Jiri Panyr, München
Harald H. Zimmermann, Saarbrücken

Information Retrieval : Überblick über aktive Systeme und Entwicklungstendenzen

In: Istvan S. Batori, Winfried Lenders, Wolfgang Putschke (Hrsg, 1989): Computational Linguistics - Computerlinguistik. Berlin / New York: Walter de Gruyter, Thema 56, S. 696-708

1. Definition und Abgrenzung

Es ist bezeichnend, dass es zu dem englischen Terminus *Information Retrieval* (im folgenden kurz IR) keinen schlüssigen deutschen Begriff gibt. Als eine deutsche Variante findet sich gelegentlich die Schreibweise Informations-Retrieval, hin und wieder auch die Umschreibung *Ablage und (Wieder-)Auffinden von Information* i. S. des weitergehenden englischen Terminus *Information storage and retrieval*. Damit wird zugleich ein grundlegender - auch in der vorliegenden Darstellung wesentlicher - Zusammenhang deutlich: es gibt kein IR ohne Verknüpfung mit der (vorherigen) Analyse, Verarbeitung und Speicherung bzw. Ablage von Daten. Eine Darstellung der Systeme und Entwicklung des IR muss daher stets vor dem Hintergrund der Datenerschließung und -ablage gesehen werden.

Mit dem Begriff 'Information' wird zugleich ein inzwischen ziemlich ausgetretenes Wortfeld angesprochen. Auch dieser Begriff bedarf daher hier einer knappen Präzisierung: in der informationswissenschaftlichen Forschung wird unter 'Information' der Transfer von Wissen (d. h. der *Prozess* der Wissensvermittlung) bzw. das Ergebnis des Wissenstransfers (also der veränderte Wissens*zustand*) verstanden. Das (menschliche) Wissen ist bzw. wird u. a. in Form von gesprochener oder geschriebener Sprache festgehalten (gespeichert) und vermittelt. Auf diesen Teilaspekt, genauer: auf *textbezogenes IR* wird sich die vorliegende Erörterung im methodischen wie im anwendungsorientierten Teil beschränken.

Mit dieser Beschränkung ist jedoch ein Problem verbunden: In der praktischen Anwendung gibt es kaum ein Retrieval, das allein auf Sprachdaten aufbaut. Bereits in der Literaturdokumentation (d. h. dem sog. *Referenz*retrieval: Ziel ist das Auffinden von Fachliteratur auf der Basis von Titeln und Abstracts) spielen 'Fakten', d. h. formal festgelegte Kategorien mit numerischen oder alphabetischen Werten (Jahreszahl, Autor ...) eine wichtige Rolle; auch die (Voll-)Textdokumentation ist in der Praxis (man vgl. z. B. die juristischen Informationssysteme) mit kategorial-numerischen Fakten kombinierbar.

Schließlich muss festgehalten werden: je tiefer und präziser das in Texten, d. h. in *natürlicher (Schrift-)Sprache* 'oberflächig' festgehaltene Wissen über geeignete Verfahren *faktenbezogen* erschlossen wird, um so weniger wird das auf die *oberflächenorientierte Sprachdatendarstellung* bezogene Retrieval relevant: linguistische bzw. computerlinguistische Verfahren dienen dabei allenfalls noch der Erschließung des Faktenwissens für hochkomplizierte Expertensysteme; die natürliche (gesprochene oder geschriebene) Sprache wird daneben als ein mögliches Interaktionsinstrument Bedeutung behalten (vgl. Art. 57). Es scheint jedoch, dass es bis zur Verwirklichung dieser Ziele noch ein weiter Weg ist. Nach wie vor sind Frage-Antwort-Systeme und Expertensysteme, die natürlichsprachige Daten verarbeiten, nur für einen eng umgrenzten Bereich und zudem meist nur experimentell einsetzbar. Was dem Menschen scheinbar so leicht fällt, nämlich der Umgang mit natürlicher Sprache, ist für den Computer eine Sisyphusarbeit: Kaum erscheint ein Problem in der Sprachdatenverarbeitung gelöst, so zeigen sich neue Gegenbeispiele oder auch weitere unüberwindlich erscheinende Hürden. Da in der Praxis zudem nicht einfach mit wenigen Satzbeispielen 'gerechnet' werden kann, sondern eine Fülle von Daten kodiert und manipuliert werden muss, klafft inzwischen eine große Lücke zwischen den modellorientierten, in der Theorie, vielleicht auch noch im Labor experimentell vorstellbaren Konzepten einer tief gehenden, wissens- und faktenorientierten Sprachdatenerschließung und der Wirklichkeit, d. h. der oberflächenorientierten Sprachdatenverarbeitung in der praktischen Information und Dokumentation bzw. der Bürokommunikation.

Die Bedeutung natürlichsprachiger Texte ist jedoch nicht vollständig analysierbar und beschreibbar, so dass die Texte bei ihrer Repräsentation (z. B. auch für ein Frage-Antwort-System) immer einer Vergröberung unterliegen. Mit anderen Worten: Dokumente (d. h. natürlichsprachige Texte) können nie ohne Informationsverlust in eine formale Sprache überführt werden (vgl. z. B. Cherniavsky 1978). Aufgrund dieser Tatsache unterscheiden sich DokumentRetrieval-Systeme essentiell von Fakten-Retrieval-Systemen (vgl. Cherniavsky/Schneider 1978): Bei letzteren liegt das Datenmaterial bereits in einer formalisierten Sprache vor, so dass sowohl die Speicherung der Daten als auch die Festlegung der Suchaspekte immer eindeutig erfolgen kann. (Dies schließt nicht aus, dass *Teilaspekte* eines Fakten-Retrieval in ein IR-System integriert werden können. (s. u.)). Hier ist jedoch der Gegensatz zu Dokument-Retrieval-Systemen festzuhalten: Bei Dokument-Retrieval-Systemen sind weder Informationen aus den Dokumenten noch das Informationsbedürfnis der Benutzer exakt als formaler Ausdruck in einer Repräsentationssprache dargestellt bzw. darstellbar (vgl. Jochum 1982, 18 ff.).

In der Praxis der Textinformation und Kommunikation stellt sich u. a. das Problem der Bewältigung von Massendaten. *Technisch* betrachtet bildet die elektronische Erstellung, Speicherung und Vermittlung von Textdaten keine entscheidende Barriere mehr: Mit den (elektronischen) Festplattenspeichern, zunehmend auch (Image-orientierten) Bildplatte bzw. der (optischen) Compact Disk (CD-ROM) werden 'vor Ort', d. h. am Arbeitsplatz, ausreichende Speichermöglichkeiten verfügbar; mit der sog. 'Paketvermittlung' und den sich ausweitenden digitalen Vermittlungssystemen (vgl. z. B. das ISDN-Konzept der Deutschen Bundespost) wird daneben eine Datenfernübertragung (relativ) kostengünstig; die Satellitenkommunikation lässt weltweite Datenverbünde u. a. in der Fachkommunikation entstehen (ein Beispiel dafür ist das Scientific and Technical Network - STN - mit den z. Zt. 3 vernetzten Zentren in USA, Japan und der BRD). Die Technik des optischen Lesens unter Identifikation von Schriftzeichen ist daneben so weit fortgeschritten, dass Sprachtexte praktisch in beliebiger Menge zur Weiterverarbeitung verfügbar werden; zunehmend werden textuelle Daten zudem bereits über Texteditoren (auf PC) maschinenlesbar (und weiterverarbeitbar) erfasst. Die noch bestehenden Probleme u. a. in Hinsicht auf den Austausch

von Textdaten über Netze unter Bewahrung der 'logischen' Datenstruktur (*Office Document Architecture* -ODA-; *Standard Generalized Markup Language* -SGML-) werden unter dem Einfluss des notwendigen weltweiten oder auch regional-branchenspezifischen Datenaustausches in Kürze gelöst sein. Dies alles bekräftigt den Eindruck, dass es nach wie vor für das IR von Text- und Sprachdaten eine praxisrelevante Aufgabe gibt: die in Massen anfallenden bzw. verfügbaren Daten in endlicher (d. h. ökonomisch vertretbarer) Zeit strukturell-inhaltlich so aufzubereiten und verfügbar zu machen, dass das darin gespeicherte Wissen bei informationellen Problemlösungen angemessen eingebracht werden kann.

