier gegeben, jedoch lässt sich dieses Problem nicht durch eine einfache Wortformenreduktion lösen, da die Pluralbildung durch eine Stammformveränderung hervorgebracht (unregelmässige Pluralbildung) wurde. „Hausmakler“ und „Bürohaus“ sind Komposita, bei denen wiederum eine entsprechende Zerlegung vorgenommen werden sollte, falls diese als sinnvoll erachtet wird.

Beispiel 3:

In diesem Beispiel ist zunächst nicht offensichtlich, wie sich die einzelnen Wörter zueinander verhalten, da ihre Bedeutung ohne Kontext jeweils uneindeutig ist. Nehmen wir folgende Kontexte an: „... muss mit Niederschlag in Form von Schnee gerechnet werden.“, „... wird er den Prozess niederschlagen.“, „... nach Niederschlagung der Revolte ...“, „... wurde er durch seinen Gegner niedergeschlagen“. Einfache informationslinguistische Verfahren stossen hier an ihre Grenzen.

Bei informationslinguistischen Verfahren geht es hauptsächlich darum, die nachfolgenden Aufgaben zu lösen (Stock 1998, Zimmermann 1983):

- nicht sinntragende Wörter eliminieren um sie damit aus der Indexierung auszuschliessen
- grammatische Flexionsformen auf eine Grundform oder Stammform zu bringen (Wortstammanalysen)
- Komposita sinnvoll zu zerlegen
- Phrasierungen (Mehrwortbegriffe) zu erkennen
- Pronomina korrekt zuzuordnen

Die Informationslinguistik untersucht sprachliche Probleme der Textanalyse, wie sie typischerweise im Kontext des Information Retrieval auftreten. Sie befasst sich mit der Verarbeitung natürlicher Sprache in bzw. für Informationssysteme, sie ist somit eine Disziplin auf der Schnittstelle zwischen Informationswissenschaft und Computerlinguistik (Luckhardt 1998). Zwar wird auch bei statistischen Verfahren eine Verarbeitung sprachlicher Einheiten (Texttermen) vorgenommen. Da allerdings keine wirkliche sprachliche Analyse erfolgt, werden statistische Ansätze nicht zu den informationslinguistischen Verfahren gezählt. Informationslinguistik im Allgemeinen und die Indexierungsverfahren auf ihrer Grundlage im Besonderen sind abhängig vom gegebenen Sprachsystem. So treten für die deutsche Sprache bspw. Problemstellungen auf, die für die englische Sprache keine Rolle spielen (vgl. Stock 2000): bspw. Lemmatisierung, Kompositazerlegung oder die Erkennung und Behandlung von Wortbindestrichergänzungen.

Zum Stand und den Möglichkeiten der Informationslinguistik vgl. Winiwarter (1996), zu Evaluationen von Retrieval-Systemen mit linguistischen Komponenten Ruge/Goeser (1998).

Linguistisch basierte Indexierungsverfahren können in ihrer Analyse auf drei Ebenen der Sprache ansetzen (Kaiser 1993):

- Morphologische Analyse
- Syntaktische Analyse
- Semantische Analyse

Diese Aufgaben fallen sowohl bei der Indexierung von Dokumenten als auch bei der entsprechenden Aufbereitung von Retrievalfragen an. Durch die Lösung dieser Aufgaben erreichen informationslinguistische Verfahren bis zu einem gewissen Grade eine Unabhängigkeit von der jeweils verwendeten sprachlichen Ausdrucksform zur Darstellung von Sachverhalten. Dieser Ansatz trägt damit dem Variantenreichtum sprachlicher Ausdrucksformen Rechnung.

Heute realisierte Indexierungssysteme bieten keine erschöpfenden Lösungen.

Bei der Vielfalt und der Komplexität der Probleme, die die natürliche Sprache stellt, sind perfekte Lösungen entweder unverhältnismässig aufwendig oder gegenwärtig gar nicht erreichbar. (Knorz 1994, S. 149)

Perfekte Lösungen für komplexere sprachliche Analysen bedürften der Realisierung aller drei oben erwähneter Analyseschritte, dabei entstehen jedoch unverhältnismässig aufwendige Verfahren. Das oben behandelte dritte Beispiel, mit seinen kontextabhängigen semantischen Differenzierungen, wäre nur unter Einsatz aufwendiger Verfahren lösbar.

Bestrebungen, eine Annäherung an perfekte Lösungen zu erreichen, können vor allem in den Ansätzen gesehen werden, eine weitgehende Syntaxanalyse (Parsing) durchzuführen (Werner 1982; Schneider 1985). Syntaktische Analysen sollten insbesondere der im jeweilig vorliegenden Kontext korrekten Grundformenreduktion dienen. Zudem wurden syntaxanalytische Verfahren für die Identifizierung von Homographen eingesetzt (vgl. obiges Beispiel 3). Der wohl weitestgehende Anspruch syntaxanalytischer Verfahren liegt in der Erschliessung kompletter syntaktischer Strukturen. Dabei wird versucht über durch Zwischenräume gekennzeichnete Beschreibungselemente („Wort“) hinaus, Einheiten der Sprache zu identifizieren, die aus mehreren Elementen („Mehrwortgruppen“) bestehen (Schneider 1985). Einen solchen Ansatz verfolgte bspw. das Indexierungssystem CTX in der Anwendung JUDO (Werner 1982; Zimmermann et al. 1983; Schneider 1985; Keitz 1996).

Zimmermann et al. (1983) geben folgende zu identifizierende ober-flächensyntaktischen Strukturen an:

1. Adjektivattribut – Substantiv
2. Substantiv – Genitivattribut
3. Substantiv – Präpositionalattribut
4. Substantiv – beigeordnetes Substantiv

Dabei galt es u.a. aus einem Satz: „... wenn sie *unvollständige* oder entstellende *Angaben* enthalten.“, die Adjektivattribut-Substantiv Struktur „unvollständige Angabe“ zu identifizieren und als Indexterm vorzusehen. Der Ausdruck „Künstliche Intelligenz“ sollte als mehrgliedrige Nominalphrase erkannt werden.

In der Syntaxanalyse können Ansätze eines partiellen Parsing (wie im System CTX) unterschieden werden von den Bemühungen, ein vollständiges Parsing (System CONDOR) durchzuführen. Diese syntaxorientierten Forschungsansätze hatten eine Blütezeit in den 70er und 80er Jahren, haben sich letztlich jedoch nicht bewährt. Syntaxanalysen führen schnell zu unverhältnismässig aufwendigen und komplexen Lösungen, ohne eine wirklich befriedigende Indexierungslösung erreichen zu können (Reimer 1992).

Als wesentliches Forschungsergebnis bleibt festzuhalten, dass Mehrwortdeskriptoren nicht rein syntaktisch bestimmt werden können. (Goeser 1994, S. 22)

Semantische Analyseschritte finden auf der Ebene des ganzen Dokuments statt. Über semantische Analysen wird versucht, kontextuelles Wissen zu verarbeiten und einen Text in bedeutungsabhängige Einheiten zu zerlegen.

Gleichwohl sind informationslinguistische Lösungen in Kombination mit anderen Indexierungs- und Retrieval-Verfahren (statistischen oder begriffsorientierten) von grösstem Interesse. So wird im System OSIRIS (Ronthaler/Sauer 1997) ein u.a. partielles Parsing auf Nominalphrasen eingesetzt. OSIRIS bildet letztlich natürlichsprachigen Input aus Dokumenten oder Systemanfragen auf die Klassen eines Klassifikationssystems ab.

Erfolgreiche Indexierungssysteme, die allein aufgrund informationslinguistischer Verfahren Resultate erzielen, sind heute meist pragmatisch ausgelegt und agieren hinsichtlich einer Bereitstellung „bereinigter“ Wortformen für ein Retrieval. Dies beinhaltet dann vornehmlich – aber bereits in eingeschränktem Masse – die ersten vier von Stock (1998) angeführten Aufgaben (s.o.). Dafür stellt die Morphologie den wichtigsten theoretischen Hintergrund zur Verfügung. Die Morphologie beschäftigt sich mit den Regularien der inneren Struktur von Wörtern und der

Bildung von Wortklassen und strukturellen Gesichtspunkten (Crystal 1995, S. 90). Morphologische Analysen finden damit auf der Wortebene statt. Das Indexierungsverfahren IDX (s.u.) kodiert bspw. unter diesem Gesichtspunkt die Wörter eines Textes. Die Internet-Suchmaschine ScoutMaster⁹ setzt das informationslinguistische System EXTRAKT ein, um Wortformreduktionen bzw. -expansionen bei der Recherche durchzuführen. Die Sucheingabe „automatische indexierung“ wird durch die Funktion *Wortformen* in folgende Recherche umgesetzt (ODER-Verknüpfung innerhalb einer Zeile, UND-Verknüpfung zwischen den Zeilen):

Suche im ganzen Text

automatische indexierung, automatisch, automatische, automatischem,
automatischen, automatischer, automatisches

und in +/- 1 Zeile automatische indexierung, Indexierung, Indexierungen

Grundsätzlich lassen sich die informationslinguistischen Indexierungssysteme hinsichtlich ihrer Verfahrensgrundlage unterscheiden in

- a) regelbasierte Verfahren und
- b) wörterbuchbasierte Verfahren

Die bereits beschriebenen syntaxanalytischen Verfahren gehören in die erste Gruppe. Mischformen aus regel- und wörterbuchbasierten Ansätzen sind möglich und in der Praxis vorhanden.

Regelbasierte Ansätze versuchen das für Indexierungen notwendige Regelgerüst einer Sprache in Algorithmen zu fassen. Die Implementierung von Regeln ist eine „einmalige“ und damit generalisierende Aufgabe, alle Terme in Dokumenten werden anhand der Regeln bearbeitet. Dieses informationslinguistische Verfahren gilt damit quantitativ als wenig aufwendig, da keine individuelle Auseinandersetzung mit einzelnen Wörtern oder Wortgruppen stattfindet und eine Pflege einmal implementierter Lösungen ausbleiben kann. Neue Wörter werden normalerweise durch vorhandene Regeln korrekt analysiert und bearbeitet. Gleichwohl kann die informationslinguistische Umsetzung sprachlicher Regeln in Algorithmen ein aufwendiges Problem darstellen. Probleme regelbasierter Verfahren bestehen bspw. bei der Bearbeitung unregelmässiger Pluralbildungen („Haus“ – „Häuser“), da diese eine Stammformveränderung beinhalten. Kaum lösbar ist die regelbasierte Zerlegung auftretender Komposita (Versuchen Sie sich einmal an der Zerlegung der Komposita „Glücksautomaten“ oder „Staatsexamen“ sowie der Suche nach einer Regel für ihre korrekte Zerlegung.).

Wörterbuchgestützte Verfahren beruhen auf „Einzelfalllösungen“. D.h. jeder zu analysierende Term muss mit allen Möglichkeiten der Behandlung in einem Wörterbuch abgelegt sein. Dies gilt auch für Mehrwortbegriffe. Damit sind diese Verfahren durch die umfangreiche und kontinuierlich zu betreibende Wörterbuchpflege in hohem Masse arbeits-, zeit- und kostenaufwendig, in der Regel jedoch wesentlich zuverlässiger, da individuelle Entscheidungen getroffen werden. Es wird bspw. festgelegt, dass „Staatsexamen“ folgendermassen zu zerlegen ist: „Staat | s | Examen“.

Für das Englische sind regelbasierte Verfahren (graphematische Verfahren) mit grossem Erfolg entwickelt und eingesetzt worden (Kuhlen 1974). Die morphologisch wenig komplexe englische Sprache kennt kaum Wortstammveränderungen, d.h. sie ist flektionsarm (vgl. Indexing and Morphology o.J.). Entsprechende Algorithmen sind für morphologisch komplexere Sprachen, wie die flektionsreiche und kompositumsträchtige deutsche Sprache, nicht hinreichend. So hat sich der erfolgreich im britischen Okapi-Retrievalsystem eingesetzte Algorithmus von Porter

⁹ <http://www.scoutmaster.de>, die Funktion ist unter „Recherche“ verfügbar.

(Porter 1980) in einem Test als nur sehr eingeschränkt für deutsche Texte anwendbar erwiesen (Ahlfeld 1995). Die von Kuhlen (1974) beschriebenen Algorithmen führen unterschiedlich weitgehende Reduktionen durch. Reduktionen können auf die *Formale Grundform*, auf die *Lexikalische Grundform* oder auf die *Stammform* ausgeführt werden.

Die Grundformenreduktion dient dazu, entweder Wörter durch das Abtrennen der Flexionsendung auf ihre formale Grundform zurückzuführen. Ergänzend kann anschliessend noch auf die lexikalische Grundform reduziert werden, indem bei Substantiven der Nominativ Singular und bei Verben der Infinitiv gebildet wird. Die Reduktion auf die Stammform findet nach der Grundformreduktion statt und ist eine Entfernung der Derivationsendungen und damit eine Rückführung der Wörter auf ihren Wortstamm.

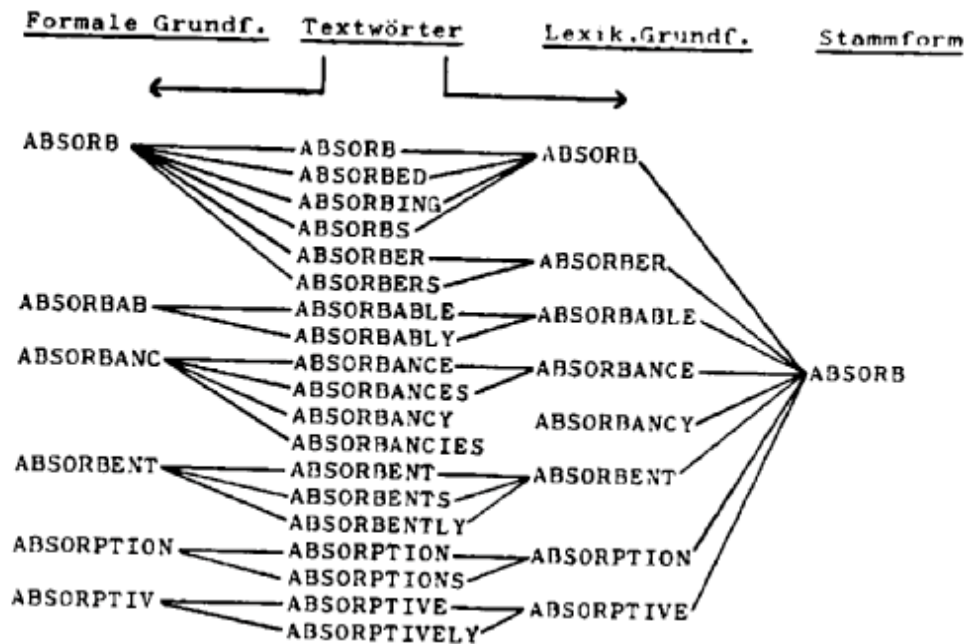


Abb. 8: Reduktionsalgorithmen nach Kuhlen (1974)

Nachfolgend wird für die englische Sprache ein einfacher aber bereits wirkungsvoller Algorithmus für eine lexikalische Grundformenreduktion angegeben:

Notation

%	alle Vokale, einschl. Y
*	alle Konsonanten
!	Länge des Wortes
/	„oder“
§	Leerzeichen
→	„zu“
←	„aus“
\	„nicht“

Regeln des Algorithmus

- IES → Y
- ES → § nach * O/CH/SH/SS/ZZ/X
- S → § nach * /E/%Y/%O/OA/EA
- S' → §
IES' → Y
ES' → §
- 'S → §
' → §
- ING → § nach **/%/X
ING → E nach %*
- IED → Y
- ED → § nach **/%/X
ED → E nach %*

Die Reduktionsalgorithmen behandeln Fälle nicht individuell, sondern wenden implementierte Regeln auf alle vorkommenden Fälle an. Bei dieser generalisierenden Vorgehensweise ist ein ge-

wisser Grad an Fehlern unvermeidlich. Die auftretenden Fehler resultieren aus zwei bekannten Fehlerklassen: Entweder aus einer zu weitgehenden Reduktion (dem *Overstemming*) oder einer Reduktion, die nicht weit genug operiert (dem *Understemming*) (Knorz 1994).

