Sprache und Information

S+I

Sprache und Information
Beiträge zur philologischen und linguistischen Datenverarbeitung, Informatik und Informationswissenschaft

Herausgegeben von
István Bátori, Walther von Hahn, Rainer Kuhlen, Winfried Lenders, Wolfgang Putschke, Hans Jochen Schneider, Harald Zimmermann

Band 12

Jiri Panyr

Automatische Klassifikation und Information Retrieval

Anwendung und Entwicklung komplexer Verfahren in Information-Retrieval-Systemen und ihre Evaluierung

Max Niemeyer Verlag
Tübingen 1986

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Panyr, Jiri:
Automatische Klassifikation und Information retrieval : Anwendung u. Entwicklung komplexer Verfahren in Information-retrieval-Systemen u. ihre Evaluierung / Jiri Panyr. - Tübingen : Niemeyer, 1986.
(Sprache und Information ; Bd. 12)
NE: GT

ISBN 3-484-31912-7 ISSN 0722-298-X

© Max Niemeyer Verlag Tübingen 1986
Alle Rechte vorbehalten. Ohne Genehmigung des Verlages ist es nicht gestattet, dieses Buch oder Teile daraus photomechanisch zu vervielfältigen.
Printed in Germany. Druck: Weihert-Druck GmbH, Darmstadt.

Meiner Frau Irmhild
und unseren Kindern
Sylva und Jirka

Vorwort

Die Arbeit entstand in mehreren Phasen im Zusammenhang mit meiner Tätigkeit am Forschungsprojekt CONDOR der Firma SIEMENS. Das Projekt CONDOR (COmmunikation in Natürlicher Sprache mit Dialog-Orientierten Retrievalsystemen) wurde zwischen 1973 und 1981 vom BMFT als Teilvorhaben 5.0.0.6 gefördert.

Im Rahmen des Projekts ist das Verfahren zur automatischen Klassifikation der Deskriptoren und Texte STEINADLER sowie die Retrieval- und Relevanzfeedback-Komponente (Präzisierung) entstanden. Die Updatingkomponente wurde erst nach dem Abschluß des Projekts entwickelt. Die vorliegende Arbeit wurde in den Jahren 1981-1984 geschrieben und dabei um generelle Überlegungen ergänzt, die mit der Anwendung von automatischen Klassifikationsverfahren in Information Retrieval Systemen zusammenhängen. Sie bietet deshalb einen umfassenden Überblick über die gesamte Problematik der Clusteranalyse in Informationssystemen. Sie wurde Ende 1984 an der Philosophischen Fakultät der Universität des Saarlandes (Fachrichtung Informationswissenschaft) als Dissertation angenommen.

Mein Dank gilt vor allem Herrn Professor Dr. Harald H. Zimmermann (Universität des Saarlandes), der mit seinem Rat und seiner Ermunterung entscheidenden Einfluß auf das Entstehen dieser Arbeit hatte, sowie den Herren Martin F. Wolters und Dr. N. Banerjee, die meine Arbeit (insbesondere die Entstehung des Verfahrens STEINADLER und der Retrievalstrategien) unterstützt und gefördert haben. Für wertvolle Hinweise danke ich auch Herrn Professor Dr. Werner H. Tack (Universität des Saarlandes).

Mein Dank gilt auch den Herren Dr. Heinz Schwärtzel und Professor Dr. Dieter Schütt, die durch ihr Verständnis und entsprechende Bereitstellung der Hilfsmittel in der Schlußphase der schriftlichen Arbeit, ihre Beendigung ermöglicht haben.

Ein besonderer Dank gebührt Frau Martha Stocker, die das Manuskript in seiner jetzigen Form geschrieben hat.

Meinen Freunden Fritz Fleischhacker, Günter Raiss und Wolfgang Schwarz danke ich für ihren Einsatz beim Korrekturlesen.

Schließlich danke ich meinen Eltern für ihre jahrelange Unterstützung.