Der Bezug auf 'aktive' Systeme im Titel des Beitrags deutet an, dass es eher um die Praxis als um theoretische Möglichkeiten gehen soll. Eine reine Deskription ist jedoch aus verschiedenen Gründen fehl am Platz: Einerseits soll ein Handbuch wesentliche Aspekte eher systematisch aufzeigen und an typischen Beispielen belegen. Andererseits versteht sich dieser Beitrag auch als ein Vermittlungsversuch zwischen der Praxis einerseits und der (informationswissenschaftlichen) Forschung und Entwicklung andererseits: es müssen daher auch konkrete Möglichkeiten aufgezeigt werden, die den Praktikern - aus welchen Gründen auch immer - nicht ausreichend bewusst sind (oder als nicht machbar erscheinen).

2. Funktionen des linguistischen Information Retrieval, Anwendungsrahmen

Es wurde schon darauf hingewiesen, dass sich die Überlegungen auf die Speicherung und das Wiederfinden geschriebener Sprache (d. h. auf textuelle Information und Kommunikation) beschränken. Der in der Forschung zunehmend interessanter werdende Bereich der Analyse und Ablage gesprochener Sprache wird also ausgeklammert: Hierzu wird vielleicht in den 90er Jahren etwas Sinnvolles gesagt werden können.

Die Anwendung *linguistischer* Verfahren muss jedoch allgemein im Zusammenhang mit *statistischen* Methoden gesehen werden. Geht man von den zwei wichtigsten Bewertungskriterien für ein IR-System (bzw. für die Recherche in einem solchen System) aus, d. h. vom *Recall*-Koeffizienten (dem Maß für die Vollständigkeit des Suchergebnisses) und von der *Precision* (dem Maß für die Genauigkeit des Suchergebnisses), sollen die statistischen Methoden für die Indexierung und Recherche vor allem den *Recall*-Parameter verbessern, während von den linguistischen (syntaktischen) Methoden die Verbesserung der *Precision* erwartet wird (vgl. auch Salton 1986). Unterscheidet man z. B. bei der Recherche zwischen einer Grob- und einer Feinrecherche, spielen die statistischen Methoden vor allem bei der Grobrecherche eine wichtige Rolle, während u. a. die komplexeren linguistischen Verfahren in der sog. Feinrecherche verwendet werden, d. h. nach einer ersten - groben - Vorauswahl zum Einsatz kommen.

Aus der Sicht der praktischen Anwendung lässt sich der Einsatzbereich computerlinguistischer Verfahren im IR etwa wie folgt abgrenzen: An dem einen 'Ende' des Spektrums steht das sog. *Referenz-Retrieval*, d. h. die Integration von Funktionen der Sprachdatenverarbeitung bei der Suche nach Texten oder genauer: 'Dokumenten'. Bei der Suche werden Wörter (Stichwörter) benutzt, die im *Titel* oder auch *Abstract* eines Dokuments vorliegen, bzw. *Schlagwörter* verwendet, die zuvor durch einen (menschlichen) Indexierer - evtl. unter Heranziehung eines 'Thesaurus' intellektuell (z. B. durch Lesen des Textes) als wichtig für die Suche ermittelt werden. In der Mitte des Funktionsspektrums liegt das Voll- oder Freitextretrieval: hierbei ist der gesamte (Quell-)Text elektronisch gespeichert und kann mit geeigneten Verfahren zur Suche erschlossen werden. Am anderen Ende des Anwendungsfeldes stehen die Frage-Antwort-Systeme und Verfahren der

Textkondensation bzw. -komprimierung (z. B. des automatischen Abstracting). Hierbei werden - dem Anspruch nach - Texte 'wissensbasiert' erschlossen bzw. auch Fakten im natürlichsprachigen Dialog erfasst. (Der letztgenante Komplex wird hier ausgeklammert; vgl. dazu Art. 55, 57). Für diesen Beitrag verbleibt also der Bereich der eher 'oberflächigen' computerlinguistischen Methoden und Verfahren, deren Nachteile (geringere linguistische Fundierung) durch die Vorteile der relativen Praxisnähe (d. h. der Machbarkeit auch unter Kosten- und Zeitaspekten) einigermaßen aufgewogen werden. Hierbei lassen sich im wesentlichen die folgenden Funktionsbereiche unterscheiden

- *Indexierung* von Dokumenten/Suchanfragen
- *Gewichtung* von Dokumenten und Deskriptoren;
- *Relationierung* von Begriffen und Dokumenten;
- *Interaktion* zwischen System und Benutzern.

2.1. Indexierungsfunktion

Unter Indexierung wird eine einfache Abbildungs- und Kondensierungsfunktion verstanden. Ziel ist es, den Inhalt eines Textes/Dokuments auf (wesentliche) Begriffe abzubilden. Bei Retrieval, d. h. der Suche nach Texten/Dokumenten, können die Begriffe (Terme/Termini) einzeln oder auch kombiniert verwendet werden.

Formal gesehen ist damit ein Text im einfachsten Falle repräsentiert durch eine Menge von Begriffen; die Texte klassifizieren sich als zu einer Menge von Suchbegriffen zugehörig oder nicht. Zum Suchen werden in den sog. kommerziellen IR-Systemen in der Regel die booleschen Operatoren AND, NOT, OR oder XOR und/oder die sog. Kontextoperatoren verwendet. Die Kontextoperatoren stellen dabei eine Einengung der AND-Verknüpfung dar. Eine Verbesserung gegenüber der booleschen Suchlogik bringt die sog. lineare pseudo-boolesche Suchlogik (vgl. z. B. Panyr 1986 a).

Die Indexierung (d. h. die Text- bzw. Dokumentrepräsentation) muss sich jedoch nicht auf eine einfache Aneinanderreihung von Termen beschränken, vielmehr lassen sich Begriffe auch verknüpfen; die Verknüpfungen können dabei einfach einen engen (assoziativen) Zusammenhang zwischen den jeweiligen Begriffen bezeichnen oder aber durch entsprechende Relatoren typisiert werden (vgl. u. a. Konzepte von Fugman 1975 und Austin 1976) sowie die Vorgehensweise bei der Pressedokumentation von Gruner & Jahr). In der Dokumentation (vgl. DIN 31623) wird die Indexierung sogar etwas weiter gefasst und (sinngemäß) als die Gesamtheit der Methoden verstanden, die der Zuordnung von Deskriptoren und Notationen zu Dokumenten *zwecks ihrer inhaltlichen Erschließung und gezielten Wiederauffindung dient*. Indexierung kann also mehr sein als eine Wortselektion bzw. -extraktion: die Abbildung auf (künstliche) *Notationen*, die ganze Themenbereiche charakterisieren (z. B. die Notation der Internationalen Patentklassifikation oder die Dezimalklassifikation) gehört ebenso dazu wie die Vergabe von Begriffen, die nicht notwendig in 'physischer' Form in dem Text, Abstract oder Titel eines Werkes auftreten (*Schlagwörter*).