Overstemming:

Verschiedene Wortformen mit gleicher Grund- bzw. Stammform werden falsch zusammengeführt:

den Buch~en \Rightarrow buch	das Eis~en \Rightarrow eis	die Rind~en \Rightarrow rind
des Buch~es \Rightarrow buch	des Eis~es \Rightarrow eis	die Rind~er \Rightarrow rind

Understemming:

Verschiedene Wortformen mit gleicher Grund- bzw. Stammform werden nicht zusammengeführt:

des schlecht(est~en) \Rightarrow schlechtest	die Them~en \Rightarrow them
den schlecht(~en) \Rightarrow schlecht	des Thema~s \Rightarrow thema
der schlecht(er~e) \Rightarrow schlechter	

Eine Mehrwortgruppenerkennung (Phrasenerkennung) ist – wie bereits mehrfach erwähnt – eine der wichtigsten Aufgaben innerhalb der informationslinguistischen Analyse. Im Beispieltext I wäre etwa „elektronischer Marktplatz“ eine entsprechende Mehrwortgruppe, im Beispieltext II „Frankfurter Neuer Markt“. Eine solche Erkennung kann über die Implementierung von Mehrwortgruppen-Wörterbüchern erreicht werden (s.u. die Beschreibung des IDX-Verfahrens). Eine wörterbuchgestützte Lösung hat jedoch den Nachteil, nur jeweils enthaltene Mehrwortgruppen zu erkennen und überaus pflegeaufwendig zu sein. Eine wörterbuchabhängige Mehrworterkennung funktioniert insbesondere bei neuen Ausdrückern nicht.

Eine andere Lösung ist die Zerlegung eines Textes in „Klumpen“ (vgl. Stock 2000). Um dies zu erreichen wird eine ausführliche Stoppwortliste erstellt, die u.a. alle Adverbien, alle Hilfsverben sowie viele weitere Verben enthält. Diese werden als Begrenzer interpretiert, die eine mögliche Mehrwortgruppe einleiten oder abschliessen (Begrenzerverfahren). Im betrachteten Text bleiben zwischen den definierten Stoppwörtern (den Begrenzern) Textklumpen übrig. Sind dies Einzelwörter spielen sie für die Ermittlung von Mehrwortgruppen keine Rolle. Bleiben mehrere nacheinander stehende Wörter übrig, so sind dies mögliche relevante Mehrwortbegriffe für die Indexierung. Ein Begrenzerverfahren wird u.a. im System FIPRAN eingesetzt (Volk et al. 1992).

Ein Beispiel aus dem Beispieldokument I:

Das **Konzept** der drei **Konkurrenz-Konzerne** zielt hingegen auf einen **elektronischen Marktplatz** für die **Zulieferindustrie**.

Der Mehrwortbegriff „elektronischen Marktplatz“ wird links- und rechtsbündig von Stoppwörtern eingeschlossen (begrenzt). Er ist damit Kandidat für einen relevanten Indexbegriff. Tritt dieser Mehrwortbegriff häufiger in diesem oder anderen Dokumenten auf und überschreitet er einen definierten Schwellenwert, wird er als Indexbegriff aufgenommen. Der Mehrwortbegriff wird automatisch in ein entsprechendes Lexikon aufgenommen. Dieses Verfahren ist bei Lexis-Nexis (für die englische Sprache) erfolgreich implementiert (Stock 1998, 2000).

Im gerade beschriebenen Verfahren wird die Welt der Mehrwortgruppen jedoch auf einen einfachen Ausschnitt ihrer möglichen sprachlichen Vorkommensformen reduziert. Stellen wir uns vor, der o.g. Beispielsatz lautete:

Das Konzept der drei Konkurrenz-Konzerne zielt hingegen auf einen **Marktplatz** für die Zulieferindustrie in **elektronischer** Form.

Nicht geändert hat sich die inhaltliche Aussage, beide Sätze geben die inhaltlich identische Information wieder. Geändert hat sich jedoch die Stellung der Komponenten der Mehrwortgruppe im Text. Das geschilderte Verfahren zur Identifizierung von Textklumpen führt hier nicht mehr zum Erfolg. Zur Lösung dieses Problems kann eine Verfahrenskombination bestehend aus einem Mehrwortgruppen-Wörterbuch und einem einfachen Parsing eingesetzt werden:

Zunächst werden einzelne Komponenten einer Mehrwortgruppe identifiziert. Anschliessend wird überprüft, ob die weiteren Komponenten innerhalb eines definierten maximalen Abstandes ebenfalls im Text auftreten. Ist dies der Fall, muss eine Bewertung anhand formaler Kriterien vorgenommen werden:

- Dem Abstand zwischen den einzelnen Komponenten der potenziellen Mehrwortgruppe.
- Für jede Komponente der potenziellen Mehrwortgruppe wird geprüft, ob die gefundene Wortform mit der Grundform im Wörterbuch oder nur mit der Stammform übereinstimmt.
- Es wird geprüft, ob die Reihenfolge der Komponenten im Text gleich der Reihenfolge der Mehrwortgruppe im Wörterbuch ist
- Es wird geprüft, ob die Komponenten im gleichen Satz aufgefunden werden.

Die Ergebnisse dieser Prüfung sind Werte, auf deren Basis Klassen von Vorkommensformen gebildet werden. Im folgendem wird die Wahrscheinlichkeit bestimmt, dass die potentielle Mehrwortgruppe auch syntaktisch korrekt ist. Dabei kommen wiederum Schwellenwerte zum Einsatz.

Die Qualität informationslinguistisch basierter Indexierungen hängt in hohem Masse von der Qualität der verwendeten Wörterbücher bzw. des grundlegenden Regelsystems ab. Insbesondere wörterbuchbasierte Verfahren implizieren einen hohen manuellen Aufwand für die permanente Pflege der Wörterbücher (Keitz 1996). Die erzielbaren Resultate der Indexierung werden durch den Umfang und die Qualität der Wörterbücher bestimmt.

3.2.1 Beispiel: Indexierungsverfahren im IZIS-ET

Nachfolgend wird ein einfacher Verfahrensablauf der automatischen Indexierung elektrotechnischer Texte im „Internationalen Zweiginformationssystem Elektrotechnik (IZIS-ET)“ nach Alliger/Richter (1978) auszugsweise beschrieben:

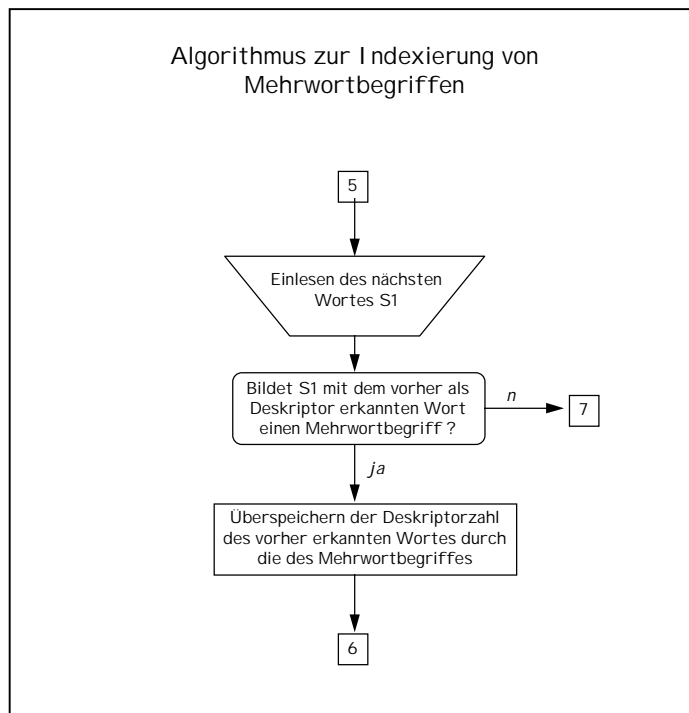
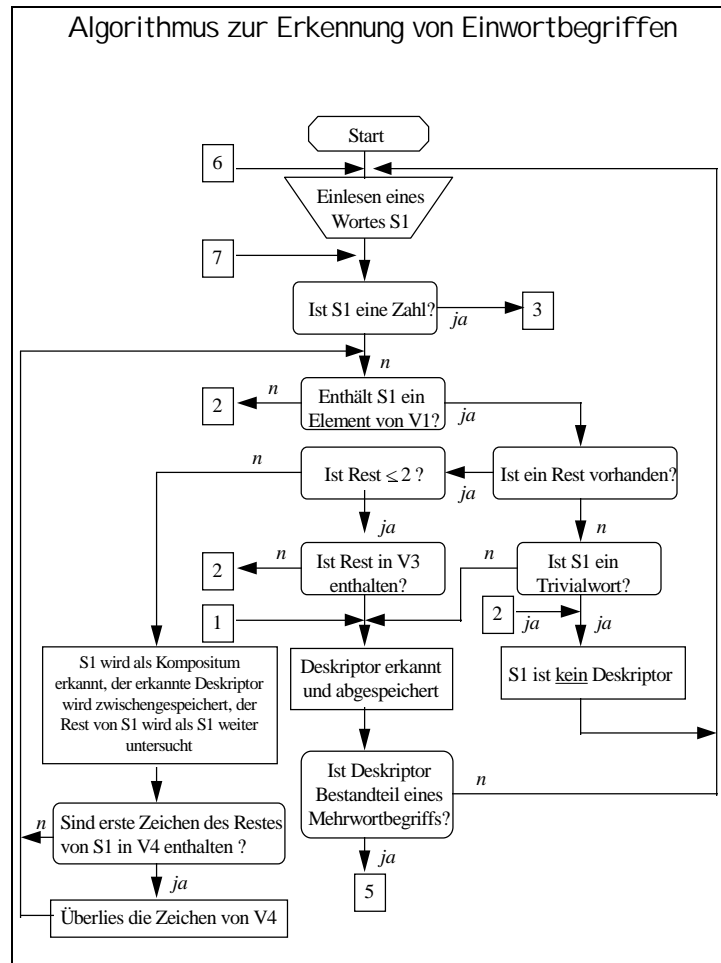


Abb. 9: *Verfahrensablauf informationslinguistischer Indexierung nach Alliger/Richter (1978)*

Legende:

Anschlusskonnektoren:

- 3:** Anschlusskonnektor für die Faktenindexierung
- 5:** Anschlusskonnektor für die Indexierung von Mehrwortbegriffen
- 6:** Abschlusskonnektor für den Rücksprung aus den Algorithmen für die Faktenindexierung und die Indexierung von Mehrwortbegriffen
- 7:** Rücksprungkonnektor der Faktenindexierung

Wörterbücher

- V1:** Wörterbuch deutschsprachiger Begriffe und deren Deskriptorzahlen (Thesaurus)
- V2:** Liste der Endungen *end, los, frei* (Wörter mit diesen Endungen werden nicht zerlegt)
- V3:** Verzeichnis der Flexionsendungen (Endungen aus 1 oder 2 Buchstaben)
- V4:** Verzeichnis der Bindelaute innerhalb von Komposita
- V5:** Verzeichnis von Endungen als Hilfsmittel zur Zerlegung von Komposita
- V6:** Verzeichnis von Mehrwortbegriffen und deren Deskriptorzahlen
- V7:** Verzeichnis von Fakten und deren Deskriptorzahlen (elektrische und andere physikalische Größen)
- V8:** Verzeichnis der Masseneinheiten (einschl. deren Umrechnungsfaktoren)

Das System indexiert Fachbegriffe, die im Text in folgender Form auftreten können:

- einfache Wörter (Simplizia): Substantive, Adjektive
- Komposita
- Abkürzungen
- mehrere Wörter, die einen Fachbegriff darstellen (Mehrwortbegriffe)
- Zahlen und Masseinheiten

Zunächst wird geprüft, ob Buchstabenfolgen oder Zahlen zu verarbeiten sind. Für Zahlen steht ein entsprechender Algorithmus für die Faktenindexierung bereit. Für die Verarbeitung von Buchstabenfolgen verschiedene Algorithmen für Einwortbegriffe oder Mehrwortbegriffe zuständig. Die Verarbeitung von Buchstabenfolgen erfolgt über einen zeichenweise ablaufenden Vergleich mit insgesamt neun verschiedenen Wörterbüchern nach dem Prinzip des „longest matching“ (vgl. Abb. 9):

- das Wort stimmt mit einem Eintrag im Thesaurus (V1) überein und ist damit ein Deskriptor
- das Wort stimmt mit einem Eintrag im Thesaurus (V1) überein und der Rest wird als Flexionsendung im Verzeichnis V3 erkannt. Das Wort wird Deskriptor, der Rest eliminiert.
- das Wort lässt sich in mehrere Eintragungen im Thesaurus (V1) zerlegen. Die Teilwörter werden Deskriptoren. Mögliche verbleibende Reste werden als Flexionsendungen im Verzeichnis V3 erkannt und eliminiert.

Für die Erkennung und Verarbeitung von Mehrwortbegriffen wird ein gesonderter Algorithmus angewandt. Der Anschlusskonnektor 5 springt in diesen Algorithmus.

Die Indexterme werden in Form eines Zahlencodes abgespeichert. Das Informationssystem (aus der ehemaligen DDR) arbeitete mehrsprachig (deutsch, russisch), die Zahlencodes dienten als Drehscheibe für die jeweilige Übersetzung in jeweils nationale Thesauri der Mitglieder in diesem Dokumentationsverbund.

Der Thesaurus enthält Wörter in ihrer Stammform, bspw. „prüf“. Die Textterme „Prüfen“ oder „Prüfung“ werden erkannt und rückgeführt. „Prüfgerät“ wird als Kompositum erkannt. Komposita werden zerlegt (ausführlich beschrieben in Alliger/Richter 1978).

3.2.2 Beispiel: IDX – Ein wörterbuchbasiertes Verfahren

Bei IDX handelt es sich um ein wörterbuchbasiertes Indexierungssystem, d.h. alle Funktionen beruhen auf Festlegungen in einem oder mehreren Wörterbüchern. IDX ist in funktionaler Hinsicht grundsätzlich mit dem älteren System PASSAT von Siemens (vgl. Gräbnitz 1987) vergleichbar.

IDX (Automatische ... 1997) ist ein Produkt der Firma SOFTEX von Prof. Harald H. Zimmermann. Die Grundlagenentwicklungen wurden an der Fachrichtung Informationswissenschaft der Universität des Saarlandes durchgeführt.

Das System IDX benötigt eine Reihe sehr umfangreicher Wörterbücher, da alle Erkennungs- und Indexierungsfunktionen an entsprechende Einträge und Festlegungen in Wörterbüchern gebunden sind. Auf syntaxstrukturelle Analysen wird zunächst vollends verzichtet.