München, den 1. April 1985 Jiri Panyr

INHALTSVERZEICHNIS

Vorwort v

0. Einleitung 1

0.1 Motivation der Arbeit und der Vorgehensweise 1
0.2 Kurze Zusammenfassung 6

1. Einige Grundaspekte der Informationsverarbeitung 9

1.1 Informationswissenschaft 11
1.2 Abgrenzung des Begriffs: Information Retrieval System (IR-System) 15
1.3 Relevanzproblematik 24
1.4 Informationsgenerierungprozeß 28
1.5 Bezug der Arbeit zur Informationslinguistik 31
1.5.1 Probleme der linguistischen Forschung 31
1.5.2 Quantitative Linguistik 35
1.5.3 Gegenstand der Informationslinguistik 36
1.6 Forschungs- und Labormodelle von Information Retrieval Systemen 37

2. Automatische Klassifikation 44

2.1 Allgemeine Problemstellung 45
2.2 Theoretische Grundlagen 47
2.2.1 Objekte und Merkmale der Klassifikation 47
2.2.2 Objektmengen der automatischen Klassifikation (Cluster) 57
2.2.3 Klassifikationstypen 61
2.3 Methoden der Klassifikation 63
2.3.1 Direkte und rekursive Verfahren 63
2.3.2 Iterative Verfahren 69
2.3.3 Graphentheoretische Verfahren 73
2.3.4 Hierarchische Verfahren 80

2.4 Anwendung der automatischen Klassifikation in IR-Systemen 87
2.4.1 Problemstellung und Motivation 87
2.4.2 Aspekte des Einsatzes von automatischen Klassifikationsverfahren 89
2.4.3 Dokumentenklassifikation (Dokumentenclustering) 90
2.4.4 Termklassifikation 96
2.4.5 Gemischte Verfahren 99
2.5 Notiz zur dynamischen Klassifikation 103
2.6 Evaluierungsproblematik 104

3. Prinzip des STEINADLER-Ansatzes 109

3.1 Motivation und Grundgedanke des Verfahrens 110
3.2 Notiz zur Voranalyse 112
3.3 Thesaurusaufbau 114
3.4 Klassifikationsprozeß 116
3.4.1 Prioritätsklassenbildung 116
3.4.2 Clusteranalyse innerhalb der einzelnen Prioritätsklassen 118
3.4.3 Verknüpfungen zwischen benachbarten Prioritätsklassen 120
3.5 Updatingprozeß 122
3.6 Übersicht über Dateiorganisation 123

4. Thesaurusaufbau 125

4.1 Zum Thesaurusbegriff 126
4.2 Informationsaufbereitung 131
4.2.1 Automatische linguistische Analyse (im IR-System CONDOR) 131
4.2.2 Statistische Voranalyse 134
4.3 Thesaurusorganisation 137
4.3.1 Hash-organisierter Stammthesaurus (Thesaurushauptteil) 137
4.3.2 Stammthesauruseintrag 139

5. Ermittlung der Hierarchieebenen der Klassifikation (Prioritätsklassenbildung) 141

5.1 Problemstellung 142
5.2 Prinzip der Prioritätsklassenbildung 143
5.2.1 Theoretische Grundlagen und Voraussetzungen (bei Objekten mit binären Merkmalen) 143
5.2.2 Präordnung auf Prioritätsklassen 150
5.2.3 Bestimmung der Prioritätsklassen bei Objekten mit beliebigen Merkmalen 151
5.3 Beschreibung des Algorithmus 153
5.3.1 Einige Hinweise zur Prioritätsklassenbildung 153
5.3.2 Aufbereitung des Klassifikationspools 155
5.3.3 Schematischer Algorithmusablauf 157

6. Clusteranalyse innerhalb der einzelnen Hierarchieebenen der Klassifikation 159

6.1 Problemstellung 160
6.2 Prinzip des Verfahrens (INPRIOR) 161
6.2.1 Theoretische Grundlagen und Voraussetzungen 161
6.2.2 Operationale Algorithmusbeschreibung 168
6.2.3 Formale Beschreibung des Verfahrensablaufs 176
6.2.4 Thesauruseinträge nach der Clusteranalyse 182
6.3 Alternative Verfahren 184
6.4 Verfahrenseigenschaften 185