Für die *maschinelle bzw. maschinengestützte Indexierung* liegt es zunächst nahe, ähnliche Kriterien anzusetzen, auch um den Vergleich unterschiedlicher Verfahrensweisen zu erleichtern. Doch gerade hier wird der Zusammenhang zwischen Indexierung und Retrieval besonders deutlich: Die Kombination zwischen Erschließungs- und Suchverfahren charakterisiert erst im eigentlichen Sinne ein IR-System. Sofern - etwa im Rahmen einer syntaktischen Analyse - bereits bei der Tex-

terschließung (morphologisch-syntaxorientierte) Begriffsbeziehungen ermittelt werden, kann in gewissen Grenzen auch von einer 'Präkoordination' gesprochen werden, die beim Retrieval im Blick auf eine höhere Präzisierung der Suchanfrage (mit) genutzt werden kann. Maschinelle Verfahren und die intellektuelle Vorgehensweise müssen demnach zu gleichartigen Ergebnissen führen.

Dafür lässt sich ein einfaches Beispiel anführen: Nehmen wir an, ein Thesaurus verzeichne als zu benutzenden Deskriptor das Wort "Nießbrauchrecht" (ohne Fugen-s), im Text steht jedoch "Nießbrauchsrecht" (mit Fugen-s). Bei der intellektuellen Indexierung wird nun - um die Suche mit Hilfe eines Deskriptors von Belegstellen zu erleichtern - eine (hier rein pragmatische) Normierung vorgenommen, ggf. steht im (gedruckten) Register auch ein Verweis auf die normierte Form. Bei der maschinellen Indexierung könnte jeweils die Originalschreibweise durchaus als Registereintrag beibehalten werden, wenn beim Retrieval bei Verwendung der alternativen Schreibweise auch das Synonym (unter Verwendung einer Begriffsrelation) mit berücksichtigt würde. Hierdurch verlagert sich der Entscheidungsprozess auf den *Retrievalvorgang* mit dem Vorteil, dass ggf. vom Benutzer zusätzlich noch zwischen den Alternativen variiert werden kann, und mit dem Nachteil, dass bei jeder Interaktion (d. h. bei wiederholtem Suchen) jeweils dieser Prozess erneut angestoßen werden muss, was notgedrungen zu größeren (wenn auch heutzutage fast vernachlässigbaren) Rechenzeiten führen kann. Man könnte natürlich beide Komponenten auch kombinieren und z. B. Deskriptoren differenzieren nach extrahierten Elementen ('Stichwörtern') und addierten Begriffen ('Schlagwörtern').

2.2. Gewichtungs- und Klassifikationsfunktion

Was für das Prinzip der Deskriptorvergabe gilt, ist analog auch auf die intensivere Verarbeitung von Dokumenten/Texten anzuwenden.

2.2.1. Gewichtung

Unter Gewichtung wird zunächst die Differenzierung von Deskriptoren nach der Bedeutung verstanden, die sie für die Suche bzw. den Suchenden erlangen können. Dies setzt im Prinzip voraus, dass Erfahrungen (von Experten bzw. Nutzern) eingebracht, dass sog. Erwartens-Erwartungen (ein Begriff, der üblicherweise in der Massenkommunikation auf den Publikumsgeschmack angewendet wird) einbezogen werden. Man bewegt sich hier bei der intellektuellen Indexierung auf schwankendem Boden; umgekehrt gilt jedoch, dass der Nutzer großer Datenbestände ohne eine Gewichtung überhaupt nicht mehr in der Lage wäre, sich in der Materialfülle auch nur halbwegs zurechtzufinden, d. h. zu in angemessener Zeit verwertbaren Ergebnissen zu kommen.

Die einfachste Methode einer Gewichtung - zugleich eine nicht-revidierbare - wird bei der traditionellen intellektuellen Indexierung verwendet: die Vergabe bzw. Nichtvergabe eines möglichen Deskriptors. Dem entspricht (ohne dass man inhaltlich vergleichen darf) in der maschinellen Indexierung die Tilgung hochfrequenter bzw. trivialer (d. h. auch im Fachgebiet kaum selektionswirksamer) Wörter (sog. Stoppwörter): meist lassen sich auf diese Weise gut 50 % der Wörter eines Textes (als trivial, d. h. für die spätere Suche nicht entscheidend) entfernen, den Rest bilden - bei dieser häufig angewendeten Verfahrensweise - die (maschinellen) 'Deskriptoren'.

Dieses Verfahren lässt sich zunächst rein statistisch verfeinern: Man kann Textwörtern in Abhängigkeit von der Häufigkeit ihre Auftretens im Text im Verhältnis zu ihrem Auftreten im Doku-

mentenbestand ein Gewicht zuordnen (vgl. bereits Harter 1975). Auch hierbei kann das Ergebnis zur physischen Tilgung von Wörtern, d. h. zu eine Nichtvergabe, führen (wenn sie z. B. unter einem bestimmten *Schwellenwert* - Cut-Off-Wert - liegen) oder aber zu einer entsprechenden Markierung, die dann beim späteren Suchen für ein *Ranking*, d. h. eine Ranganordnung der Ergebnisse, genutzt werden kann.

Bei der am Dokumentenbestand orientierten Gewichtung muss berücksichtigt werden, dass sich die Gewichte mit dem Anwachsen dieses Bestandes verschieben können. Dies führt technisch gesehen zu einigen Problemen, da die Werte im Prinzip bei jedem neu hinzukommenden Dokument/Text auch neu berechnet werden müssten. Zumindest müssen in gewissen Zeitabständen bei wachsenden Beständen solche Berechnungen neu durchgeführt werden. Ein Nebeneffekt liegt aber gerade darin, dass dadurch - vorausgesetzt, der Datenbestand ist repräsentativ für ein Fachgebiet - Trends und Tendenzen der Forschung und Entwicklung formal festgestellt werden können, eine Methode, die bislang allerdings wegen der technischen Grenzen noch wenig praktisch eingesetzt wurde. Dies wird sich jedoch mit größter Wahrscheinlichkeit ändern, da die Barrieren 'Zeit' und 'Kosten' aufgrund der wachsenden Leistungsfähigkeit der Computer zunehmend entfallen.

Ein interessanter Aspekt - sowohl in der Forschung als auch für die Praxis - ist die Frage, das *Kondensierungsprinzip* zu simulieren bzw. nachzubilden, das der intellektuellen Indexierung zugrundeliegt. Ziel eines solchen Verfahrens (vgl. Lustig 1979 und - darauf bezogen - Knorz 1983 in Bezug auf das AIR-Projekt) ist es, mit automatischen (vorwiegend heuristisch-mathematischen) Abbildungsverfahren in etwa auf die gleichen Deskriptoren zu kommen, die bei dem intellektuellen Textverstehen gegeben werden. Dies ist insofern mehr als ein reines 'Gewichten' von im Text vorhandenen Termini, als (über intellektuell erstellte Thesauri, d. h. ein Netzwerk von Begriffsrelationen) auch *nicht* im Text vorgefundene Deskriptoren erstellt werden. Auch hierbei wird eine Rangfolge erstellt und ein Schwellenwert benutzt, um Begriffe zu qualifizieren bzw. von der Vorgabe auszuschließen. Im o. g. AIR-Projekt wird zu dieser Nachbildung von Relationen eine repräsentative *intellektuell* indexierte Dokumentenmenge benötigt.