Grundsätzlich geht IDX von einer Freitextindexierung aus, wobei jedoch eine Reihe informationslinguistischer Funktionen bereitstehen, um auftretende Textwortformen zu bearbeiten. Z.Zt. werden für die Sprachen Deutsch, Englisch, Französisch, Italienisch und Spanisch folgende Funktionen angeboten (Zimmermann 1996, Automatische ... 1997):

- Markierung von Stoppwörtern und ihre weitere Eliminierung aus dem Indexierungsprozess
Bsp.: Funktionswörter wie bspw. der, mit
→ Stoppwortwörterbuch
- Textwortformen werden über ein Deflexionsverfahren auf ihre relevanten Grundformen reduziert
Bsp.: Kinder → Kind; Häuser → Haus; schlugst → schlagen
→ Identifikationswörterbuch
- Komposita werden *zusätzlich* mit ihren sinnvollen Bestandteilen für die Indexierung bereitgestellt (Dekomposition)
Bsp.: Wissensrepräsentation → Wissen, Repräsentation; Haustüren → Haustür, Haus, Tür
→ Identifikationswörterbuch, vorrangig Relationenwörterbuch
- Wortableitungen (Derivationen) werden zusätzlich in ihrer Grundform bereitgestellt
Bsp.: Besichtigung → besichtigen
→ Identifikationswörterbuch, vorrangig Relationenwörterbuch
- Soweit lexikalisiert, identifiziert IDX Mehrwortbegriffe und Wortbindestrichergänzungen
Bsp.: automatische Indexierung; juristische Person; Wirtschafts- und Sozialordnung → Wirtschaftsordnung, Sozialordnung
→ Mehrwort- und Übersetzungswörterbuch, Relationenwörterbuch
- Soweit lexikalisiert, können diskontinuierliche Verbteile ihrem Hauptbestandteil zugeordnet werden
Bsp.: steht ... zur Verfügung → zur Verfügung stehen; kamen ... an → ankommen
→ Mehrwort- und Übersetzungswörterbuch, Relationenwörterbuch

Bei den beiden letztgenannten Funktionen kann IDX durch ein vorgeschaltetes Parsingverfahren auch nicht-lexikalische Wortgruppen identifizieren.

- Wortrelationierungen lassen sich in IDX einbinden, bishin zu der Funktionsweise eines „echten“ Thesaurus

Bsp.: Photo ↔ Foto (Schreibvarianten); Klavier ↔ Piano („echte“ Synonyme); TQM ↔ Total Quality Management (Abkürzungen und Langform)

Über die Integration eines echten Thesaurus sind auch andere Relationierungen möglich, bspw. Hierarchie

→ Relationenwörterbuch

- Für die bereits genannten Sprachen bietet das IDX-Verfahren eine wortbezogene Übersetzung an. Damit sind einsprachige Indexierungen für eine mehrsprachige Dokumentensammlung möglich

Der Ablauf einer Indexierung mit IDX läuft in mehreren Phasen (Textdurchläufen) ab, wobei jede Phase Zwischendateien hinsichtlich der nächsten Phase erzeugt. Welche Phasen (Indexierungsschritte) konkret durchlaufen werden, hängt von Systemkonfiguration des Anwenders ab.

Phase 0: Grundformermittlung auf Einzelwortebene

Der Text wird in der Reihenfolge durchlaufen und wortweise nummeriert. Ebenfalls nummeriert werden Stammwortklassen und Wortformenwortklassen. Wörter, die für die Ermittlung diskontinuierlicher Wortgruppen relevant sein könnten, werden gekennzeichnet. Die Grundformenermittlung erfolgt aufgrund einer morphologischen Analyse.

Die Identifikation der Wörter erfolgt nach dem Prinzip des „longest matching“, d.h. längere Wörter stehen in den Identifikationswörterbüchern vor kürzeren Wörtern, wenn deren Zeichenfolge enthalten ist („Drucker“ steht also vor „Druck“).

Phase M: Mehrwortbegriffe

In diesem Schritt werden kontinuierliche Mehrwortbegriffe anhand des Mehrwortwörterbuchs identifiziert. Wortbindestrichergänzungen werden gekennzeichnet.

Phase B: Ermittlung getilgter Wörter

In dieser Phase erfolgt die Ermittlung getilgter Teilwörter über die in der Phase M vergebenen Kennzeichnungen.

Phase 1: Strukturanalyse, Stoppwortermittlung

In dieser Phase werden Stoppwörter markiert.

Die Strukturanalyse fasst die Ergebnisse der vorausgegangenen Phasen (aus den Zwischendateien) zusammen, ermittelt alternative Lemmanamen (messen, Messe), führt diskontinuierliche Verbalgruppenteile und „feste Wendungen“ zusammen.

Phase 2: Derivation und Dekomposition, Mehrwortkontrolle

In dieser Phase werden die Einträge hinzugefügt, die durch Derivation und Dekomposition als zusätzliche Indexterme gewonnen werden.

Phase X: Alphabetische Sortierung

In diesem Schritt werden die erzeugten Stichwörter alphabetisch sortiert, markierte Stoppwörter, Nichtworteinträge und Dubletten eliminiert.

Phase 3: Aufbau der Indexierungsergebnisdatei

In dieser Phase wird durch Zusammenfassung der vorausgegangenen Schritte das Ergebnis der Indexierung erzeugt und in eine Ergebnisdatei gestellt.

Phase G: Aufbau der Relationendatei

In dieser abschliessenden Phase wird eine Kombination der Indexierungsergebnisse mit den zugehörigen Wortrelationen vorgenommen. Die Relationendatei ist der Abschluss des Indexierungslaufs.

Im Anschluss an den Indexierungslauf kann im Bedarfsfall ein Übersetzungslauf vorgenommen werden.

Das Beispiel einer Indexierung des nachstehenden kurzen Textes, mit anschliessender Wiedergabe der Relationendatei mag die Funktionen von IDX verdeutlichen:

Beispieltext:

Patentschriften stehen im Mittelpunkt des Interesses der Computerhersteller.

Relationendatei mit (auszugsweiser Legende):

1 Patentschriften → Patentschrift <6> :23: Patent <8>
 1 Patentschriften → Patentschrift <6> :23 t: Schrift <6>
 *2 stehen <5>
 *3 im <1>
 4 Mittelpunkt <7> :3 t: Punkt <7>
 4 Mittelpunkt <7> :4: Mitte <6>
 *5 des <1>
 6 Interesses → Interesse <8> :4: interessieren <5>
 6 Interesses → Interesse <8> :4: Interessiertheit <6>
 *7 der <1>
 8 Computerhersteller <6> :23: Computer <6>
 8 Computerhersteller <6> :23 t: Hersteller <6>
 8 Computerhersteller <6> :23 t: herstellen <5>
 9 .

- 1 Wortform (klein)
- 3 Kompositum/Teilwort
Relation
- 4 Derivationsrelation
- 5 Infinitiv (Verb)
- 6 Substantiv feminin
- 7 Substantiv maskulin
- 8 Substantiv neutrum
- 23 Kompositum/Teilwort,
gewonnen durch
Wortzerlegung
- t Ergänzung zur Relations-
angabe, gibt an, dass es sich
um den letzten Bestandteil
des Wortes handelt
- * Markierung von
Stoppwörtern

Das Indexierungssystem IDX wurde im Rahmen zweier grosser DFG-geförderter Projekte – MILOS I und MILOS II – an der Universitäts- und Landesbibliothek Düsseldorf eingeführt und evaluiert (Lepsky 1996a, b; 1998; Sachse et al. 1998).¹⁰ Im Rahmen dieser Anwendung werden bibliothekarische Katalogdaten automatisch indexiert. Bemerkenswert an dieser bibliothekarischen Anwendung ist einerseits der diskursunabhängige Einsatz der zugrundeliegenden Wörterbücher in einem stark heterogenen Bestand einer Universallbibliothek. Weiterhin beeindrucken die positiven Ergebnisse der Evaluierung im Rahmen zweier grosser Retrievaltests (Sachse et al. 1998), deren Zusammenfassung in den beiden folgenden Tabellen wiedergegeben ist¹¹:

Retrievaltest zum Projekt MILOS I:

¹⁰ Über die Homepage der MILOS-Projekte ist u.a. der Zugriff auf Dokumente, Produktbeschreibungen, Testergebnisse sowie den indexierten Bibliothekskatalog möglich: http://www.uni-duesseldorf.de/WWW/ulb/mil_home.htm. Informationen zum Nachfolgeprojekt KASCADE sind unter http://www.uni-duesseldorf.de/ulb/kas_home.htm zu finden.

¹¹ Ausführlich werden die Retrievaltests und ihre Resultate behandelt in Lepsky et al. 1996 und Sachse et al. 1998.

Retrievaltest mit 40.000 Datensätzen und 50 Suchfragen

Methode	Recall	Precision	Einheitswert ¹²
Stichwort (Freitextinvertierung)	14 %	59 %	0.84
Stichwort + Automatische Indexierung (IDX)	51 %	83 %	0.46
Stichwort + RSWK-Schlagwörter ¹³ (Verstichwortet)	39 %	83 %	0.58

Retrievaltest zum Projekt MILOS II:

Retrievaltest mit 190.000 Datensätze und 100 Suchfragen

Methode	"0-Treffer-Erg." ¹⁴	Precision
Stichwort (Freitextinvertierung)	15	0,82
Automatische Indexierung (IDX)	3	0,75
RSWK (Verstichwortet)	30	0,95
Basic Index	0	0,803

Als Fazit der Erforschung informationslinguistischer Verfahren kann festgehalten werden, dass pragmatische Lösungen auf hauptsächlich morphologischer Ebene dominieren. Ihre Einsatzfähigkeit ist durch Retrievaltests unter Beweis gestellt.

Vor allem aber spielen informationslinguistische Verfahren eine wichtige Rolle in Kombination mit anderen Verfahrensansätzen, wobei sie in diesen Fällen eine vorbereitende Rolle einnehmen.

¹² Rijsbergen 1979.

¹³ Die RSWK (Regeln für den Schlagwortkatalog) sind eine in deutschen Bibliotheken verbreitetes intellektuelles Indexierungsverfahren.

¹⁴ Die hohe Zahl sog. „0-Treffer-Ergebnisse“ bei manueller Indexierung nach RSWK waren eine der Gründe für den Test automatischer Indexierung.

3.3 Pattern-Matching-Verfahren

Verfahren automatischer Indexierung, deren hauptsächlicher Analyseansatz auf dem Pattern-Matching (Mustererkennung) beruht, sind sowohl in der Forschung als auch in der Praxis nur wenig verbreitet. Bezogen auf fest umrissene Diskursbereiche, ist ein Pattern-Matching ein mächtiges Analyseverfahren, um Informationen aus Texten zu filtern. Mit einer gewissen Berechtigung könnten die Ansätze des Pattern-Matching auch den informationslinguistischen Verfahren zugerechnet werden, zumal meist eine Kombination mit linguistischen Analysetechniken verfolgt wird (wie im unten beschriebenen System FIPRAN) bzw. die Erkennung von Mustern sich auf sprachliche Muster oder Indikatoren bezieht. Da jedoch mit dem Pattern-Matching ein durchaus anderer Analyseansatz verfolgt wird, sollen diese Verfahren hier gesondert behandelt werden.

Das Pattern-Matching führt einen Abgleich zwischen sprachlichen (Wort-)Mustern aus den vorliegenden Texten mit den Einträgen (Mustern) in einer Wissensbasis aus. Dabei unterscheidet sich dieses Verfahren durchaus von den wörterbuchorientierten informationslinguistischen Verfahren. Zunächst besteht keine Notwendigkeit die Einträge (Schlüssel) in der Wissensbasis auf morphologische Elemente zu beziehen. Zudem sind in der zugrundeliegenden Wissensbasis über die abgelegten Schlüssel hinaus in der Regel eine Reihe von Erkennungsparametern enthalten. Neben der erfolgreichen Mustererkennung – auf der Basis vorliegender Schlüssel –, müssen überdies auch die definierten Erkennungsparameter für ein Matching erfüllt sein.

Das Volltextanalysesystem FIPRAN (Firmen und PRodukt ANalyse) wurde entwickelt für die semi-automatische Auswertung wehrtechnischer Zeitschriftenartikel. Mit dem Einsatz von FIPRAN wird ein Ansatz verfolgt, der über eine reine Indexierung von Texten hinaus reicht. Vielmehr wird versucht, über die Kombination von Pattern-Matching und linguistischen Analysen, Informationen aus den vorliegenden Texten zu filtern. Diese Informationen werden in einer Datenbank abgelegt. Mit FIPRAN werden also aus Texten Informationen gefiltert, die anschliessend in einer Datenbank zur Auswertung bereitgestellt werden. Überdies sind auch Verweise auf den Originaltext angelegt, insofern handelt es sich bei FIPRAN auch um ein Indexierungssystem. Damit wäre FIPRAN eigentlich eher den *Text Mining-Verfahren* (Goeser 1997; Gotthard et al. 1997) zuzurechnen. Da die Grenzen zwischen Indexierungssystemen und Text Mining-Systemen jedoch fließend sind, wollen wir diese Unterscheidung hier nicht treffen.

Die vorrangige Analyseaufgabe von FIPRAN besteht darin, Firmennamen, militärische Organisationen, Produktkategorien, Länder sowie mögliche Relationen zwischen diesen Entitäten (bspw. in der folgenden Form: FIRMA liefert PRODUKT an LAND) in wehrtechnischen Artikeln aufzufinden. Für die Relationen ist die Erkennung von Präpositionen, Artikel und bestimmten Verben von Bedeutung. Die Analysegegenstände sind verschiedenen Patternklassen zugeordnet, die aus Schlüsseln, bestehend aus Zeichenfolgen, sowie einer Reihe von prüfbaren Parametern bestehen. Ein Beispiel nach Volk et al. (1992):

Schlüssel	Wortanfang bündig	Wortende bündig	Über Wortgrenzen hinweg
Länder (dän)	Ja	Nein	Nein
Firmen (Kraus Maffay)	Ja	Ja	Ja
Produktkategorie (flugzeug)	Nein	Nein	Nein
Präposition (von)	Ja	Ja	Nein

Abb. 10: Beispiele für Patternklassen in FIPRAN

Länderschlüssel identifizieren sowohl adjektivische als auch nominale Formen:

dän → Dänemark, dänisch → Dänemark

Über den Schlüssel *flugzeug* wurden etwa *Flugzeuge* oder *Jagdflugzeuge* identifiziert und auf die Produktkategorie *Flugzeug* geschlossen.

Ein ähnliches Analyseverfahren wird von Lexis-Nexis innerhalb des Retrieval-Systems *Freestyle* eingesetzt (Stock 2000). *Freestyle* verfügt über Schlüssellisten für die Erkennung von Firmennamen und Personennamen. Personennamen werden über eine Liste englischer Vornamen erkannt, Firmennamen über Schlüsselbegriffe wie Inc., Ltd., Bros., Corp., oder Corporation.

FIPRAN enthält zur Unterstützung des Pattern-Matching eine auf heuristischen Regeln basierende Komponente zur Erkennung von Textblöcken, die Nominalphrasen oder Präpositionalphrasen entsprechen (Begrenzerverfahren).

Durch die folgenden sieben Regeln werden Blockgrenzen innerhalb von Texten ermittelt:

1. Bei Satzende
2. Bei Semikolon oder Doppelpunkt
3. Vor und nach Verben
4. Vor und nach Hilfsverben
5. Vor Konjunktionen
6. Vor einer Präposition
7. Vor Artikeln, wenn davor keine Präposition steht

Die ermittelte Blockstruktur wird durch drei Regeln ausgewertet:

1. Treten in einem Block Länderschlüssel und Firmenschlüssel auf, so wird darauf geschlossen, dass dieses Land Firmensitz ist.
2. Tritt in einem Block das Schlüsselwort *Firma*, gefolgt von unbekannten Wörtern auf, so wird ein Firmenname angenommen (z.B. ... Firma Siemens Nixdorf ...).
3. Tritt in einem Block ein als Produktkategorie erkanntes Wort auf (z.B. *Jagdflugzeug*), gefolgt von unbekannten Wörtern, so wird auf einen Produktnamen geschlossen (bspw.: ... Jagdflugzeug Fighter 51 ...).

Relationen werden aufgrund auftretender Verben und erkannter Schlüsselwörter identifiziert.

3.4 Begriffsorientierte Verfahren

Einen qualitativ weitergehenden Versuch der automatischen Indexierungsverfahren stellen Ansätze dar, die nicht auf einer Extraktion vorhandener Textterme – mit mehr oder weniger weitreichender Bearbeitung – beruhen. Alle bislang vorgestellten Verfahren sind angewiesen auf eine gegebene Wortwahl im vorliegenden Text. So können bspw. weder statistische noch informationslinguistische Verfahren erkennen, dass es sich bei den Termen „Klavier“ und „Piano“ um die sprachliche Repräsentation *einer Bedeutung* handelt. Informationslinguistische Ansätze erkennen „Klavier“ und „Klaviere“ durch Rückführung der Pluralform als das gleiche Wort und statistische Verfahren berücksichtigen dies bei ihrer Berechnung, eine sprachunabhängige, auf Bedeutungen abhebende Analyse leisten sie jedoch nicht.