7. Verknüpfungen zwischen benachbarten Hierarchieebenen der Klassifikation 188

7.1 Problemstellung 189
7.1.1 Theoretische Grundlagen und Voraussetzungen (Ausgangssituation) 190
7.1.2 Grundprinzipien und Basis der Vorgehensweise 198

7.2 Beschreibung des Verfahrensansatzes 203
7.2.1 Auflösung zu kleiner Dokumentengruppen 204
7.2.2 Behandlung der nichtzuzuordnenden Knotenteile 211
7.2.3 Schematischer Algorithmusablauf 212
7.3 Thesauruseinträge nach der Vernetzung 216
7.4 Verfahrensmodifikation und alternative Verfahren 218
7.5 Interpretation und Eigenschaften des Verfahrens 219

8. Updatingmöglichkeiten der Klassifikation (dynamische Klassifikation) 222

8.1 Problemstellung 223
8.2 Updatingprinzip beim STEINADLER-Verfahren 225
8.3 Funktionen der Updatingkomponente 228
8.3.1 Ausgangssituation und Funktionsübersicht 228
8.3.2 Klassieren 233
8.3.3 Korrektur der Termcluster im Klassifikationsnetz 238
8.3.4 Sonderbehandlung einiger Einzelterme 240
8.3.5 Löschen von Informationsobjekten 243
8.3.6 Sonstige Modifikationsmöglichkeiten 246
8.4 Veränderung der Thesauruseinträge nach der Updatingkomponente 247
8.5 Aufwandschätzung (der Klassifikation mit STEINADLER) 249

9. Interpretation der Ergebnisse der Klassifikation bei der Grobrecherche 251

9.1 Einführung in die Rechercheproblematik 252
9.1.1 Recherchearten und Suchtechniken 253
9.1.2 Traditionelle Suchlogiken 256
9.1.3 Prinzip der Clustersuche 258
9.2 Spezielle Problemstellung (STEINADLER) 262

9.3 Grobrecherchestrategien 264
9.3.1 Einleitung und Definitionen 265
9.3.2 Suchauftragprojektion in das Klassifikationsnetz 269
9.3.3 Retrievalfunktionen 274
9.3.4 Rankingalgorithmen 275
9.4 Rahmenfunktionen der Grobrecherche 282

10. Präzisierung der Rechercheergebnisse (als spezielles Relevanzfeedback) 283

10.1 Grundlagen zu Relevanzfeedback-Strategien (interaktive Retrievalstrategien) 284
10.2 Problemstellung der Präzisierung als eines speziellen Relevanzfeedback 288
10.3 Präzisierung 290
10.3.1 Grundgedanke und Einteilung der Vorgehensweise 291
10.3.2 Potentielle Erweiterungsmöglichkeiten 294
10.4 Rahmenbedingungen der Präzisierungsfunktion 295

11. Evaluierungsproblematik 298

11.1 Bewertungskriterien und -maße 299
11.1.1 Bewertungsproblematik: kritische Betrachtung 299
11.1.2 Übersicht über die wichtigsten Effektivitätsmaße 303
11.2 Aspekte der Bewertungsproblematik 306
11.3 Evaluierungsmethoden 310
11.3.1 Zum Vorgang der Evaluierung von Retrievalergebnissen 311
11.3.2 Zur Bewertung der Retrievalleistung 314
11.3.3 Bewertung der Klassifikationsanwendungen 319
11.3.4 Zur CONDOR-Evaluierung (bzgl. STEINADLER) 326
11.4 Evaluierungsproblematik: Schlußbemerkung 327

12. Wechselwirkungen der automatischen Klassifikation mit anderen Erschließungsverfahren und ihr Anwendungsgebiet 329

12.1 Abgrenzung der Anwendungsziele der automatischen Klassifikation in Information Retrieval Systemen 330
12.2 Grenzen der Einflüsse sonstiger Erschließungsinstrumente 333
12.2.1 Thesaurusproblematik 333
12.2.2 Automatische linguistische Analyse 334
12.2.3 Formale Dokumentanalyse 335
12.2.4 Verhältnis zwischen automatischer Klassifikation und automatischer Indexierung 336
12.3 Automatische Klassifikation und Datenmodell 337
12.4 Intellektuelle und sonstige Einflüsse und Wechselwirkungen 339
12.5 Zur Anwendung der automatischen Klassifikation in Information Retrieval Systemen 341