Einen interessanten Aspekt in die Theorie und Praxis von IR-Systemen bringen die sog. probabilistischen Modelle der Indexierung und des Retrieval mit sich (vgl. Maron/ Kuhns 1960; Robertson 1977; van Rijsbergen 1977; bzw. Robertson/Maron/Cooper 1982). Zu diesen Ansätzen können auch die Überlegungen zur nutztheoretischen Indexierung nach Cooper/Maron (1978) gezählt werden (vgl. zu beiden Ansätzen und zum o. g. Modell Harters Panyr 1986b). Diese Modelle betrachten die Relevanz der wiedergewonnenen Dokumente als eine Zufallsgröße und unterscheiden daher zwischen einem sog. Relevanzgrad und einer Relevanzwahrscheinlichkeit. Während der Relevanzgrad die Übereinstimmung (bzw. die Diskrepanz) zwischen dem gefundenen Dokument und der Suchfrage ausdrückt, soll die Relevanzwahrscheinlichkeit die Unterschiede (bzw. Übereinstimmung) zwischen der Formulierung der Suchfrage und dem tatsächlichen Benutzerbedürfnis ausdrücken.

Während die statistischen Verfahren (wie z. B. das Verfahren von Harter 1975) die Indexierung einer Dokumentenkollektion als einen ungeteilten Prozess sehen, betrachten die Vertreter der probabilistischen bzw. nutztheoretischen Indexierung diesen Prozess lediglich als Indexierung der einzelnen Dokumente (in der Reihenfolge ihrer Ankunft) in Bezug auf die potentiellen Benutzerbedürfnisse. Die Durchführung der probabilistischen Indexierung (oder auch des Retrievals - vgl. Robertson/Sparck Jones 1976) drückt sich in der Zuteilung von sog. Relevanzgewichten zu

einzelnen Termen (Deskriptoren oder Suchfragetermen) aus. Die benötigten Informationen zur Berechnung dieser Gewichte können approximativ mit Hilfe eines Relevanzfeedback-Verfahrens (s. u.) gewonnen werden. Das Modell unterliegt verschiedenen Einschränkungen (vgl. Panyr 1986b).

2.2.2. Klassifikation

Bei der (heute noch intellektuell durchgeführten) Klassifikation werden (Fach-)Texte sozusagen in eine (oder auch mehrere) Themenschublade(n) abgelegt. Dies ist in der Regel eine extrem grobe Zuordnung: Man kann es sich weitgehend so vorstellen, dass jemand so tut, als sei er gezwungen, eine Akte in einem einzigen Ordner abzulegen (wie jeder weiß, ein fast unmögliches Unterfangen, wenn ein Thema in der Veröffentlichung unter alternativen Gesichtspunkten behandelt wird). Dieser Ordner ist nun allerdings in der Regel weiter strukturiert (hierarchische Gliederung), gelegentlich muss auch nach Aspekten differenziert werden (Facettenklassifikation). In der Theorie lässt sich zeigen, dass der Prozess der intellektuellen Zuordnung von Texten zu einem Themenbereich (dessen Formalisierung allerdings noch weitgehend aussteht) durch maschinelle Verfahren relativ gut nachgebildet werden kann.

Eine in diese Richtung weisende Verfahrensweise wird (meist rein intellektuell, was diese Stufe angeht) bereits bei der sogenannten Inhaltsanalyse (*Content Analysis*) zugrundegelegt: Die Wörter, die z. B. in einem psychotherapeutischen Gespräch von einem Klienten/Patienten gebraucht werden, werden mit einem Inhaltsanalyse-Wörterbuch verglichen. Dieses Wörterbuch ist meist intellektuell erstellt und gepflegt und enthält im wesentlichen lexikalisierte Hypothesen derart: wer die Wörter al, a2 oder a3 gebraucht, drückt damit (versteckt) eine bestimmte Emotion E (z. B. Angst) aus, wer b1 oder b2 oder b3 sagt, befindet sich im Zustand Z (fühlt sich z. B. unter Druck gesetzt) usw. Somit lassen sich - vorausgesetzt, die Hypothesen stimmen und die Datenmenge ist ausreichend - durch statistische Auswertungen Themen- und Problemfelder ermitteln. Analog könnte man etwa in der Patentklassifikation (IPC) verfahren; Durch eine Auswertung der zu einer bestimmten IPC-Kategorie zugeordneten, im Text vorhandenen Stichwörter lässt sich eine Klassifikation, zumindest aber ein Klassifikationsvorschlag ermitteln.

Bei thematisch heterogenen Dokumentenmengen wird es sinnvoll zu versuchen, die Texte (Dokumente) automatisch zu klassifizieren. Man spricht in diesem Zusammenhang auch vom sog. Dokumenten-Clustering. Als Motivation für diese Vorgehensweise gibt Salton (1968) Effizienzgründe an. Nach von Rijsbergen (1979) bringt die Anwendung von Dokumenten-Clustering prinzipiell auch eine Effektivitätsverbesserung mit sich. Da die einzelnen Dokumenten-Cluster durch Deskriptoren identifiziert sind, können sie auch als Ersatz für intellektuell gewählte Klassen im obigen Sinne dienen. Eine Kombination des Dokumenten-Clustering mit einer intellektuell vorgegebenen Klassifikation ist denkbar. Der Benutzer kann auch die gewonnene Clustermenge durch ein Relevanzfeedback-Verfahren (s. u.) nachträglich modifizieren (vgl. Panyr (1986 a)).

2.3. Relationierung von Begriffen und Dokumenten

Die Herstellung bzw. Darstellung von Beziehungen zwischen Begriffen (genauer: ihren Benennungen) und zwischen Dokumenten (genauer: zwischen Texten/Titeln/Abstracts) dient dazu, thematisch identische bzw. ähnliche Sachverhalte in einem (textuellen) Datenbestand zu identifizieren und damit bei der Recherche die *Quantität* der (relevanten) Ergebnisse - bezogen auf eine Suchanfrage bzw. gewünschte Problemlösung - zu steigern.

2.3.1. Begriffsrelationierung (Thesaurus)

Ein klassisches - d. h. schon in der traditionellen Dokumentation genutztes - Instrument der Begriffsrelationierung ist der Thesaurus. Dabei lassen sich - grob gerechnet - zwei Grundlinien aufzeigen: eine linguistisch-sprachphilosophisch motivierte und eine praktisch-ablagetechnisch begründete Richtung (die sich im Endeffekt überlappen können).

Es gab in der Linguistik schon immer Versuche, die Begriffswelt zu gliedern. Roget's Thesaurus ist hier u. a. zu nennen, aus dem sich viele Variationen (z. B. fürs Deutsche: Wehrle-Eggers, Dornseiff) entwickelten. Ein derartiges System begrifflicher Vernetzungen, meist kombiniert mit einer hierarchischen (Welt-)Gliederung, ist naturgemäß (ähnlich den traditionellen Klassifikationssystemen) geprägt von einer (meist wenig empirisch abgesicherten) spezifischen Weltansicht und läuft Gefahr, sich letztlich im Detail, d. h. in der Performanz der lexikalischen Realisierung, in wenig definierten, eher subjektiv-assoziativen Verknüpfungen zu verlieren.