Begriffsorientierte Verfahren abstrahieren von der gegebenen Wortwahl vorliegender Dokumente auf die Bedeutung der Texte. Die erkannte Bedeutung (der Inhalt) eines Dokuments wird anschliessend durch Ausdrücke einer kontrollierten Indexierungssprache repräsentiert. Indexierungssprachen können dabei sowohl Thesauri, als auch Klassifikationen sein. Insofern diese Verfahren inhalts- und nicht termzentriert arbeiten, kommen sie einer intellektuellen Indexierung näher als Extraktionsverfahren. Tatsächlich *simulieren* begriffsorientierte Verfahren die Arbeitsweise eines menschlichen Indexierers insofern, als sie versuchen die *Bedeutung* eines Textes zu ermitteln und durch entsprechende Indexterme diese Bedeutung zu repräsentieren (Darstellung des begrifflichen Gehalts eines Dokuments). Da jedoch auch diesen Verfahren ein wirkliches *Verstehen* (eine Inhaltsanalyse im engeren Sinne) vorliegender Dokumente nicht implementiert werden kann, muss letztlich über die Sprachoberfläche auf Bedeutungen geschlossen werden. Für diese Analyseaufgaben wird meist wiederum auf statistische und/oder informationslinguistische Methoden zurückgegriffen. Die *Simulation* eines menschlichen Indexierers ist damit lediglich eine Simulation des Arbeitsergebnisses, nicht jedoch des Arbeitsprozesses zur Erreichung dieses Ergebnisses.

Die Korrelationsannahme zwischen sprachlicher Ausdrucksweise und der Bedeutung des Ausgesagten, gewinnt insbesondere bei diesen Verfahren an Gewicht, da sie explizit eine Repräsentation von Dokumenten*inhalten* anstreben auf der Grundlage sprachoberflächlicher Analysen. Die moderne sprachwissenschaftliche Position geht hingegen von der Annahme aus, die Bedeutung von Wörtern könne nur aus dem Kontext ihres jeweiligen Gebrauchs erschlossen werden (Crystal 1995, S. 102). Analyseverfahren müssten im Idealfall daher den Kontext auftretender Wörter berücksichtigen. Statistische Analysen erfüllen diesen Anspruch gar nicht. Informationslinguistische Methoden allein – auch weiterführende syntaxanalytische Ansätze – können kontextbedingte Bedeutungsanalysen nicht hinreichend leisten (Reimer 1992).

Im Bereich der begriffsorientierten Verfahren wird daher auch in eine wissensbasierte Richtung geforscht. Diese Analysemodelle aus dem Forschungsumfeld der Künstlichen Intelligenz zeichnen sich durch die Einbeziehung von Weltwissen aus und sind zudem in der Lage, ihre Wissensbasis selbst zu erweitern (Wissensakquisition). Ihr Einsatz reduziert die Notwendigkeit der problematischen syntaktischen Vollanalysen. Praktisch sind diese Ansätze heute noch mit dem Nachteil behaftet, nur für begrenzte und homogene Diskursbereiche geeignet zu sein, da die diskursunabhängige Implementierung von Weltwissen bislang nicht gelungen ist (Lehmann 1988; Görz 1991). Bereits für kleine Themengebiete sind diese Systeme extrem aufwendig. Gleiches gilt für Methoden des „maschinellen Lernens“ in diesem offenen Kontext. In der Praxis spielen diese Ansätze daher z.Zt. kaum eine Rolle. Ein Indexierungssystem dieses Typs ist TCS (Text Categorization Shell) (Knorz 1994). Im Bereich des Text Summarization folgt das System FRUMP (Endres-Niggemeyer 2000, S. 313-314) einem wissensbasierten Ansatz.

In der Folge wollen wir uns daher auf einige „pragmatische“ Ansätze und ihre jeweiligen Möglichkeiten beschränken. Ein probabilistisches Modell liegt dem Ansatz AIR/X zugrunde. Als eine erfolgreiche kommerzielle Anwendung ist vor allem das „categorization system“ InfoSort von Profound zu nennen (Maller 1998; Stock 2000). Alle Profound Informationsdatenbanken werden mit InfoSort erschlossen, Agenturmeldungen und Zeitungsartikel ausschliesslich automatisch. Ein Beispiel für die Zuordnung von Dokumenten (in diesem Falle WWW-Dokumente) zu Klassen eines Klassifikationssystems ist das System GERHARD. AIR/X und GERHARD werden nachfolgend ausführlicher beschrieben.

3.4.1 Beispiel: Das Verfahren AIR/X

AIR/X ist ein probabilistisches Indexierungsmodell, entwickelt zwischen 1978 und 1985 an der TH Darmstadt unter der Leitung von Gerhard Lustig (Lustig 1986, 1989; Knorz 1994; Biebricher et al. 1988). Die Forschungsansätze dieses Modells gehen jedoch bereits zurück bis in die 60er Jahre (Lustig 1969). Eine Pilotanwendung wurde seit 1985 unter dem Namen AIR/PHYS beim Fachinformationszentrum Karlsruhe zur Indexierung der Datenbank PHYS (heute Teil von INSPEC) betrieben.

Neben der automatischen Indexierung englischsprachiger Abstracts, beinhaltet der AIR-Ansatz auch einen weitgehend automatischen Aufbau der zur Indexierung benötigten Wörterbücher. Voraussetzung dafür ist, dass bereits eine umfangreiche Kollektion mit manuell indexierten Dokumenten vorliegt, aus dem das automatische Verfahren „lernen“ kann.

Eine Indexierung läuft nach dem AIR-Ansatz im wesentlichen in den folgenden Schritten ab:

1. Die Erkennung aller in einem Abstract enthaltenen Terme, die einen relevanten Begriff darstellen könnten.
2. Die Auswahl der Terme, die tatsächlich einen relevanten Begriff darstellen.
3. Die Repräsentation dieser Begriffe durch Deskriptoren des Thesaurus.

Im ersten Schritt werden alle auftretende Terme eines Abstracts mit einem Wörterbuch verglichen. Das Wörterbuch von AIR besteht aus Termen (Einzelwörter und Mehrwortgruppen) und dem kontrollierten Vokabular eines Thesaurus (Deskriptoren). Zwischen den Termen und den Deskriptoren bestehen verschiedenartige Relationen:

- Deskriptor-Deskriptor-Relation
- Term-Deskriptor-Relation (Use)
- Term-Deskriptor-Relation auf Grundlage der statistischen Relation Z (s.u.)

Im Zuge der Indexierung eines Dokuments wird zunächst für jeden Term des Textes geprüft, ob im Wörterbuch eine Relation auf einen Deskriptor eingetragen ist. Diese Informationen werden gesammelt. Nicht jede Relation zwischen einem Term und einem Deskriptor bedingt automatisch die Deskriptorenzuteilung. Im 2. Indexierungsschritt werden nicht-relevante Begriffe erkannt. Die Zuteilungsentscheidung beruht auf einer Berechnung der *Wahrscheinlichkeit* (probabilistischer Ansatz), dass ein Deskriptor zuzuteilen ist, wenn ein Term im Dokument auftritt. An diesem Ansatz wird die Problematik der Simulation eines menschlichen Indexierers deutlich: Ein Indexierer entscheidet nach einer auf einem Verstehensprozess basierenden Inhaltsanalyse, während AIR diese Entscheidung am Vorhandensein bestimmter Textterme festmacht. Die Analyse beruht auf statistischen und heuristischen Ansätzen (Lustig 1989), nicht auf dem Verstehen eines Dokuments. Die wichtigste Analysefunktion ist dabei die statistisch ermittelte

Relation Z. Folgende Beschreibung ist übernommen von Lustig (1989, S. 142) (vgl. auch Schwantner 1987):

Lustig nennt als entscheidende Funktion

die Generierung der gewichteten Relation Z, die beliebige – d.h. einfache oder zusammengesetzte – Fachausdrücke mit Deskriptoren verbindet. Sie beruht auf dem auf einer möglichst grossen Menge D intellektuell indexierter Dokumente berechneten Assoziationsfaktor

$$z(t,s) = \frac{h(t,s)}{f(t)}$$

wobei

$f(t)$ die Anzahl der Dokumente, in deren Referatetext der Fachausdruck t vorkommt, und

$h(t,s)$ die Anzahl derjenigen unter diesen Dokumenten, denen der Deskriptor s intellektuell zugeteilt ist,

bezeichnet.

$z(t,s)$ kann als Näherung für die bedingte Wahrscheinlichkeit, dass ein bestimmter Deskriptor s zuzuteilen ist, wenn ein Term t auftritt, interpretiert werden. Z-Werte, für die $h(t,s)$ bzw. $z(t,s)$ sehr klein ist, sind statistisch sehr unsicher bzw. tragen zur Indexierungsentscheidung kaum bei.

In einer Beschreibung von G. Lustig aus dem Jahre 1969 (S. 250) wird die Bezugnahme auf frühere Forschungen im Rahmen einer europäischen kerntechnische Dokumentation deutlich:

Dieser Ansatz wurde von CETIS¹⁵ wie folgt verallgemeinert. Man geht aus von einer Kollektion von Referaten und den zugehörigen manuellen Schlagwortzuteilungen im System ENDS¹⁶. Für einen beliebigen in den Referaten vorkommenden Ausdruck E und ein beliebiges Schlagwort D bezeichnen

- $f(E)$ die Anzahl der Referate, in denen E vorkommt und

- $h(E,D)$ die Anzahl der Referate, denen *ausserdem* der Ausdruck D als Schlagwort zugeteilt worden ist.

Dann wird durch

$$z(E,D) = \frac{h(E,D)}{f(E)}$$

angenähert die Wahrscheinlichkeit dargestellt, dass im EURATOM-System das Schlagwort D zugeteilt wird, wenn das Wort E in dem Referat vorkommt. Mann kann dann für alle hinreichend grossen Werte von $z(E,D)$ eine Relation $E \rightarrow D$ in das

¹⁵ Europäische Forschungsanstalt für wissenschaftliche Datenverarbeitung.

¹⁶ EURATOM Nuclear Documentation System.

Wörterbuch aufnehmen und einen in dem Referat identifizierten Ausdruck E_0 als relevant ansehen, wenn es wenigstens eine solche Relation $E_0 \rightarrow D$ gibt.

Auf dieser Relation Z beruhen in der Pilotanwendung des AIR-Verfahrens, AIR/PHYS, rund 57% der Wörterbucheinträge!

Die Pilotanwendung von AIR/PHYS

Die Pilotanwendung von AIR ist die Indexierung für die Datenbank PHYS (heute Teil von INSPEC) beim Fachinformationszentrum Karlsruhe. Dabei wurden zunächst in einem Pilotprojekt, anschliessend im Routinebetrieb, auf der Basis englischsprachiger Abstracts Dokumente aus dem Fach Physik für die Datenbank PHYS indexiert. Die Pilotphase wurde durch einen Retrievaltest begleitet (s.u.).

Das Indexierungswörterbuch enthält 200.000 Terme (einschl. der 22.700 Deskriptoren des Thesaurus), 190.000 Relationen zwischen zwei Deskriptoren und 620.000 Relationen zwischen beliebigen Termen und den Deskriptoren, ermittelt hauptsächlich über die Relation Z .

Um die Relation Z berechnen zu können, wird eine grosse Anzahl bereits manuell indexierter Dokumente benötigt (s.o.). Für die Pilotanwendung AIR/PHYS gingen 400.000 Dokumente in die Auswertung ein. Von den 22.700 Deskriptoren des Thesaurus konnten auf dieser Grundlage jedoch nur für ca. 10.000 Deskriptoren die entsprechenden Z -Werte berechnet werden! Damit ist für mehr als die Hälfte der Deskriptoren die wichtigste Relationierung nicht ermittelt (Schwantner 1987).

Der Indexierungsvorgang

Der Indexierungsprozess besteht aus einem *Beschreibungsschritt* und einem anschliessenden *Entscheidungsschritt*. Der Beschreibungsschritt besteht aus der Textaufbereitung durch die Eliminierung von Stoppwörtern, eine Grundformenermittlung und Mehrworterkennung, einer Formelerkennung und -transformation (Formeln werden in normierte Terme umgesetzt) sowie der Erstellung von Relevanzbeschreibungen. Im zweiten Schritt, dem Entscheidungsschritt, werden aus der Relevanzbeschreibung die Gewichtungen errechnet. Liegt die Gewichtung für einen Deskriptor über einem zu definierenden Schwellenwert, so wird er zugeteilt.

Da alle Dokumente im Anschluss an die automatische Indexierung intellektuell klassiert werden, findet bei dieser Gelegenheit eine Überprüfung der automatischen Indexate statt. AIR/PHYS erstellt durchschnittlich 12 Deskriptoren pro Dokument, gegenüber 9 Deskriptoren die vor seinem Einsatz auf manuellem Wege zugeteilt wurden (Schwantner 1987).

Etwa ein Drittel der einem Dokument zugeteilten Deskriptoren werden in dieser Nachbearbeitung gestrichen, da es sich um Fehlzuteilungen handelt. Etwa gleichviel Deskriptoren werden durch den Indexierer neu zugeteilt. In der Praxis der Pilotanwendung ist AIR damit als semi-automatisches Verfahren eingesetzt.

Der Retrievaltest

Das Pilotprojekt wurde durch einen Retrievaltest begleitet, wobei einer Kollektion von 15.000 Dokumenten aus PHYS 300 Originalfragen unterzogen wurde (Lustig 1986, 1989). Dabei wurden jeweils die Werte für Recall (r) und Precision (p) der automatischen Indexierung und der intellektuellen Indexierung ermittelt:

Durchschnittswerte der *automatischen* Indexierung:

$$p = 0,46; r = 0,57$$

Durchschnittswerte der *intellektuellen* Indexierung:

$$p = 0,53; r = 0,51$$

AIR/dpa

Die Anpassung des AIR-Ansatzes im beschriebenen Umfang scheint für andere Anwendungen kaum möglich, da ein zu hoher Aufwand für die Vorbereitung der Indexierung notwendig ist.

Reduzierte Voraussetzungen könnten den Ansatz jedoch auch für weitere Anwendungen interessant werden lassen. In einer Untersuchung wurde eine sehr grobe Kategorisierung von Nachrichtmeldungen der dpa erprobt (Brilmayer et al. 1997). Dabei waren lediglich 40 Klassen zu berücksichtigen, bspw.:

INLA - Inland	APOL - Aussenpolitik	SOZP - Sozialpolitik
IPLO - Innenpolitik	MILT - Militär	HIST - Geschichte
JUST - Justiz	INDU - Industrie	WETT - Wetter
KULT - Kultur	ENER - Energie	WIFI - Wirtschaft, Finanzen

Einem Dokument können mehrere Klassen zugeteilt werden.

Da zudem zunächst nur Substantive, Adjektive, Verben und Eigennamen als Terme berücksichtigt wurden, hielt sich der Aufwand für die Berechnung der Z-Relation für das Wörterbuch in engeren Grenzen. Die Dokumente wurden mit dem Programm GERTWOL (Haapalainen/Majorin 1995), einem System für die morphologische Analyse der deutschen Sprache, insbesondere für die Wortformerkennung, analysiert. Ein Beispiel für die Indexierung:

Mehrheit der Brandenburger hält PDS für regierungsfähig

Potsdam (dpa) - Gut zwei Drittel der Brandenburger halten die PDS für regierungsfähig, auch wenn 59 Prozent sie niemals wählen würden. Dies ergab eine Infas-Umfrage im Auftrag der „Märkischen Allgemeinen“ im Dezember 1995 unter 500 Brandenburgern. Bemerkenswert ist, dass die PDS bei jüngeren Wählern deutlich besser ankommt als bei älteren: Während der Umfrage zufolge 75 Prozent der Brandenburger über 65 Jahre die PDS niemals wählen würden, sind dies bei den bis zu 34jährigen nur 57 Prozent und bei den 35- bis 65jährigen sogar nur 56 Prozent.

Die PDS werde von 61 Prozent als Integrationsfaktor in Ostdeutschland für notwendig gehalten. Von den Befragten meinen zudem 52 Prozent, die PDS werde von den anderen Parteien unfair behandelt.

Extrahierte Terme und ihre informationslinguistische Analyse (Auswahl) und Aufbereitung:

ostdeutschland: ost ostdeutschland deutsch land
deutschland

Parteien: partei

PDS: pds

regierungsfähig: regierungsfähig regierung

Das Ergebnis der automatische Indexierung:

	Deskriptor	Gewichtung
Schwellenwert: 0.5	INLA	0.649819
	IPOL	0.548205
	PART	0.385417
	PERS	0.312776
	JUST	0.230031

Die intellektuelle Indexierung des Beispieldokuments bei der DPA ergab eine Zuordnung zu folgenden drei Klassen: INLA – IPOL – PART. Eine automatische Indexierung mit AIR ermittelt in diesem Beispiel die gleichen drei Kategorien mit der höchsten Gewichtung, PART wurde lediglich aufgrund des definierten Schwellenwertes nicht zugeteilt.

3.4.2 Beispiel: GERHARD

GERHARD¹⁷ (GERman Harvest Automated Retrieval and Directory) ist ein Web-basiertes Informationssystem für den Nachweis deutscher wissenschaftlicher WWW-Seiten (z.Zt. nur HTML-formatierte Dokumente). Die nachgewiesenen Seiten werden automatisch einer oder mehreren Klassen der Universellen Dezimalklassifikation (UDK) zugeordnet. Such- und Navigationsprozesse basieren auf der UDK. Die folgende Darstellung der automatischen Indexierungsprozesse in GERHARD folgt einer Diplomarbeit von Carmen Krüger (1999).

Die automatische Indexierung in GERHARD basiert auf informationslinguistischen und statistischen Methoden. Ziel dabei ist, den Inhalt der Dokumente auf die UDK abzubilden und damit eine Klassifikation der Dokumente zu erreichen.

Der Indexierungsprozess besteht aus folgenden Verfahrensschritten:

- Erstellung eines UDK-Lexikons
- Aufbereitung der zu indexierenden (klassierenden) Dokumente
- Analyse der Notationen

Die UDK wurde dabei zunächst für eine Abbildbarkeit der Dokumente auf die Einträge überarbeitet. Dafür waren bspw. Klassenbenennungen der Form „Übersetzungen / Technische u. naturwissenschaftliche“ zu wandeln in „Technische und naturwissenschaftliche Übersetzungen“, Aufzählungen waren auf Einzelbegriffe zurückzuführen, die auf ihre entsprechende Notation verweisen. Umlaute wurden normiert sowie eine einheitliche Kleinschreibung eingeführt.

Einträge im UDK-Lexikon haben die folgende Form:

Natürlichsprachlicher_Schlüssel Trennsymbol Notation

Beispiele (# = Trunkierungszeichen)

Esperanto:=089.2

Umwelt# frau#:396,5.00.504

¹⁷ <http://www.gerhard.de>

Die zu indexierenden Texte werden zunächst analog der UDK-Einträge aufbereitet (Kleinschreibung, Umlaute). Anschliessend werden die Terme des Dokuments durch ein iteratives look-up von Präfixen analysiert. Dabei wird jeweils der längste Präfix gesucht und als Ergebnis geliefert. Die Textanalyse liefert so eine bestimmte Anzahl von Übereinstimmungen mit dem UDK-Lexikon und ermittelt auf diese Weise passende Notationen für das Dokument. Diese werden – zusammen mit Angaben über die Häufigkeit des Auftretens der jeweiligen Begriffe – an die auf statistischen Methoden basierende Analyse der Notationen weitergegeben. Dabei wird die strukturelle Transparenz der Notationen der UDK ausgenutzt. Die Sicherheit der Zuordnung eines Dokuments zu einem Themenbereich steigt mit der Anzahl der vorliegenden Notationen mit einem gemeinsamen „Notations-Präfix“. Je länger dieser Präfix ist, desto spezifischer ist die Klassifizierung. Beide Faktoren werden miteinander verrechnet.

Notationen, die aufgrund der Titelanalyse vergeben werden, gehen mit einem höheren Relevanzwert in die Berechnung ein, als Notationen, die aus der Analyse des restlichen Dokuments gewonnen werden. Der Relevanzfaktor gibt an, wie exakt die Zuordnung eines Dokuments zu einer Klasse der UDK ist. Durch dieses Ranking wird erreicht, dass die für eine Klasse relevanten Dokumente vor den weniger relevanten positioniert werden.

Ein unabhängiger, am Studiengang Informationswirtschaft der Fachhochschule Stuttgart durchgeführter, Retrievaltest¹⁸ ergab, dass 83,75 % der Dokumente durch die automatische Indexierung der richtigen Klasse der UDK zugeordnet werden. Dabei wiesen die 20 untersuchten Klassen jedoch erhebliche Schwankungen in der zuverlässigen Klassifizierung auf.

¹⁸ Der Retrievaltest wurde von Carmen Krüger im Rahmen einer Diplomarbeit durchgeführt, vgl. Krüger 1999.

4. Keyphrase Extraction

Eine Zwischenstufe auf dem Weg von der automatischen Indexierung hin zur automatischen Generierung textueller Zusammenfassungen (Automatic Text Summarization) stellen Ansätze dar, die Schlüsselphrasen aus Dokumenten extrahieren (Keyphrase Extraction). Die Grenzen zwischen den automatischen Verfahren der Indexierung und des Text Summarization sind fließend. Die Extraktion von Keyphrases kann sowohl der Indexierung dienen, als auch der Aufgabe, Dokumente in einem Abstract zusammenzufassen. Entsprechende Software kann häufig für beide Aufgaben herangezogen werden.

Text Summarization (Mani/Maybury 1999, Endres-Niggemeyer 1994, 1998) in einem fortgeschrittenen Sinne strebt die Zusammenfassung und Wiedergabe des Bedeutungsgehalts eines Dokuments durch einen kohärenten Text an. Sparck Jones (1999) charakterisiert diesen Ansatz als *fact extraction*. Die Wiedergabe der Fakten, die im Originalbeitrag dargestellt werden, erfolgt durch die Generierung eines neuen, kohärenten und zusammenfassenden Textes. Diese Ansätze sind mit einer Vielzahl komplexer Probleme behaftet, bspw. des automatischen „Textverstehens“ und der Produktion von Texten. Dem Summarization-Prozess als einem besonderen kognitiven Verstehens- und Kommunikationsprozess (Endres-Niggemeyer 1998) wird in diesen Ansätzen jedoch versucht Rechnung zu tragen.

Zusammenfassungen, die hingegen auf der Basis von *text extraction* beruhen, versuchen durch unterschiedliche Analyseschritte Schlüsselsätze, keyphrases oder Topic-Sätze im Originalbeitrag zu identifizieren, um diese anschließend extrahieren zu können. Die Analyseverfahren beruhen auf formalen, nicht inhaltlichen Ansätzen. Die Zusammenfassungen entstehen durch die Aneinanderreihung der ermittelten und extrahierten Textpassagen. Kohärente Texte entstehen auf diese Weise nur im begrenzten Sinne. Diese Form des Text Summarization bzw. Abstracting ist der historisch ältere und in der Praxis weitgehend verbreitete Ansatz. Die Forschung verfeinert diesen Verfahrensansatz weiterhin (Endres-Niggemeyer 1998, S. 333)

Diesen klassischen Ansatz bietet bspw. Microsoft's Textverarbeitungsprogramm Word ab der Version 97 mit der Funktion AutoZusammenfassung an (Turney 1997). Diese Funktion identifiziert „wichtige Sätze“ und gibt diese als automatisch erstellte Zusammenfassung aus. Der Anwender kann dabei festlegen, welchen Umfang diese AutoZusammenfassung annehmen soll (prozentual im Verhältnis zum Original oder als Anzahl der Sätze).

Das folgende Beispiel einer Zusammenfassung ist mit der Funktion AutoZusammenfassung von Word erzeugt worden:

Because of that an important element in most knowledge management programs is the identification of personal and organizational knowledge.

To identify knowledge it is necessary to create a codified and organized form of it. Knowledge codification is the representation of knowledge such it can be accessed by each member of an organization.

An excellent way to codify knowledge is to visualize it. Visualizing knowledge of an organization leads to knowledge maps (see knowldgWORKS News, Volume 1 Number 5). „What knowledge is important to do your companies work?“ is the starting question of each knowledge mapping project.

The creation of knowledge maps isn't an information technology project mainly! First of all it is a project of analyzing and systematizing knowledge resources and knowledge driven processes. To identify important knowledge and knowledge-based processes in your company is the starting point of each knowledge codification project.

Knowledge cartographers should pay attention to how knowledge is categorized.

In addition to the guiding function knowledge maps also support the identification of knowledge gaps in a company.
Not collecting and storing but using knowledge is the aim.

Obiger Text ist die AutoZusammenfassung (25% des Originalbeitrages) eines kurzen Artikels mit dem Titel „Knowledge Codification“¹⁹. Das Summary vermittelt einen guten Eindruck über den Inhalt des Originals, offensichtlich lässt der Text jedoch sowohl in Hinsicht auf Kohäsion als auch Kohärenz zu wünschen übrig. Besonders augenfällig wird dies bereits im ersten Satz der Zusammenfassung („Because of that ...“).

Diese Form der Zusammenfassung textueller Dokumente beruht auf formalen Analyseschritten. Erste statistische Ansätze gehen wiederum auf H.P. Luhn (1958) zurück (Endres-Niggemeyer 1998, S. 304-306). Aufbauend auf statistische Indexierungsansätze (bspw. dem Termhäufigkeitsansatz), wird die Konzentration signifikant häufiger Terme in einem Satz ermittelt. Eine solche Ermittlung bedarf wiederum einer linguistischen Unterstützung, da eine Reduktion auf Grund- oder Stammformen für die statistischen Häufigkeitsanalysen auch in dieser Anwendung das Resultat bedeutend verbessern (bereits Luhn sah ein stemming vor!). Sätze mit einer hohen Konzentration solcher Terme gelten als signifikant für den Inhalt des Dokuments. Diese Sätze werden aus dem Original *extrahiert* und in das Summary eingestellt.

Andere bzw. ergänzende Extraktionsmethoden gründen ihre Analyse auf bestimmte Indikatoren die geeignet sind, signifikante Sätze zu identifizieren. So werden bspw. Schlüsselwörter oder Signalwörter definiert oder bestimmte Phrasen als Indikatoren angesehen. Solche Phrasen könnten bspw. sein: „Der Zweck dieses Artikels ...“, „Dieser Aufsatz behandelt ...“ oder „This paper reviews ...“. Sätze, die diese Indikatoren enthalten, werden extrahiert. Als Schlüsselwörter am Beginn eines Satzes können bspw. „Zusammenfassend ...“ oder „Finally ...“ definiert werden. Statistische Häufigkeitsansätze und die Identifizierung von Indikatoren können bei der Extraktion in ergänzender Weise eingesetzt werden.

Mit begrenzter Zuverlässigkeit lassen sich auch bestimmte „Topic-Sätze“ aus Texten extrahieren. Dabei wird aus der Stellung eines Satzes im Text auf seine inhaltliche Bedeutung geschlossen: Kapitelüberschriften, Kapitelanfänge, Kapitelende, Bildunterschriften usf. werden aufgrund ihrer Stellung als signifikanter für den Inhalt eines Textes angesehen. Diese Form der Auswertung bedarf jedoch einer Auszeichnungssprache für die Struktur eines Dokuments wie sie bspw. durch SGML, HTML oder XML angeboten wird. Ausgezeichnete Dokumente können strukturabhängig ausgewertet werden, d.h. einzelne Elemente eines Dokuments können identifiziert werden. Dieser Auswertung liegt die Annahme zugrunde, dass bestimmte Passagen von Texten eine in inhaltlicher Hinsicht höhere Bedeutung und Aussagekraft haben als andere. So wird bspw. angenommen, dass am Anfang eines Abschnitts oder Kapitels das Thema eingeführt wird und am Ende von Kapiteln eine Zusammenfassung gegeben wird.

Extrahierte Sätze werden nach diesen Verfahren in der Reihenfolge ihres Auftretens im Originaldokument ausgegeben, sie stehen allerdings meist unverbunden zusammen. Summaries in diesem Sinne fehlt weitgehend die Textkohärenz.

Extraktionssysteme der jüngeren Generation sind z.T. lernfähige Systeme, d.h. sie besitzen die Fähigkeit aus der Analyse von Dokumenten ihre Extraktionsfähigkeit zu verbessern. Diese Funktion ermöglicht es Anwendern, Extraktionssysteme anhand typischer Textkorpora, intellektuell erstellter Abstracts sowie Relevanzentscheidungen hinsichtlich ihrer konkreten Bedürfnisse zu

¹⁹ Veröffentlicht unter <http://www.hbi-stuttgart.de/nohr/publ/KWN.pdf>

trainieren. Ein typisches und in seiner Funktionalität ausgereiftes Beispiel für solche Systeme ist der Extractor des National Research Council of Canada (NRC) (Turney 1999, 2000)²⁰.

Automatisches Abstracting als ein spezielles Anwendungsfeld für Verfahren des Automatic Text Summarization wird u.a. von Kuhlen (1989) und Paice (1990) behandelt. Das menschliche Referieren kann kaum als Vorbild für automatische Verfahren dienen, da dieser kognitive Prozess wohl letztlich nicht abbildbar ist.

Alle neueren und weitergehenden Ansätze unter Einbeziehung wissensbasierter Verfahren sowie kognitionswissenschaftlicher Erkenntnisse (Kuhlen 1989, Endres-Niggemeyer 1994), sind bislang ohne wirklichen Erfolg hinsichtlich einer Umsetzung in praktische Anwendungen geblieben.

4.1 Beispiel: NRC's Extractor

Der Extractor erstellt aus einem Input-Dokument Indexterme sowie eine Liste mit Keyphrases (Schlüsselsätzen). Der Algorithmus lernt anhand von Beispielen und Relevanzbeurteilungen der Nutzer. Das System Extractor generiert aus einem Text mit dem Titel „Technisierung von Wissen – eine Herausforderung für die Technikfolgenforschung?“²¹ folgende Indexterme: Geisteswissenschaften, Technisierung, Technikfolgenforschung, Technik, Wirkungen, Bestimmung, Ansatz.

Anschliessend werden die folgenden extrahierten Textpassagen ausgegeben:

- ◆ Technisierung von Wissen – eine Herausforderung für die **Technikfolgenforschung**?
- ◆ Sehr allgemein lässt sich – je nach **Ansatz** und Auftrag – das Ziel von Technikfolgenforschung definieren als die Steuerung, die Überwachung oder die Eindämmung technischer Innovation, basierend auf Früherkennung bzw. Prognosen möglicher **technischer Wirkungen** und Folgen.
- ◆ Eine **genaue Bestimmung** der Aufgaben und Methoden der Technikfolgenforschung bleibt der jeweiligen Forschungsrichtung bzw. dem konkreten Auftrag überlassen und kann nur in diesem Kontext erfolgen.
- ◆ **Technik** wird hier in ihrer Wirkung auf soziale Umwelten untersucht.
- ◆ Beiden aussertechischen Richtungen der Technikfolgenforschung ist nicht selten eine Technikferne eigen – insbesondere aber den **Geisteswissenschaften**.
- ◆ In ihrer bisher stärksten Form ist diese **Technisierung** von Wissen anzutreffen bei den sogenannten wissensbasierten („intelligenten“) Informationssystemen, sie aber sind nur vorläufiger Höhepunkt einer langen Entwicklung.