Demgegenüber gibt es sehr wohl thematisch-formal begründete Definitionen begrifflicher Relationierungen wie Synonymie, Ober- und Unterbegriff usf. (vgl. Hutchins 1975). In der praktischen Dokumentation und Information haben *fachspezifische* Thesauri einen auch bis in die Datenbankwelt reichenden Stellenwert erhalten bzw. bewahrt. In der traditionellen Dokumentationspraxis erfüllt ein Thesaurus zunächst eine *Normierungs*funktion, z. T. aufgrund der volumenmäßigen Begrenzungen für den Registerteil einer Dokumentation, z. T. wohl auch aus fachterminologischen ('didaktischen') Erwägungen. Andererseits unterstützt er das Prinzip und den Prozess der Indexierung, nämlich über eine hierarchische Relationierung von wenig spezifischen, d. h. kaum dokumentdifferenzierenden Begriffen zu spezifischen Begriffen und umgekehrt von allzu spezifischen (den Themenbereich in der Praxis zu sehr einschränkenden) Begriffen zu weniger spezifischen Begriffen hinzuführen. Hinzu kommt die Synonym-Relation, deren Effizienz angesichts der (oberflächigen) Ausdrucksalternativen in den Texten ebenso unbestritten ist wie - zumeist in der Theorie - die Vermeidung von Mehrdeutigkeiten eines (Fach-)Begriffs durch Überführung in eine quasi-kunstsprachliche Form. Dies geschieht i. d. R. durch Selektion einer 'Bedeutungsvariante' als *verbindliche* Interpretation und 'Verbot' der Verwendung des Wortes (genauer: der Benennung) als Deskriptor, wenn es im Dokument in einer anderen Bedeutung vorkommt.

Mit der Verwendung der elektronischen Speicherung und Datenverarbeitung wird diese Art der Thesaurus-Nutzung zunehmend relativiert. Die intellektuelle Pflege eines Fachthesaurus ist zeitaufwendig, bislang wird die Zuordnung der (meist geringen) Zahl von Deskriptoren rein intellektuell vorgenommen (vgl. auch das AIR-Experiment von Knorz und Lustig: Knorz 1983). Demgegenüber sind Freitext- bzw. Volltextverfahren nahezu in allen Datenbanken verfügbar und erbringen offenbar (dies zeigten schon frühere Studien von Cleverdon/Keen 1966) beim Retrieval vergleichbare Ergebnisse.

2.3.2. Dokument- und Deskriptor-Clustering

Die Klassifikation (vgl. 2.2.2.) besitzt bereits im Dokument-Clustering, d. h. der *thematischen In-Beziehung-Setzung* von Dokumenten, ein zentrales Instrument. Mit der hierarchischen Klassifikation lassen sich daher entsprechend klassierte Dokumente als (thematisch) synonym (d. h. der gleichen Klasse zugeordnet) oder nebengeordnet (d. h. der gleichen Klasse untergeordnet) usw.

charakterisieren. Aber auch ohne eine derartige (intellektuelle) Vorabzuordnung sind Methoden vorstellbar, die z. B. zum Zeitpunkt des Retrieval die Suche nach ähnlichen Dokumenten im Datenbestand ermöglichen. Hierbei werden verschiedene Funktionen genutzt, z. B. das Auftreten bzw. auch die Auftretenshäufigkeit eines Wortes/Deskriptors.

Dieses Ziel verbirgt sich hinter dem Konzept eines Deskriptoren-Clustering (d. h. der Termklassifikation): Hierbei werden die Deskriptoren anhand ihres Auftretens in Dokumenten in sog. Deskriptoren-Clustern (Termcluster) zusammengeschlossen. Diese Vorgehensweise sollte die Effektivität des IR-Systems erhöhen.

Die Anwendung der Termklassifikation ist allerdings umstritten. Die Erweiterung des Suchauftrags um die ermittelten Termklassen führte nämlich häufig zu schwer interpretierbaren und unerwarteten Ergebnissen. Durch die gleichzeitige Term- und Dokumentenklassifikation wurde dann ein Versuch gemacht, die beiden Motivationsgründe (d. h. Effektivitätsverbesserung beim Termclustering und Effizienzverbesserung beim Dokumentenclustering) in Einklang zu bringen (vgl. Panyr 1986 a). Eine modifizierte Clusteranalyse wird unter der Bezeichnung 'Konzeptionelles Clustering' (*conceptual clustering*) zunehmend auch für die automatische Wissensextraktion in den sog. Expertensystemen verwendet (vgl. Panyr 1987 c.)

2.4. Relevanzfeedback

Bei der Evaluierung von Rechercheergebnissen mit dem MEDLARS-System der *National Library of Medicine* wurde festgestellt, dass ca. 60 % der *Recall*-Fehler (d. h. der Fehler, die zur Nicht-Anzeige relevanter Dokumente führen) mit der Recherchevorgehensweise zusammenhängen, während bei der *Precision* (d. h. der Fehler, die zur Angabe irrelevanter Dokumente führen) der Anteil der Recherchefehler ca. 50 % ausmacht (vgl. Lancaster 1979, 147 ff.). Der größte Anteil dieser Fehler ist der mangelhaften bis fehlenden Interaktion zwischen Benutzer und System oder einer falschen bzw. unzureichenden Suchfrageformulierung zuzuschreiben. Um die Effektivität der Recherche zu verbessern, sind Verfahren entwickelt worden, die Relevanzurteile von Benutzern als Feedback-Information (zur erneuten Suche also) verwenden. Diese sog. Relevanzfeedback-Strategien (RF-Strategien) stellen somit einen Versuch dar, den IR-Benutzer direkt in die Retrievalstrategie einzubeziehen. Der Benutzer muss dabei die wiedergewonnenen Dokumente ihrer Relevanz nach beurteilen. Auf dieser Grundlage führt das IR-System ein modifiziertes Retrieval durch und bietet weitere Dokumente zur Bewertung an. Dieser Prozess kann mehrmals wiederholt werden, bis der Benutzer zufrieden ist oder bis kein
weiteres relevantes Dokument mehr gefunden wird.

Bei Relevanzfeedback-Verfahren wird im allgemeinen zwischen der Modifikation der Suchfrage und der Modifikation des Dokumentenraumes unterschieden. Die beiden Ansätze können auch zu einer sog. hybriden RF-Strategie kombiniert werden. Bei der Modifikation der Suchfrage wird unterstellt, dass die Benutzerfrage (Suchfrage) unklar oder ungenau ist (z. B. infolge nichteindeutiger Suchargumente). Bei der Modifikation des Dokumentenraumes wird darüber hinaus noch angenommen, dass die unbefriedigenden Retrievalergebnisse aufgrund der Auswahl eines fehlerhaften Dokumentenraumes (z. B. durch falsche Indexierung) zustande gekommen sind. Bei den Modifikationsstrategien wird weiterhin zwischen RF-Strategien für einen unklassifizierten und RF-Strategien für einen vorklassifizierten Dokumentenraum unterschieden. Es wird ferner angenommen, dass die wiedergewonnenen Dokumente in der Reihenfolge ihrer (systemseitigen) Relevanz (= Relevanzgrad) dem Benutzer angeboten werden (*Ranking*). Diese (letztere) Vorausset-

zung ist bei Systemen, die nur mit der klassischen booleschen Suchlogik ausgestattet sind, nicht erfüllt.

Bei den Relevanzfeedback-Strategien wird daneben zwischen einem sog. positiven und einem negativen Feedback unterschieden. Beim positiven Feedback werden zur Modifikation nur die Dokumente verwendet, die als relevant eingestuft wurden. Beim negativen Feedback erfolgt die Modifikation auch mit der Hilfe der als nichtrelevant bezeichneten Dokumente. Eine solch 'negative' Technik kommt insbesondere dann zur Anwendung, wenn bei der Suche nur nichtrelevante Dokumente gefunden werden.