Der Extractor-Algorithmus bearbeitet Dokumente in 10 Schritten (Turney 2000 für eine ausführlichere Beschreibung des Algorithmus):

1. Find Single Stems

Zunächst wird eine Liste mit allen Wörtern des Textes erstellt unter Eliminierung von Wörtern mit weniger als drei Zeichen sowie Stoppwörtern. Anschliessend werden die Wörter auf ihre Stammform rückgeführt.

2. Score Single Stems

²⁰ Der Extractor steht unter <http://extractor.iit.nrc.ca> in einer Demoversion zur Verfügung, die auch deutsche Texte bearbeiten kann.

²¹ Im Internet unter <http://www.hbi-stuttgart.de/nohr/publ/technik.htm>

Für jede Stammform wird die Häufigkeit ihres Auftretens im Text ermittelt. Gleichfalls wird die Position ihres ersten Auftretens ermittelt. Zur Errechnung der Termgewichtung wird die Häufigkeit mit einem definierten Faktor multipliziert.

3. Select Top Single Stems

Über die ermittelte Gewichtung wird die Liste der Stammwörter einem Ranking unterzogen. Ein Schwellenwert dient der Ermittlung geeigneter Stammwörter für die weitere Bearbeitung.

4. Find Stem Phrases

In diesem Schritt wird eine Liste aller Phrasen des Textes erstellt. Phrasen sind per Definition Wortfolgen von einem, zwei oder drei Wörtern, die im Text aufeinander folgen ohne von Stoppwörtern oder Interpunktion unterbrochen zu sein. Anschliessend werden die Wörter der Phrasen auf ihre Stammform rückgeführt.

5. Score Stem Phrases

Für jede Phrase (in Stammform) wird die Häufigkeit ihres Auftretens sowie die Position ihres ersten Auftretens im Text ermittelt. Anschliessend wird eine Gewichtung errechnet, analog zum Schritt zwei.

6. Expand Single Stems

Für jedes Stammwort aus der Rankingliste (Schritt 3) wird die höchstgewichtete „Stammwort-Phrase“ ermittelt. Die somit erstellte Liste der Phrasen wird anhand der Gewichtungen aus Schritt zwei einem Ranking unterzogen. Die Gewichtung der „Wortstamm-Phrasen“ (Schritt 5) wird damit ersetzt durch die Gewichtungen, die sich durch die enthaltenen einzelnen Stammwörter ergeben (ermittelt in Schritt 2)

7. Drop Duplicates

Die Rankingliste der „Wortstamm-Phrasen“ kann durch das beschriebene Verfahren Dubletten enthalten. Nur die höchstgewichtete Phrase wird beibehalten, andere Einträge in der Liste werden eliminiert.

8. Add Suffixes

Für jede „Wortstamm-Phrase“ aus der Rankingliste wird die häufigste „Vollphrase“ im Text ermittelt und eingesetzt. Beispiel: „evolu psych“ korrespondiert mit „evolutionary psychology“ (10mal im Text) und „evolutionary psychologist“ (3mal im Text).

9. Add Capitals

In diesem Schritt werden Grossschreibungen für Ausgabe eingesetzt.

10. Final Output

Ausgabe der vorliegenden Rankingliste. Das Ranking der Phrasen basiert auf der Errechnung der Gewichtung der höchstgewichteten einzelnen Wortstämme.

5. Indexierung und Retrievalverfahren

Indexierung ist ein Teilprozess des Information Retrieval, d.h. automatische Indexierungsverfahren können nicht als geschlossener Prozess betrachtet werden. Vielmehr sind Indexierung und Recherche als aufeinander abzustimmende Teilprozesse anzusehen, die – sollen sie optimale Resultate erzielen – einem gemeinsamen und aufeinander abgestimmten Modell des Information Retrieval unterworfen sein müssen.

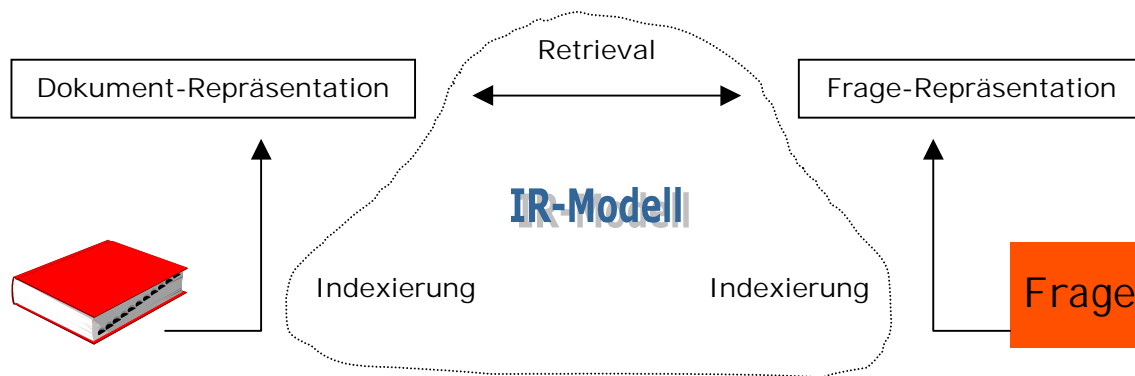


Abb. 11: Indexierung und Retrieval im IR-Modell

Während intellektuelle Inhaltserschließung traditionell in Zusammenhang mit sog. „exact-match-Retrievalverfahren“ (meist unter Anwendung der Booleschen Algebra) diskutiert und eingesetzt wird, ermöglicht und benötigt eine automatische Indexierung ein „intelligenteres“ Retrieval-System (Lustig 1989, Nohr 1991). Werden, wie beispielsweise beim Einsatz von IDX, die vorkommenden Wortformen bei der Indexierung auf ihre Grundform reduziert, Mehrwortbegriffe behandelt, Komposita zerlegt usw., so ist damit die Notwendigkeit verbunden, diese Analyse- und Behandlungsverfahren in entsprechender Weise auf die Anfrageformulierungen der Benutzer anzuwenden. Wir haben hier das Prinzip der Gleichbehandlung von Dokumenten und Suchanfragen innerhalb eines IR-Systems (Abb. 11).

Auf der anderen Seite ermöglicht erst die Anwendung automatischer Indexierungsverfahren bestimmte fortschrittliche Retrievalverfahren. Erst die Implementierung gewichteter Indexierung auf statistischer Grundlage (siehe Kap. 3.1) kann die Einführung von „best-match-Retrievalverfahren“ hinreichend und sinnvoll unterstützen. „Exact-match-Retrievalverfahren“ teilen die Dokumentensammlung – ohne jede Zwischenstufen – in zwei diskrete Untermengen: in Dokumente, die den „exact match“ erfüllen (= relevante Dokumente), und solche, die es nicht tun (= nicht-relevante Dokumente). Dokumente mit drei enthaltenen Termen werden in einer aus vier, mit dem logischen Operator UND verknüpften Termen bestehenden Suchanfrage genauso zurückgewiesen, wie solche mit 0.

Dagegen werden bei „best-match-Verfahren“ die durch eine Suchanfrage nachgewiesenen Dokumente in einer Rangfolge ausgegeben, die der Ähnlichkeit der Dokumente mit der Suchanfrage entspricht („relevance ranking“). In dieser Sicht gibt es nur besser oder schlechter passende Dokumente hinsichtlich einer Suchfrage. Vektorraumorientierte Ansätze behandeln eine Frageformulierung lediglich wie ein weiteres Dokument im Rahmen einer Dokumentensammlung (vgl. die Beschreibung des Vektorraummodells in Kap. 3.1). Über eine Gewichtung der Indexterme wird eine Ähnlichkeit zwischen den Dokumenten ermittelt. Wird eine Frageformulierung als ein weiteres Dokument der Kollektion betrachtet, kann die Relevanz von Dokumenten in bezug auf eine gestellte Anfrage ermittelt werden.

Die Ähnlichkeiten werden über den Merkmalsbesitz ermittelt. Merkmale von Dokumenten und Fragen sind typischerweise die Terme. Die vom System ermittelten Ähnlichkeiten legt die Reihenfolge der Dokumente in der Ergebnisliste fest. Für die Berechnung der Ähnlichkeit wird häufig das sog. Vektorprodukt angewendet, bei dem sich die Ähnlichkeit aus der Produktsumme der Termgewichte errechnet, die in Anfrage und Dokument gemeinsam vorkommen (Salton/McGill 1987). Je höher der ermittelte Wert ist, um so weiter oben steht das Dokument in der Ergebnisliste. Ein definierter Schwellenwert sollte eine untere Grenze bestimmen.

Ohne eine Indexierung mit Termgewichtung (binäre Indexierung) wird der Grad der Übereinstimmung zwischen Anfrage und Dokument über die bloße Anzahl der gemeinsamen Terme bestimmt. Die jeweiligen Terme werden gleichrangig behandelt, obwohl ihre Bedeutung innerhalb der Dokumente und für die Intention einer Anfrage naturgemäß sehr unterschiedlich sein kann. Ein einfaches Beispiel:

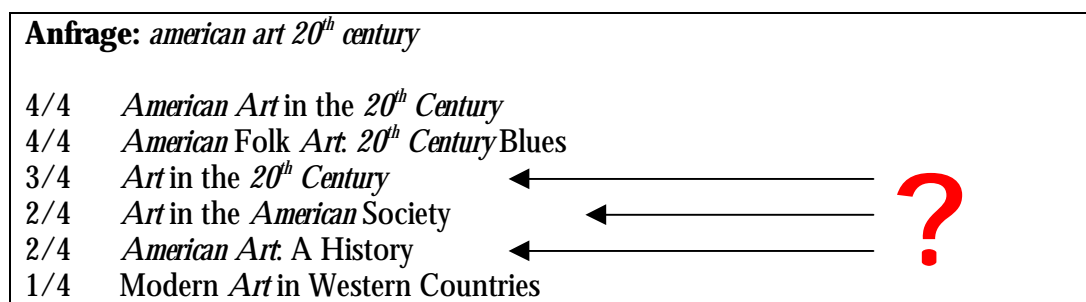


Abb. 12: *Relevance Ranking ohne Gewichtung*

Werden die Terme hingegen gewichtet, können in einer Suchanfrage Grade der Bedeutung für ein aufzufindendes Dokument berücksichtigt werden. Die binäre Indexierung wird durch eine gewichtete ersetzt. Das Ranking der Dokumente in einer Ausgabeliste kann höhere Gewichtungen in der Rangfolge berücksichtigen. Im Beispiel der Abbildung 12 würden etwa *american* und *art* eine höhere Bedeutung für die Anfrage besitzen als *20th* oder *century*. Die Titel *Art in the American Society* und *American Art: A History* würden in der Ausgabeliste vor *Art in the 20th Century* rücken.

Eine einfache Gewichtung der Terme für das Relevance Ranking kann durch die Berücksichtigung der Termfrequenz im Dokument (TF_{td}) und der inversen Dokumenthäufigkeit (IDF_t) erreicht werden. Beide Werte werden durch Multiplikation in Beziehung gesetzt:

$$\text{Gewicht}_{td} = TF_{td} * IDF_t$$

Sollen mehrere Terme in einer Suchfrage betrachtet werden, so werden die einzelnen Termgewichte für jedes Dokument zu einer Summe addiert. Die Summen werden für das Ranking in der Ausgabeliste herangezogen. Auf diese Weise erhalten „gute Indexterme“ ein höheres Gewicht für die Festlegung der Reihenfolge der ausgegebenen Dokumente. Neben den o.g. Werten können weitere Kriterien in die Berechnung der Gewichte einbezogen werden, bspw. die Position des Terms im Dokument oder die Abstände zwischen den betrachteten Termen.

In einem solchen auf Termgewichtungen beruhenden Ergebnisranking werden die angezeigten Dokumente häufig mit einem Mass für die Relevanz-Beurteilung versehen. Welche Kriterien genau in diese Bewertung eingehen, unterliegt bei Suchmaschinen des Internet dem Geschäftsgeheimnis. Für eine eingehendere Untersuchung des Relevance-Ranking wichtiger Suchmaschinen siehe Courtois/Berry (1999). Auf den Hilfeseiten der deutschen Version von Excite wird der Service folgendermassen beschrieben:

Excite listet jeweils 10 Suchergebnisse in absteigender Relevanz auf einmal auf. Das jeweils auf der linken Seite erscheinende Prozentzeichen gibt die Relevanz-Beurteilung an. Je näher dieser Wert bei 100% liegt, desto mehr entspricht (nach Auffassung der Suchmaschine) das Dokument der von Ihnen aufgegebenen Suche. Diese Werte werden automatisch von unserer Suchmaschine generiert, indem sie die Informationen auf der Site mit den Informationen in Ihrer Suchanfrage vergleicht.

Der „Auffassung der Suchmaschine“ liegen natürlich Kriterien zugrunde, die durch den Betreiber der Suchmaschine über einen Algorithmus implementiert wurden.

Für die Suchmaschine Lycos wird ein ähnliches Verfahren folgendermassen beschrieben:

Häufigkeit der Wörter: Dieses Relevanzkriterium vergleicht die Häufigkeit mit der ein Suchwort in einer einzelnen Ergebnisseite auftaucht, mit der durchschnittlichen Häufigkeit dieses Wortes in dem Lycos Katalog. Wenn zum Beispiel das Wort „Computer“ durchschnittlich 10 mal in einem Dokument des Suchkatalogs auftaucht, dann werden Dokumente, die „Computer“ mehr als 10 mal enthalten als wichtiger eingestuft als Dokumente, die dieses Suchwort weniger als 10 mal enthalten.

Problematisch an dieser Form der Relevanz ist die fachliche multidimensionalität der Lycos-Datenbank. Es lässt sich bspw. vermuten, dass Dokumente aus dem Informatik-Sektor das Wort „Computer“ relativ häufig enthalten, ohne das dieses Wort für die einzelnen Dokumente eine hohe inhaltliche Signifikanz besitzt. Diese Häufigkeit beeinflusst nun jedoch auch die Relevanzbewertung für Dokumente aus anderen Disziplinen.

Fortschrittliche (nicht-Boole'sche) Retrievalmodelle lassen sich hinsichtlich ihres theoretischen Hintergrundes in probabilistische (statistische Wahrscheinlichkeitstheorie), vektorielle (Vektorraummodell) und Fuzzy-Retrievalmodelle (Theorie unscharfer Mengen) unterscheiden, die die Ähnlichkeitsfunktion jeweils verschieden interpretieren. Wie jedoch die Untersuchungen der TREC-Conferences²² zeigten, wirken sich diese theoretischen Unterschiede kaum auf die Resultate im Retrievalprozess aus. Die einzelnen Modelle sollen daher an dieser Stelle auch nicht besprochen werden (vgl. dazu Salton/McGill 1987, Fuhr 1997).

Auch Verfahren, die eine schrittweise Verbesserung der Suchergebnisse zum Ziel haben und auf ein bereits erzieltes erstes Resultat aufsetzen bedienen sich der Gewichtungungsverfahren. Über Gewichtungen kann eine Reformulierung von Anfragen aufgrund bereits erzielter Suchergebnisse durchgeführt werden. In manchen Internet-Suchmaschinen finden wir diese Funktion bspw. unter der Bezeichnung „more like this“, „Things like this“ oder „Ähnliche Sites“. Diese Retrievalverfahren werden unter dem Namen *Relevance Feedback* (Abb. 13) geführt (Salton/McGill 1987, S. 150-155). Dabei wird ein auf die Intention der Anfrage „passendes“ Dokument vom Suchenden ausgewählt und als „Vorlage“ für einen weiteren Suchlauf genommen, der eine Verfeinerung zum Ziel hat.

Relevance feedback is a process where users identify relevant documents in an initial list of retrieved documents, and the system then creates a new query based on those sample relevant documents. (Croft 1995)

Schauen wir uns wieder die Beschreibung aus der Hilfe zu Excite an:

²² TREC steht für „Text REtrieval Conference“. Informationen zu den bisherigen Conferences, inkl. Der Proceedings online, finden Sie unter <http://trec.nist.gov>.

Befindet sich unter den aufgelisteten Suchergebnissen ein Dokument, das genau dem von Ihnen gesuchten Thema entspricht? In einem derartigen Fall gehen Sie zur Liste der Suchergebnisse und zur betreffenden dort aufgeführten Site zurück. Klicken Sie hier auf den Link Ähnliche Sites, den Sie direkt neben dem Titel finden. Automatisch wird dieses Dokument jetzt als beispielhafter Ausgangspunkt für einen neuen Suchdurchgang genommen, bei dem weitere Sites gefunden werden, die dieser, von Ihnen als zutreffend bewerteten Seite inhaltlich gleichen.

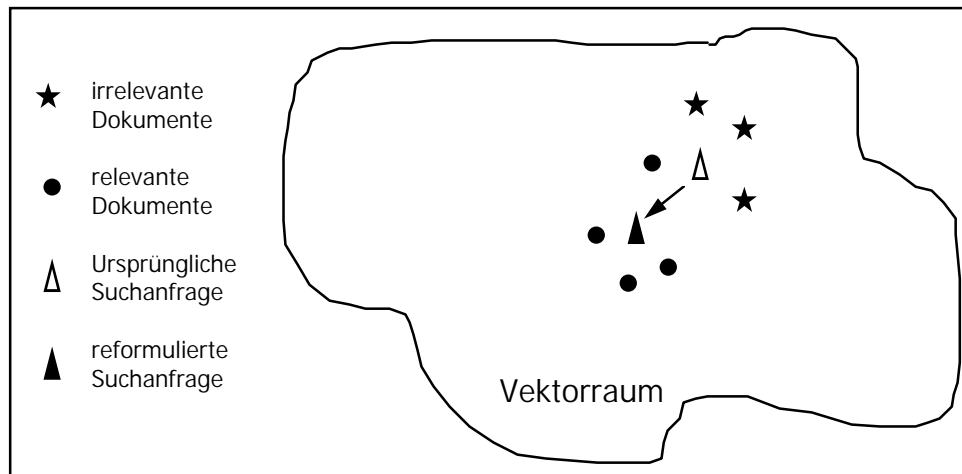


Abb. 13: Modell des Relevance Feedback

Die Terme des ausgewählten Dokuments werden in der Regel den ursprünglichen Suchbegriffen hinzugefügt. Terme aus der ersten Suchformulierung, die in diesem „optimalen“ Dokument nicht enthalten sind, werden für den erneuten Suchlauf nicht weiter berücksichtigt. Im Vektorraum verändert die erneute Suchanfrage damit die Position im Vergleich mit der ursprünglichen Anfrage (Abb. 13) und nähert sich auf diese Weise einem als optimal empfundenen Retrieval-ergebnis an. Dieser Vorgang kann sich mehrfach wiederholen. Die iterative Eigenschaft des Information Retrieval kommt hier im besonderem Masse zum Ausdruck.

Den Ablauf einer Recherche unter Anwendung informationslinguistischer Verfahren (crossreference, weak stemming, strong stemming) im IR-System Okapi (Schröder 1990) zeigt die folgende Abbildung 14:

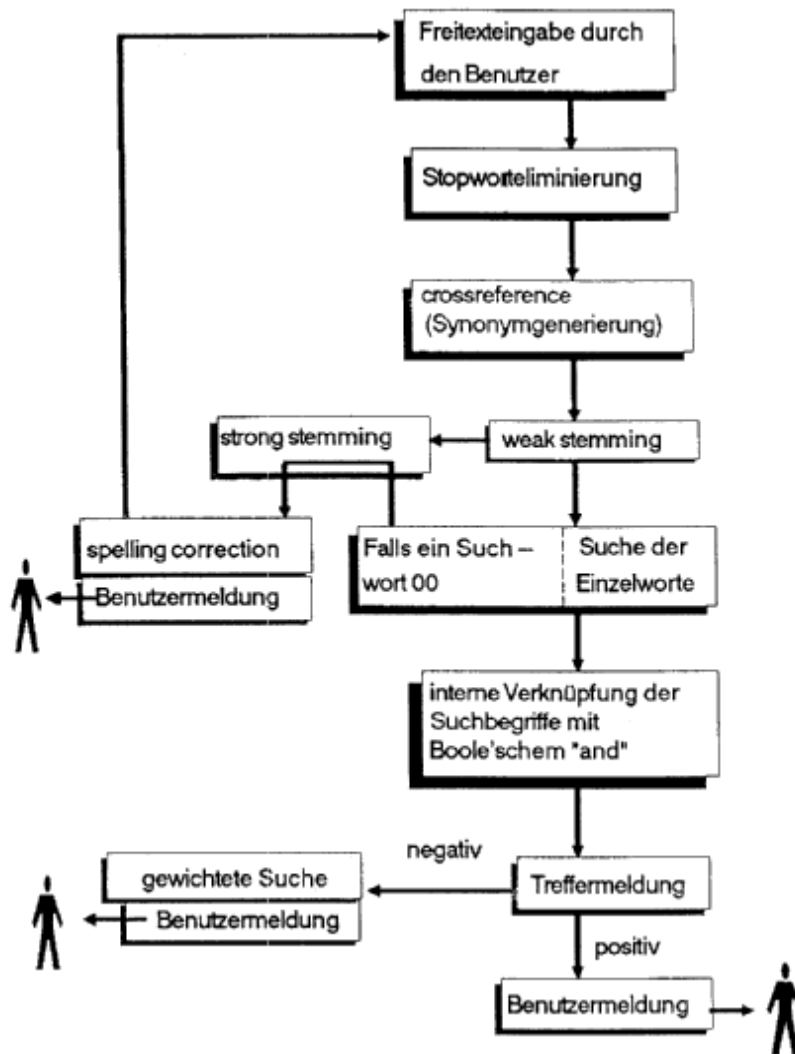


Abb. 14: Retrievalablauf in Okapi

In der Funktion crossreference sind eine Reihe von informationslinguistischen Verfahren zusammengefasst. Dabei werden u.a. unregelmässige Pluralformen, adjektivierte Substantive, Abkürzungen oder englische und amerikanische Schreibweisen sowie echte Synonyme zusammengeführt.

Auf Porters Algorithmus (Porter 1980) beruht die Wortstammreduktion. Zunächst wird eine schwache Reduktion (weak stemming) vorgenommen (-ing, -ed usw.) und – bei unzureichender Treffermenge – eine starke Reduktion (strong stemming) angeschlossen (-ization, -ability usw.).

Bei einem Scheitern der exact-match-Suche wird ein Ranking-Verfahren auf Grundlage des Algorithmus von Harper zur Ermittlung und Ausgabe ähnlicher Dokumente angewendet. Die Berechnung beruht auf der Häufigkeit der einzelnen Suchbegriffe in der Dokumentensammlung. Dabei werden Suchbegriffe höher gewichtet, je seltener sie in der Kollektion enthalten sind.

6. Resümee

Für die Informationswirtschaft ist die effiziente und effektive Verarbeitung unstrukturierter Informationen eine zentrale Aufgabe. Interne wie externe Informationssysteme werden betrieben um dieser Aufgabe gerecht zu werden.

Entscheidend für den Erfolg aller dieser Informationssysteme ist die automatische Erschliessung der vorliegenden Information, die automatische Indexierung. Eine Reihe im Grundsatz verschiedener Ansätze der automatischen Indexierung sind durch die IR-Forschung entwickelt worden, praktische Relevanz haben heute vor allem statistische sowie informationslinguistische Verfahren – meist in Kombination – erlangt.

Anhang

Nachfolgend sind zwei Meldungen der Agentur Reuters vom 20. März 2000 wiedergegeben. Diese Meldungen dienen als Beispieltex te.

Beispiel I:

BMW plant eigenen Internet-Marktplatz

Hamburg (Reuters) – Der BMW-Konzern plant einem Zeitungsbericht zufolge eine eigene Plattform für einen elektronischen Marktplatz. Die "Financial Times Deutschland" berichtete am Mittwoch, BMW wolle bei dem elektronischen Marktplatz aber nicht mit der Plattform der drei Hersteller DaimlerChrysler, Ford Motors und General Motors konkurrieren. Das BMW-Modell soll sich dem Bericht zufolge auf sogenanntes indirektes Material beziehen, dass BMW nicht unmittelbar für die Produktion seiner Fahrzeuge benötigt. Das Konzept der drei Konkurrenz-Konzerne zielt hingegen auf einen elektronischen Marktplatz für die Zulieferindustrie.

Beispiel II:

Freenet steigt bei Musik-Portal ein

Hamburg (Reuters) – Der Hamburger Online-Dienst Freenet hat sich mit 50 Prozent an der music@lines AG beteiligt. Ziel der Zusammenarbeit sei es, das Musik-Portal der neuen Beteiligung zu einem führenden Angebot im Internet auszubauen, teilte die am Frankfurter Neuen Markt gelistete Mobilcom-Tochter am Mittwoch in einer Pflichtveröffentlichung mit. Von April an werde das gemeinsam betriebene Portal zehn Online-Channels mit abspielbaren Titeln, Internet-Radio und redaktionellen Beiträgen zum Thema Musik sowie der Möglichkeit zum Ticket-Kauf anbieten.

Literatur

- Ahlfeld, C.: Ein Reduktionsalgorithmus für deutsche Wortformen als Softwarewerkzeug zur Optimierung des Recalls bei Literaturrecherchen durch OPAC-Benutzer – ein Schritt auf dem Weg zur dritten OPAC-Generation. Fachhochschule Hamburg, Fachbereich Bibliothek und Information, 1995 (Diplomarbeit)
- Alliger, W.; Richter, W.: Ein Verfahren zur automatischen Indexierung deutschsprachiger Texte im Rahmen des Internationalen Zweiginformationssystems Elektrotechnik. In: Informatik 25 (1978) 4, S. 10-15
- Automatische Indexierung IDX mit Übersetzungsfunktion. Hrsg.: SOFTEX GmbH. Saarbrücken 1997
- Biebricher, P.; Fuhr, N.; Lustig, G.; Schwantner, M.; Knorz, G.: Das automatische Indexierungssystem AIR/PHYS. In: Deutscher Dokumentartag 1987: Von der Information zum Wissen – vom Wissen zur Information: Traditionelle und moderne Informationssysteme für Wissenschaft und Praxis. Weinheim: VCH, 1988. S. 319-328
- Brilmayer, I.; Schellkes, W.; Schlotte, M.; Seitner, P.: Experiment mit automatischer Indexierung. FH Darmstadt, Fachbereich IuD, 1997
(<http://www.iud.fh-darmstadt.de/iud/wwwmeth/lv/ss97/wpai/grpair/ausarb1.htm>)
- Browne, G.: Automatic Indexing and Abstracting. In: AusSI Newsletter 20 (1996) July 1996, S. 4-9
- Croft, W.B.: What Do People Want from Information Retrieval? The Top Ten Research Issues for Companies that Use and Sell IR Systems. In: D-Lib Magazine, November 1995
(<http://www.dlib.org/dlib/november95/11croft.html>)
- Crystal, D.: Die Cambridge Enzyklopädie der Sprache. Frankfurt a.M.: Campus, 1995
- Courtois, M.P.; Berry, M.W.: Results Ranking in Web Search Engines. In: Online, May 1999.
(<http://www.onlineinc.com/onlinemag/OL1999/courtois5.html>)
- Endres-Niggemeyer, B.: Summarizing Text for Intelligent Communication. In: Knowledge Organization 21 (1994) 4, S. 213-223
- Endres-Niggemeyer, B.: Summarizing Information. Berlin: Springer, 1998
- Foltz, P.W.; Dumais, S.T.: Personalized Information Delivery: An Analysis of Information Filtering Methods. In: Communications of the ACM 35 (1992) 12, S. 51-60
- Forst, A.: Dokumente speichern, indizieren und wiederfinden. In: Wissensmanagement 2 (1999) 2, S. 23-28
- Fugmann, R.: Theoretische Grundlagen der Indexierungspraxis. Frankfurt a.M.: Indeks Verlag, 1992
- Fuhr, N.: Information Retrieval: Skriptum zur Vorlesung. Universität Dortmund, 1997
- Gaus, W.: Dokumentations- und Ordnungslehre: Theorie und Praxis des Information Retrieval. 2. Auflage. Berlin: Springer, 1995

- Gebhardt, F.: Dokumentationssysteme. Berlin: Springer, 1981
- Geis, A.: Computergestützte Inhaltsanalyse – Hilfe oder Hinterhalt. In: C. Züll und P.Ph. Mohler (Hrsg.): Textanalyse: Anwendungen der computerunterstützten Inhaltsanalyse. Opladen: Westdeutscher Verlag, 1992. S. 7-32
- Genzmer, H.: Deutsche Grammatik. Frankfurt a.M.: Insel Verlag, 1995
- Gerick, Th.: Recherchetechniken: Suchen und Finden sind zweierlei. Dokumenten-Management / Intelligentes Information Retrieval als KM-Basistechnik. In: Computerwoche Nr. 7/2000, S. 90-92
- Gödert, W.; Liebig, M.: Maschinelle Indexierung auf dem Prüfstand: Ergebnisse eines Retrievaltests zum MILOS II Projekt. In: Bibliotheksdienst 31 (1997) 1, S. 59-68
- Görz, G.: Wissensrepräsentation und die Verarbeitung natürlicher Sprache. In: Wissensrepräsentation / Hrsg.: P. Struss. München: Oldenbourg, 1991. S. 87-101
- Goeser, S.: Linguistik und Wissensrepräsentation im Information Retrieval. In: it + ti – Informationstechnik und Technische Informatik 36 (1994) 2, S. 19-26
- Goeser, S. (1997): Inhaltsbasiertes Information Retrieval: Die TextMining-Technologie. In: LDV-Forum (1997) 1
- Gotthard, W.; Marwick, A.; Seiffert, R.: Mining Text Data. In: DB2 Magazine online, Winter 1997
(<http://www.db2mag.com/97wiGot.htm>)
- Gräbnitz, V.: PASSAT: Programm zur Automatischen Selektion von Stichwörtern aus Texten. In: Inhaltserschliessung von Massendaten: Zur Wirksamkeit informationslinguistischer Verfahren am Beispiel des Deutschen Patentinformationssystems / Hrsg.: J. Krause. Hildesheim: Olms, 1987. S. 36-55
- Haapaleinen, M.; Majorin, A.: GERTWOL und Morphologische Disambiguierung für das Deutsche. In: Proc. of the 10th Nordic Conference on Computational Linguistics; May 30-31 1995, Helsinki (NODALIDA-95)
(<http://www.lingsoft.fi/doc/gercg/NODALIDA-poster.html>)
- Haller, J.; Wieland, U.: Die Erschliessung natürlichsprachiger Information im Informationssystem CONDOR. In: Nachrichten für Dokumentation 29 (1978) 4/5, S. 177-183
- Hauer, M.: Automatische Indexierung. In: Schmidt, R. (Hrsg.): Wissen in Aktion: Wege des Knowledge Managements; 22. Online-Tagung der DGI, Frankfurt, 2. bis 4. Mai 2000. Frankfurt: DGD. S. 203-212
- Henzler, R.G.: Free or Controlled Vocabularies. In: International Classification 5 (1978), S. 21-26
- Hovy, E./Lin, C.-Y.: Automated Text Summarization in SUMMARIST. In: Mani, I.; Maybury, M.T. (Ed.): Advances in Automatic Text Summarization. Cambridge, MA: MIT Press, 1999. S. 81-94
- Indexing and Morphology: Why You Need Morphology. Lingsoft, o.J.
(<http://www.lingsoft.fi/doc/indexing/morph.html>)