Die Begründung für die RF-Vorgehensweise zur Modifikation des Suchauftrags liegt in der Interpretation der Anforderungen an die sog. optimale Suchfrage; diese kann durch Kenntnis der Mengen von relevanten und nichtrelevanten Dokumenten konstruiert werden. In der Praxis sind diese Mengen jedoch unbekannt. Der Benutzer kann also keine solche optimale Suchfrage stellen. Sie kann jedoch approximiert werden, indem man einen Teil der relevanten und nichtrelevanten Dokumente identifiziert. Es gilt also, eine Prozedur zu finden, die mit Hilfe des Relevanzurteils nach einem initialen Retrieval eine verbesserte Suchfrage iterativ konstruiert (vgl. auch Panyr 1987 b).

Ungeachtet der Unterschiede in der Auffassung des Relevanzbegriffs bei Salton und seinen Mitarbeitern (um 1965) und den Vertretern der probabilistischen Modelle (s. o.), kann das Relevanzfeedback als die typische probabilistische Retrievalstrategie bezeichnet werden (vgl. auch Harper 1980). Die eigentliche Vorgehensweise entspricht u. a. auch der Vorstellung der probabilistischen Retrievaltheorie über den approximativen Charakter der Retrievalprozesse. In den kommerziellen IR-Systemen sind die RF-Verfahren bisher nicht implementiert. Eine (wenn auch nur ansatzweise realisierte) Ausnahme stellte das IR-System TELDOK (implementiert auf TR440-Anlagen von Telefunken Siemens) dar, in dem eine rudimentäre RF-Vorgehensweise durch die Einbeziehung der Hintergrunddeskriptoren der wiedergewonnenen relevanten Dokumente realisiert ist.

3. Computerlinguistische Lösungen

Betrachtet man die derzeit aktiven (kommerziellen) Systeme, genauer gesagt: die weltweite Praxis des IR, so sind computerlinguistische Verfahren selbst dort, wo im Prinzip die technisch-linguistischen Fragestellungen gelöst erscheinen (z. B. in der Morphologie), absolut defizitär. Eine Ausnahme macht hier das System GOLEM (Wz) mit dem computerlinguistisch fundierten System PASSAT (Wz; Hersteller ist in beiden Fällen die Siemens AG). Sowohl in der deutschen Rechtsdokumentation (JURIS) als auch in der deutschen Patentdokumentation wird PASSAT bei der Freitexterschließung herangezogen. Daher ist es sinnvoll, die Funktionsweise dieses Systems kurz vorzustellen: Im Text (Titel/Abstract/Volltext) vorkommende (Einzel-)Wortformen werden auf ihre mögliche(n) Grundform(en) abgebildet (z. B. *'Häusern => Haus'*); Wortzusammensetzungen (z. B. *'Haustür'*) werden darüber hinaus zusätzlich in ihre (sinnvollen) Bestandteile zerlegt. Alle Variationen (hier *'Tür, Haus, Haustür'*); werden als Freitextdeskriptoren verfügbar gemacht.

Es ist jedoch bezeichnend für die ganze Situation, dass von dieser lokalen Ausnahme (die vielleicht mit der morphologischen Komplexität des Deutschen begründet werden konnte) u. a. in international zugänglichen Datenbasen keine erwähnenswerten *computerlinguistischen* Verfahren

zum Einsatz kommen. Die Gründe dafür sind kaum eindeutig ermittelbar. Einige Argumente sollen jedoch ohne Anspruch auf Gewicht und Vollständigkeit kurz genannt werden:

- Die Anwendung computerlinguistischer Verfahren setzt eine gewisse Durchgängigkeit voraus: Es hat im internationalen Bereich wenig Sinn, wenn für *eine* Sprache und für *ein* Softwarepaket eine Funktion, wie z. B. die Grundformenermittlung, existiert, für andere Sprachen und Rechnersysteme jedoch nicht.

- Die internationale Fach-Kommunikationssprache ist heute Englisch. Englisch ist (relativ) flexions- und kompositionsarm, so dass einfache technische Hilfen, wie z. B. die *Trunkierung* (d. h. das Weglassen von Buchstabenfolgen u. a. am Wortende) und die *Adjacency* (d. h. die logische Verknüpfung von *aufeinander folgenden Wörtern*) einen Ersatz für morphologisch-linguistische Verfahren darstellen.

- Aus (software-)ökonomischen Gründen werden in einer Datenbank *verschiedensprachige* Daten kumuliert, so dass (überwiegend) englische neben deutschen, französischen usw. Dokumenten stehen. Die intellektuelle Verschlagwortung und ggf. die intellektuelle Klassifizierung werden in der Praxis für ausreichend empfunden, wenn man davon absieht, dass nach wie vor Datenbanken noch Akzeptanzprobleme haben.

- Die Implementierung weitergehender linguistischer Verfahren ist letztlich nach wie vor eine Kostenfrage: Softwaretechnologisch sind z. B. erweiterte Systeme zur Verwaltung und Nutzung (d. h. Integration) von Thesaurusrelationen erforderlich, verbunden mit dem erweiterten Funktionsinventar angepasster Display- und Verknüpfungsalgorithmen.

- Die zur Reduktion von Fehlern (*Unterindexierung*: nicht identifizierte, aber thematisch relevante Deskriptoren, *Überindexierung*: Vergabe von Deskriptoren, die jedoch thematisch nicht relevant sind) erforderlichen Verfahren (z. B. zur Auflösung von Mehrdeutigkeiten, zur Begriffsrelationierung) liegen trotz der inzwischen in Laborsystemen und Experimentalumgebungen erzielten Fortschritte (man vgl. die Untersuchungen in Deutschland anhand der Systeme CONDOR und CTX: u. a. Panyr 1986 a und Zimmermann/Kroupa/Keil 1984) nicht für die Praxis, d. h. die breite Anwendung aufbereitet vor, ganz zu schweigen von entsprechenden Entwicklungen für andere Sprachen.

Die derzeitige Welt der IR-Systeme ist demgemäß beherrscht von Verfahren, die die 'Last' des Retrieval (im Freitextbereich) weitgehend dem Nutzer/Recherchierenden überlassen: er muss ggf. 'per Hand' mit Synonymen aus anderen Sprachen suchen, er muss z. T. komplizierte Trunkierungs- und Abstandsverfahren benutzen usw. Ein *Ranking* der gefundenen Dokumente ist allenfalls nach dem Datum der Veröffentlichung möglich, inhaltliche Vergleiche sind somit so gut wie ausgeschlossen. Aus der Sicht des Anbieters bzw. des Datenbankbetreibers belasten zudem solche Verfahren die Rechner erheblich (verglichen mit dem weitgehend passiven Suchverfahren).

Trotz der technischen Entwicklungen der letzten Jahre, die ein Retrieval nicht nur auf großen Rechenanlagen, sondern auch auf dem PC erlauben, ist ein Fortschritt unter verstärkter Einbeziehung computerlinguistischer Verfahren kaum erkennbar: Bei dieser Miniaturisierung, d. h. softwaretechnologischen Übertragung von Großrechnerlösungen auf die Mini- und Mikrorechnerwelt, werden zunächst die bestehenden Methoden, die sich auf Großrechnerebene 'bewährt' haben, migriert. Zudem bestehen noch erhebliche Beschränkungen z. B. bei der Zugriffszeit auf die

CD-ROM, auch die Festplattenkapazität der PC-Peripherie ist demgegenüber noch begrenzt, so dass die Implementierung weitergehender Verfahren z. Zt. kaum betrieben wird.