- Jacobsen, J.: Auf den Punkt gebracht: Können Computer Texte verstehen und zusammenfassen. In: Die Zeit Nr. 49/1998
- Kaiser, A.: Computer-unterstütztes Indexieren in Intelligenten Information Retrieval Systemen: Ein Relevanz-Feedback orientierter Ansatz zur Informationserschliessung in unformatierten Datenbanken. Wien: Universität Wien, 1993
- Kampffmeyer, U.: Der Markt für elektronisches Dokumenten-Management in Europa: Technologien und Lösungen. In: PROJECT CONSULT Newsletter v. 26.11.1999. S. 9-18
- Keitz, W. von: Automatic Indexing and the Dissemination of Information. In: INSPEL 20 (1986), S. 47-67
- Klosterberg, M.: Das computergestützte Gruppengedächtnis: Wissensmanagement in Sitzungen. Wiesbaden: Deutscher Universitätsverlag, 1999
- Knorz, G.: Automatische Indexierung. In: Wissensrepräsentation und Information Retrieval / Hrsg.: R.-D. Hennings et al. Potsdam, 1994. S. 138-196
- Knorz, G.: Information Retrieval-Anwendungen. In: Kleines Lexikon der Informatik und Wirtschaftsinformatik / Hrsg.: M.G. Zilahi-Szabo. München: Oldenbourg, 1995. S. 244-248
- Knorz, G.; Arz, J.; Rostek, L.; Steffen, J.: Adaptive automatische Indexierung für komplexe Dokumente. In: LDV-Forum (1997) 1
- Königer, P.; Reithmayer, W.: Management unstrukturierter Informationen: Wie Unternehmen die Informationsflut beherrschen können. Frankfurt/Main: Campus, 1998
- Krause, J.: Principles of Content Analysis for Information Retrieval Systems: An Overview. In: ZUMA-Nachrichten Spezial: Text Analysis and Computers / Hrsg.: C. Zuell, J. Harkness u. J.H.P. Hoffmeyer-Zlotnik. Mannheim: ZUMA, 1996. S. 76-100
- Krause, J.; Womser-Hacker, Ch.: PADOK-II: Retrievaltests zur Bewertung von Volltext-indexierungsvarianten für das Deutsche Patentinformationssystem. In: Nachrichten für Dokumentation 41 (1990) 1, S. 13-19
- Krause, J.; Mutschke, P.: Indexierung und Fulcrum-Evaluierung. Bonn: InformationsZentrum Sozialwissenschaften, 1999
- Krüger, C.: Evaluation des WWW-Suchdienstes GERHARD unter besonderer Beachtung der automatischen Indexierung. Fachhochschule Stuttgart – HBI 1999 (Diplomarbeit)
- Kuhlen, R.: Morphologische Relationen durch Reduktionsalgorithmen. In: Nachrichten für Dokumentation 25 (1974) 4, S. 168-172
- Kuhlen, R.: Information Retrieval: Verfahren des Abstracting. In: Computational Linguistics – Computerlinguistik: An International Handbook of Computer Oriented Language Research and Applications / Hrsg.: S. Batori, W. Lenders und W. Putschke. Berlin: de Gruyter, 1989. S. 688-696
- Lehmann, E.: Problemaspekte der Wissensrepräsentation. In: Siemens Forschungs- und Entwicklungsberichte 17 (1988) 2, S. 45-51

- Lehner, F.: Organisational Memory: Konzepte und Systeme für das organisationale Lernen und das Wissensmanagement. München: Hanser, 2000
- Lepsky, K.: Maschinelle Indexierung von Titelaufnahmen zur Verbesserung der sachlichen Erschliessung in Online-Publikumskatalogen. Köln: Greven, 1994
- Lepsky, K.: Automatisierung in der Sacherschliessung: Maschinelles Indexieren von Titeldaten. In: 85. Deutscher Bibliothekartag in Göttingen 1995: Die Herausforderung der Bibliotheken durch elektronische Medien und neue Organisationsformen. Frankfurt a.M.: Klostermann, 1996. S. 223-233 [1996a]
- Lepsky, K.: Automatische Indexierung und bibliothekarische Inhaltsererschliessung: Ergebnisse des DFG-Projekts MILOS I. In: Zukunft der Sacherschliessung im OPAC: Vorträge des 2. Düsseldorfer OPAC-Kolloquiums am 21. Juni 1995 / Hrsg.: E. Niggemann u. K. Lepsky. Düsseldorf: Universitäts- und Landesbibliothek, 1996. S. 12-36 [1996b]
- Lepsky, K.: Im Heuhaufen suchen – und finden: Automatische Erschliessung von Internetquellen: Möglichkeiten und Grenzen. In: Buch und Bibliothek 50 (1998) 5, S. 336-340
- Lepsky, K.; Siepmann, J.; Zimmermann, A.: Automatische Indexierung für Online-Kataloge: Ergebnisse eines Retrievaltests. In: Zeitschrift für Bibliothekswesen und Bibliographie 43 (1996) 1, S. 47-56
- Lepsky, K.; Zimmermann, H.H.: Katalogerweiterung durch Scanning und Automatische Dokumenterschliessung: Das DFG-Projekt KASCADE. In: ABI-Technik 18 (1998) 1, S. 56-60
- Luckhardt, H.-D.: Informationslinguistik. In: Einführung in die Informationswissenschaft / I. Harms u. H.-D. Luckhardt. Universität des Saarlandes, Fachrichtung Informationswissenschaft, 1998
(<http://www.phil.uni-sb.de/FR/Infowiss/papers/iwscript/infoling/index.html>)
- Luhn, H.P.: The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development 2 (1958) 2, S. 159-165 (Wiederabdruck in: Mani, I.; Maybury, M.T. (Ed.): Advances in Automatic Text Summarization. Cambridge, MA: MIT Press, 1999. S. 15-21)
- Lustig, G.: Die automatische Zuteilung von Schlagwörtern des EURATOM-Thesaurus. In: Neue Technik 14 (1969) A4, S. 247-256
- Lustig, G.: Automatische Indexierung zwischen Forschung und Anwendung. Hildesheim: Olms, 1986
- Lustig, G.: Automatische Indexierung und Information Retrieval – Erfahrungen und Perspektiven. In: Klassifikation und Ordnung / Hrsg.: R. Wille. Frankfurt a.M.: Indeks Verlag, 1989. S. 137-148
- Maller, S.: What is InfoSort? In: Dialect 3/1998
<http://library.dialog.com/newsletters/dialect/issue3/infosort.html>
- Mani, I.; Maybury, M.T. (Ed.): Advances in Automatic Text Summarization. Cambridge, MA: MIT Press, 1999
- Martin, G.: Dokumenten-Management: Wissensbasierte Analyse und Recherche. Mit KI-Systemen sollen Anwender Papierflut eindämmen. In: Computerwoche 17/1998, S. 24

- Mater, E.: Ziele und Methoden automatischer Inhaltserschliessung. In: Dokumentation/Information (1990), Heft 77, S. 36-50
- Mergenthaler, E.: Computer-Assisted Content Analysis. In: ZUMA-Nachrichten Spezial: Text Analysis and Computers / Hrsg.: C. Zuell, J. Harkness u. J.H.P. Hoffmeyer-Zlotnik. Mannheim: ZUMA, 1996. S. 3-32
- Nohr, H.: Maschinelle Experten für Information Retrieval. In: Informatik 38 (1991) 6, S. 234-236
- Nohr, H.: Inhaltsanalyse. In: nfd. Information – Wissenschaft und Praxis 50 (1999) 2, S. 69-78 [1999a]
- Nohr, H.: Das Projekt OurKnowledge: Wissensmanagement in einem kleinen Beratungsunternehmen. In: HBI aktuell 2/1999, S. 16-19 [1999b]
- Nohr, H.: Automatische Verfahren der Dokumentanalyse. In: Ders.: Wissensmanagement: Wie Unternehmen ihre wichtigste Ressource erschliessen und teilen. Göttingen: BusinessVillage, 2000 (eBook). S. 61-87
- Paice, C.D.: Constructing Literature Abstracts by Computer: Techniques and Prospects. In: Information Processing and Management 26 (1990) 1, S. 171-186
- Panyr, J.: Automatische Indexierung und Klassifikation. In: Automatisierung in der Klassifikation / Hrsg.: I. Dahlberg u. M. Schader. Frankfurt a.M.: Indeks Verlag, 1983. S. 90-111
- Porter, M.F.: An Algorithm for Suffix Stripping. In: Program 14 (1980) 3, S. 130-137
- Rau, L.F.; Jacobs, P.S.; Zernik, U.: Information Extraction and Text Summarization Using Linguistic Knowledge Acquisition. In: Information Processing & Management 25 (1989) 4, S. 419-428
- Recker, I.; Ronthaler, M.; Zillmann, H.: OSIRIS: Osnabrück Intelligent Research Information System – ein Hyperbase Front End System für OPACs. In: Bibliotheksdienst 30 (1996) 5, S. 833-848
- Reimer, U.: Verfahren der automatischen Indexierung. Benötigtes Vorwissen und Ansätze zu seiner automatischen Akquisition: Ein Überblick. In: Experimentelles und praktisches Information Retrieval: Festschrift für Gerhard Lustig / Hrsg.: R. Kuhlen. Konstanz: Universitätsverlag, 1992. S. 171-194
- Rijsbergen, C.J. van: Information Retrieval. 2. Auflage – London: Butterworths, 1979
- Riggert, W.: Betriebliche Informationskonzepte. Braunschweig: Vieweg, 1998
- Ronthaler, M.; Sauer, U.: OSIRIS – Computerlinguistik in der wissenschaftlichen Bibliothek. Osnabrück, 1997
- Ruge, G.; Goeser, S.: Information Retrieval ohne Linguistik? In: nfd. Information - Wissenschaft und Praxis 49 (1998) 6, S. 361-369
- Sachse, E.; Liebig, M.; Gödert, W.: Automatische Indexierung unter Einbeziehung semantischer Relationen: Ergebnisse des Retrievaltests zum MILOS-II-Projekt. Fachhochschule Köln, Fachbereich Bibliotheks- und Informationswesen 1998

- Salton, G.: Automatic Text Analysis. In: Science, Vol. 168 (1970) S. 335-343
- Salton, G.: Fast Document Classification in Automatic Information Retrieval. In: Kooperation in der Klassifikation I. Frankfurt a.M.: Indeks Verlag, 1978. S. 129-146
- Salton, G.: Another Look at Automatic Text Retrieval Systems. Ithaca: Department of Computer Science, Cornell University, 1985
- Salton, G.; McGill, M.J.: Information Retrieval – Grundlegendes für Informationswissenschaftler. Hamburg: McGraw-Hill, 1987
- Schmitz, U.: Computerlinguistik: Eine Einführung. Opladen: Westdeutscher Verlag, 1992
- Schneider, Ch.: Automatische Indexierung und Syntaxanalyse: Zur Entwicklung sprachanalytischer Komponenten von Informationssystemen auf empirischer Grundlage. Hamburg: Buske, 1985
- Schröder, K.: Zur OPAC-Diskussion, ein britisches Projekt: OKAPI. In: Bibliotheksdienst 24 (1990) 11, S. 1504-1512
- Schwantner, M.: Entwicklung und Pflege des Indexierungswörterbuches PHYS/PILOT. In: Deutscher Dokumentartag 1987: Von der Information zum Wissen – vom Wissen zur Information: Traditionelle und moderne Informationssysteme für Wissenschaft und Praxis. Weinheim: VCH, 1988. S. 329-339
- Schwarz, Ch.: Linguistische Hilfsmittel beim Information Retrieval. In: Nachrichten für Dokumentation 35 (1984) 4/5, S. 179-182
- Sparck Jones, K.: A Statistical Interpretation of Term Specificity and its Application in Retrieval. In: Journal of Documentation 28 (1972), S. 11-21
- Sparck Jones, K.: Automatic Summarizing: Factors and Directions. In: Mani, I.; Maybury, M.T. (Ed.): Advances in Automatic Text Summarization. Cambridge, MA: MIT Press, 1999. S. 1-12
- Stock, W.G.: Natürlichsprachige Suche – More like this! Lexis-Nexis' Freestyle. In: Password (1998) 11, S. 21-28
- Stock, W.G.: Informationswirtschaft: Management externen Wissens. München: Oldenbourg, 2000
- Stock, W.G.: Qualitätskriterien von Suchmaschinen. In: Password (1999) 5, S. 22-31
- Teuber, T.: Information Retrieval und Dokumentenmanagement in Büroinformationssystemen. Göttingen: Unitext Verlag, 1996
- Thiel, Th.J.: Automated Indexing of Information Stored on Optical Disk Electronic Document Image Management Systems. In: Encyclopedia of Library and Information Science. New York: Dekker. 54, Suppl. 17 (1994), S. 98-121
- Tkach, D.: Text Mining Technology: Turning Information Into Knowledge; A White Paper from IBM. IBM Software Solutions, 1997
- Turney, P.: Extraction of Keyphrases from Text: Evaluation of Four Algorithms. National Research Council of Canada, 1997 (ERB-1051)

- Turney, P.: Learning to Extract Keyphrases from Text. National Research Council of Canada, 1999 (ERB-1057)
- Turney, P.: Learning Algorithms for Keyphrase Extraction. In: Information Retrieval (im Druck), 2000
- Volk, M.; Mittermaier, H.; Schurig, A.; Biedassek, T.: Halbautomatische Volltextanalyse, Datenbankaufbau und Document Retrieval. In: Datenanalyse, Klassifikation und Informationsverarbeitung: Methoden und Anwendungen in verschiedenen Fachgebieten / Hrsg.: H. Goebel und M. Schader. Heidelberg: Physica, 1992. S. 205-214
- Wätjen, H.-J.; Diekmann, B.; Möller, G.; Carstensen, K.-U.: GERHARD – German Harvest Automated Retrieval and Directory: Bericht zum DFG-Projekt. Oldenburg: Bibliotheks- und Informationssystem der Universität, 1998
- Werner, H.: Indexierung auf linguistischer Grundlage am Beispiel von JUDO-DS(1). In: Deutscher Dokumentartag 1981: Kleincomputer in Information und Dokumentation. München: Saur, 1982. S. 599-609
- Winiwarter, W.: Bewältigung der Informationsflut: Stand der Computerlinguistik. In: Nachrichten für Dokumentation 47 (1996) 3, S. 131-150
- Zimmermann, H.H.: Automatische Indexierung – Entwicklung und Perspektiven. In: Automatisierung in der Klassifikation / Hrsg.: I. Dahlberg und M. Schader. Frankfurt a.M.: Indeks Verlag, 1983. S. 14-32
- Zimmermann, H.H.: Maschinelle Übersetzung in der Wissensvermittlung. In: etz. Elektrotechnische Zeitschrift 110 (1989) 23/24, S. 1252-1256
- Zimmermann, H.H.: Computer und Sprache im Zeitalter der Fachinformation. In: Lebende Sprachen 35 (1990) 1, S. 1-5
- Zimmermann, H.H.: Language and Language Technologie. In: International Classification 18 (1991) 4, S. 196-199
- Zimmermann, H.H.: Automatische Indexierung und elektronische Thesauri. In: Zukunft der Sacherschliessung im OPAC: Vorträge des 2. Düsseldorfer OPAC-Kolloquiums am 21 Juni 1995 / Hrsg.: E. Niggemann und K. Lepsky. Düsseldorf: Universitäts- und Landesbibliothek. S. 37-47
- Zimmermann, H.H.; Kroupa, E.; Keil, G. : CTX – Ein Verfahren zur Computergestützten Textanalyse. Saarbrücken 1983 (BMFT Forschungsbericht ID 83-006)

Bisher erschienen:Stand:
September 2000

1/2000	Wissen und Wissensprozesse visualisieren	Prof. Holger Nohr
2/2000	Automatische Dokumentindexierung – Eine Basistechnologie für das Wissensmanagement	Prof. Holger Nohr
3/2000	Einführung von Wissensmanagement in einer PR-Agentur	Prof. Holger Nohr