Sieht man daher von der Siemens-Entwicklung GOLEM (mit PASSAT) ab, so ist im kommerziellen Bereich eine wichtige Leitlinie mit dem IBM-System STAIRS gesetzt; viele andere gängige Systeme (wie STN-Messenger oder FAIRS) sind analog zu STAIRS gestaltet (mit einem meist wenig anwenderfreundlichen Änderungsdienst). Bei dem IR-System UNIDAS (Fa. UNISYS) ist eine starke Thesaurus-Funktion ausgebaut. Die Dokumente sind allerdings intellektuell zu indexieren. GRIPS/DIRS läuft wiederum auf Siemens-Anlagen. Es zeichnet sich zwar durch eine größere Flexibilität aus, der Anwender muss jedoch mit den GRIPS-Sprachmitteln die Datenstruktur und die Zusammenhänge zwischen Dokument- und Deskriptordatei für jede Datenbank sehr aufwendig beschreiben.

STAIRS - das 'Vorbild' der kommerziellen IR-Welt - kennt bereits die Trunkierung sowie die wort- und satzbezogenen Abstandsmaße als wichtige Ersatzfunktionen für sprachmorphologische Verfahren. Es ist in diesem Zusammenhang bezeichnend, dass auch die IBM-Eigenentwicklung einer Grundformsuche mit STAIRS-TLS kaum zum Einsatz kommt.

Von den IR-Systemen für Mikrocomputer soll hier noch das IR-System SIRE erwähnt werden. Das ursprünglich als ein Labormodell von der Syracuse University (N. Y.) konzipierte System wurde später von der Firma KNM, Inc., zum kommerziellen IR-System für Mikrocomputer modifiziert. SIRE ist ein System, das die Merkmale der konventionellen Systeme (d. h. invertierte Dateien, boolesche Anfragen) besitzt. Noreaut/Koll/ McGill (1978) haben gezeigt, dass auch ein solches System durch eine geringere Modifikation der Dateiorganisation Ähnlichkeitsfunktionen zu Vergleichszwecken (Dokument mit Suchfrage) verwenden kann und dadurch auch ein Ranking der wiedergewonnenen Dokumente möglich wird.

4. Laborsysteme

An Hochschulen und in Industrielabors ist demgegenüber eine Reihe von Systemen entwickelt worden, die einen Großteil der in 2. dargestellten Funktionen realisiert haben. Das in der Bundesrepublik Deutschland wohl am weitesten entwickelte (inzwischen jedoch nicht mehr weiter verfolgte) System in diesem Zusammenhang ist zwischen 1973 und 1981 von Siemens unter dem Namen CONDOR experimentell erprobt worden. Es umfasste u. a. eine syntaxorientierte Feinrecherche (mit Ranking; vgl. Wieland 1979) und ein automatisches Klassifikations- und Clusterverfahren mit einer Relevanz-Feedback-Komponente beim Retrieval (vgl. Panyr 1986 a).

Als eine z. T. (computer)linguistische Methode kann auch die Adaption der Theorie der sog. Fuzzy-Mengen (unscharfe Mengen) für eine IR-Anwendung einbezogen werden. Die Fuzzy-Logik (insbesondere in Bezug auf das Einführen der sog. linguistischen Variable im Bereich der natürlichsprachlichen Systeme) kann auch als ein Versuch betrachtet werden, die Schwachstellen der logisierenden Sprachbetrachtung, wie sie z. B. in den Arbeiten des Wiener Kreises, bei der Montague-Grammatik oder im Endeffekt auch bei der generativen Transformationsgrammatik zum Ausdruck kommen, zu mildern bzw. zu beseitigen. In IR-Systemen hängt die Anwendung der Fuzzy-Mengen mit der Unschärfe der Aussagen und Urteile bei der Formulierung der Anwenderprobleme (d. h. der Suchaufträge bzw. Suchfragen) sowie mit der Indexierungsproblematik zusammen. Das für die Anwendung der Fuzzy-Mengen typische Quantifizieren der qualitativen Aussagen drückt sich überwiegend lediglich in der Zuordnung numerischer Werte zwischen 0

und 1 zu den Index- und Suchfragentermen aus. Diese sog. Zugehörigkeitswerte (genauer: die Werte der sog. Zugehörigkeitsfunktion) werden bei den Fuzzy-theoretischen Modellen in der Regel heuristisch bestimmt; in vielerlei Anwendungen werden statistische Wahrscheinlichkeiten verwendet. Somit können jedoch die Begriffe wie 'Zugehörigkeitsgrad' im Prinzip nur als andere Bezeichnungen für die Wahrscheinlichkeitsgewichte der probabilistischen oder statistischen Modelle aufgefasst werden (eine kritische Übersicht über die Theorie und Literatur gibt Panyr 1986 c).

Das IR-System FIRST von Rank Xerox integriert - wie CONDOR - Datenbank- mit IR-Funktionen in einem System. Die textuellen Daten werden linguistisch verarbeitet und mehrstufig klassifiziert; bei der Wiedergabe der Dokumente kann die Ähnlichkeit zwischen der „Suchfrage" und den gefundenen Dokumenten zum Ranking verwendet werden (vgl. Datolla 1979).

An der ETH Zürich ist auf der PC-Ebene das experimentelle System CALIBAN entwickelt worden (vgl. Frei/Bärtschi/Jauslin 1985). Hierbei werden Ähnlichkeitsmaße beim Retrieval eingesetzt. Mit FAKYR wurde in Berlin (vgl. Bollmann 1983; Süß/ Leckermann 1981) ein Experimentalsystem entwickelt, das Methoden des Clustering und verschiedene Retrievalstrategien, u. a. ein Relevanz-Feedback-Verfahren, integriert. Besondere Bedeutung besitzen z. T. die Pionierarbeiten im Bereich der Evaluierung von IR-Systemen und der Planung von IR-Experimenten (vgl. Schneider/Bollmann/Jochum et al. 1986).

Das experimentelle IR-System SMART stellt das bisher bedeutendste Forschungsprojekt auf dem Gebiet der IR-Systeme dar. Es enthält viele Anregungen und Konzepte auf dem Gebiet der IR-Systeme. Viele Impulse auf dem Gebiet der Erschließung und Wiedergewinnung von Informationen kommen daher aus der 'SMART-Schule'. Die Bedeutung dieses Projekts liegt auch in der großen Zahl der durchgeführten Experimente (vgl. Salton (ed.) 1971; 1975; Salton/McGill 1983). Im Rahmen von SMART wurden u. a. die grundlegenden Arbeiten auf dem Gebiet der Dokumentenklassifikation und Relevanzfeedback-Strategien durchgeführt, die einen bis heute nachhaltigen Einfluss auf die Theorie der IR-Systeme ausüben. SMART wird heute an mehreren Universitäten eingesetzt und an der Cornell University in Ithaca fortgeführt.

5. Entwicklungstendenzen

Am ehesten haben in Zukunft - betrachtet man den Massenanfall von Daten - solche Entwicklungen Chancen, in die Praxis umgesetzt zu werden, die bei der *Überwindung der Sprachbarrieren* mithelfen. Mit der zunehmenden Internationalisierung des Zugangs zu Datenbasen ist jetzt bereits abzusehen, dass als Zugangs- und Kommunikationssprache Englisch von zentraler Bedeutung sein wird. Umgekehrt werden wohl aus verschiedenen Gründen die Nationalsprachen auch in der lokalen Fachkommunikation ihre Bedeutung behalten. Dies gilt insbesondere für den Transfer von Fachwissen in die Praxis.

Wenn man - auch aus ökonomischer Sicht - längerfristig mit Erfolg im Bereich der Wissensvermittlung (z. B. gerade in der internationalen Fachinformation und Bürokommunikation) mitwirken will, muss die Konsequenz sein, im internationalen Bereich mit englischsprachigen Datenbanken präsent zu sein, umgekehrt aber einen nationalsprachlichen Zugang sicherzustellen.

Auf diese Weise wird die maschinelle bzw. maschinengestützte Übersetzung in der Fachinformation, insbesondere aber beim IR-Prozess, erheblich an Bedeutung gewinnen. Dies setzt aller-

dings hierfür geeignete Werkzeuge bzw. Verfahrensweisen voraus. Zugleich wird die Entwicklung mehrsprachiger Thesauri bzw. elektronischer Übersetzungswörterbücher und deren Integration in den Retrievalprozess, schließlich die maschinelle (auch ggf. nur informationelle) Übersetzung eine wichtige Rolle spielen.

Ein *Szenario* für zukünftige IR-Prozesse - wohlgemerkt: allein bezogen auf die bisher dominante Welt der großen Informationsbanken (auch hier seien die Entwicklungen bei Expertensystemen ausgeklammert) - könnte dabei wie folgt aussehen:

(1) Titel und Schlagwörter/Deskriptoren liegen in jedem Fall in der Nationalsprache und in der internationalen Kommunikationssprache (Englisch) übersetzt vor. Damit und über die ggf. ebenfalls bereits übersetzten Freitextbegriffe erfolgt eine Recherche und Selektion im traditionellen IR-Prozess.

(2) Nach erster Sichtung der Ergebnisse unter Relevanzkriterien erfolgt ggf. ein *Translation-Ordering* (analog dem *Document-Ordering*), wobei entweder eine kostengünstige so genannte Informativ-Übersetzung oder aber eine (teuere) hochqualitative Übersetzung in der gewünschten Sprachrichtung erfolgt (die durch 'Human'-Übersetzer in einem Service-Zentrum nachbereitet wird).

(3) Das Szenario muss durch ein komplexes Relevanzfeedback-Verfahren ergänzt werden. Die Anwendung von RF-Methoden hängt auch mit der notwendigen Ablösung der konventionellen booleschen Suchlogik in kommerziellen IR-Systemen zusammen (Salton/Fox/Vorhees 1985 zeigen die Anwendung der RF-Technik auch für boolesche Suchfragen in einem erweiterten booleschen Retrieval).

Die Voraussetzungen für derartige Verfahrensweisen sind technisch weitgehend gegeben; ein *Document-Ordering*-Verfahren ist in vielen IR-Systemen bereits realisiert; das Verfahren des Packet-Switching (in der Bundesrepublik Deutschland: DATEX-P) erlaubt eine weltweit (relativ) kostengünstige Übermittlung, evtl. auch Zwischenübermittlung zum spezialisierten 'Übersetzungscomputer'. Das System, das z. Zt. hierfür am ehesten infrage kommt, ist SYSTRAN, dessen kommerzielle Rechte weltweit im wesentlichen bei Gachot S. A., Frankreich liegen. SYSTRAN liefert inzwischen Rohübersetzungen zu einer Reihe von Sprachpaaren, z. B. Englisch⇔Französisch, Englisch⇔Deutsch, Englisch⇔Italienisch, Englisch. ⇔Arabisch, Englisch⇔Japanisch. Ein großes Problem, das es hierbei zu bewältigen gilt, ist in erster Linie der Ausbau des Fachvokabulars und die fach- bzw. themenspezifische Disambiguierung. Ein weiterer Aspekt dürfte die Nutzung der Transferwörterbücher als Synonymwörterbuch bereits beim Retrievalvorgang sein. Unter internationalem Aspekt ist im Bereich der weltweiten Fachinformation eine Nutzung (englischsprachiger) (Text-)Datenbanken dann besonders wirksam, wenn auch der Zugang über die jeweilige Nationalsprache möglich ist. Die bedeutet u. a.:

- die Integration mehrsprachiger Thesauri in den Retrievalprozess;

- die Einbeziehung zunächst morphosemantischer (monolingualer) Begriffsbeziehungen in den Retrievalprozess (d. h. von Synonymen, Quasisynonymen, aber auch Flexions- und Derivationsformen zu einem Freitextelement ;

- die Verknüpfung von (Text-)Stichwörtern mit (systemseitig - intellektuell, evtl. zunehmend auch maschinengestützt - vergebenen) Klassifikationen, über die dann eine Ausweitung der Informationssuche möglich wird.

Im Bereich der betrieblichen Information und Kommunikation werden linguistische Verfahren ebenfalls eine flankierende Funktion haben. Bei betrieblichen Archivierungssystemen werden u. a. im formal-strukturellen Bereich Fortschritte erzielt, d. h. textuelle Dokumente über strukturelle-leategeniale Merkmale (z. B. Brief, Adressat, Postleitzahl) recherchierfähig. Dennoch ist auch hier zunehmend der Einsatz linguistischer Verfahren möglich. Bereits heute haben international die Stilhilfen (z. B. elektronische Synonymwörterbücher) einen gewissen Stellenwert bei der Textverarbeitung erlangt. Es ist sinnvoll, derartige Zentren bzw. lexikalische Verknüpfungen auch für die Datenablage bzw. den Retrievalvorgang zu nutzen. Dabei verwischen allerdings die Grenzen zwischen Textverarbeitung, Archivsystem und Datenbank. So ist es kein Wunder, dass z. T. linguistisch betrachtet 'schlichte' Systeme wie Q & A (in der deutschen Fassung: F & A) bereits unter dem Begriff 'Expertensystem' fungieren (vgl. die Paneldiskussion "Natural Interfaces - Ready for Commercial Success?" und den Beitrag von Hendrix 1986, 164), ohne dass im sprachdatenbezogenen Teil viel mehr geschieht als eine Trunkierung und eine boolesche Verknüpfung von Freitextwörtern.

Nach aller bisherigen Erfahrung werden die bestehenden kommerziellen bzw. marktorientierten Entwicklungen - die z. T. ihre Grundlage noch in den 60er Jahren haben - u. a. aufgrund der Sprödigkeit und Komplexität des Materials Sprache, der Vielfalt der Nationalsprachen und auch der Kostensituation die inhaltlichen Entwicklungen nur langsam fortschreiten. Daher wird es schon als ein großer Fortschritt erscheinen, wenn einmal das bereits jetzt schon weltweit technisch und einzelsprachlich verfügbare textuelle Wissen über ein Retrieval mittels der jeweiligen muttersprachlichen Terminologie zugänglich und im Ergebnis auch die Daten selbst in der jeweiligen Landessprache - wenn vielleicht auch nur informationell - verfügbar werden.

Daneben erscheint es möglich, entscheidende Fortschritte im Bereich der hochspezialisierten 'Mini-Welten' über natürlichsprachige Zugänge und Interaktionen in Frage-Antwort- und Expertensystemen zu erzielen. Darüber hinaus muss auch die Problematik der semantischen Repräsentationssprachen eingehender untersucht werden (s. o. - vgl. Jochum 1982).

6. Literatur (in Auswahl)

H. P. Frei/M. Bärtschi/J. F. Jauslin 1985 • U. Hahn/W. Reimer 1982 • P. S. Harter 1975 • H. Heilmann (ed.) 1987 • W. I. Hutchins 1985 • F. Jochum 1982 • G. Knorz 1983 • E. Kroupa 1983 • F. W. Lancaster 1979 • G. Lustig 1979 • J. Panyr 1986 a • J. Panyr 1987 • G. Salton 1968 • G. Salton (ed.) 1971 • G. Salton 1975 • G. Salton/M. McGill 1983; dt. 1987 • C. Süß/J. Leckermann 1981 • C. J. Van Rijsbergen 1979 • H. H. Zimmermann 1979.