Anmerkungen 343

Literatur 368

Anhänge 391

A. Statistische Daten 393
B. Beispiele der Knotenbildung 397
C. Recherchebeispiel 406

Verzeichnis der mathematischen Bezeichnungen, der sonstigen Notation und Abkürzungen 412

KAPITEL 0

EINLEITUNG

0.1 Motivation der Arbeit und der Vorgehensweise

Im Mittelpunkt der Arbeit stehen die Methoden der automatischen Klassifikation in Information Retrieval Systemen (IR-Systemen), speziell ihre Anwendung in Prozessen der Informationserschließung und -wiedergewinnung. Einen besonderen Schwerpunkt bildet dabei das Verfahren zur automatischen Klassifikation der Deskriptoren (Terme) und der Dokumente STEINADLER /1/, das im Rahmen des Forschungsprojekts CONDOR /2/ zwischen 1974 und 1980 konzipiert und entwickelt wurde.

Die Anwendung der automatischen Klassifikation muß in einem breiteren Spektrum der methodologischen Grundlagen untersucht werden, die mit der Verwendung komplexer Verfahren in IR-Systemem zusammenhängen. Daher ist in der vorliegenden Arbeit auch ein Einblick in die Problematik der Informationserschließung und -wiedergewinnung angeboten.

Für die Anwendung der automatischen Klassifikation in IR-Systemen wurden in bisherigen Ansätzen vor allem zwei Gründe angegeben:

- die Verbesserung der Effizienz der Suche /3/ bei der Dokumentenklassifikation;

- die Verbesserung der Effektivität des Retrievals /4/ bei der Termklassifikation.

Die Ansichten über die Erfüllung der o.g. Zielvorstellungen differieren. Während einige Fachleute der Dokumentenklassifikation neben der Effizienz- auch die Effektivitätsverbesserung bescheinigen /5/, sehen andere die Zielsetzung der Termklassifikation

nicht erfüllt /6/ oder zumindest ihre Anwendung umstritten (z.B. für die Erweiterung der Suchfrage um die Termklassen, die Suchbegriffe enthalten). Bezogen auf die gesamte Clusteringsproblematik sagt Mater (1983), daß man "den extralinguistischen Analysemethoden mehr Aufmerksamkeit schenken sollte /in bezug auf die Informationsverarbeitung in natürlicher Sprache/. Hier sei nur die Cluster-Analyse erwähnt (...), besonders wenn sie als selbstlernendes System /z.B. als dynamische Klassifikation mit Updatingmöglichkeiten/ konzipiert ist. Die Meinungen über die Eignung solcher Verfahren gehen noch weit auseinander. Das kann man als Indiz dafür interpretieren, daß die theoretischen wie praktischen Grundlagen noch nicht hinreichend erforscht sind" /7/. Die Bezeichnung "extralinguistisch" ist unpräzise, denn die automatische Klassifikation kann ohne weiteres in die Nähe der linguistischen Forschung gestellt werden, wie dies später noch aufgezeigt wird /8/.

Als ein weiteres Argument für die Anwendung der Termklassifikation in IR-Systemen wurde auch der mit einem Dokumentenzuwachs sich zunehmend stabilisierende Wortschatz einer Dokumentenkollektion angegeben. Obwohl diese Annahme in überwiegender Mehrheit der Anwendungen ihre Gültigkeit behält,verursacht der ständige Zuwachs der Merkmale (d.h. bei der Termklassifikation der Zuwachs der Dokumente) die gleichen Probleme, die auch mit einem Zuwachs der Klassifikationsobjekte verbunden sind /9/.

Die prinzipielle Unmöglichkeit eines vom eigentlichen Benutzer unabhängigen objektiven Relevanzurteils ist eines der Hauptprobleme der "objektiven" Bewertung der Effektivität. Aus diesem Grund sind Ansätze entstanden, die auf verschiedene Arten versuchten, sich den Effektivitätsvorstellungen eines Benutzers zu nähern. So fangen z.B. Jackson (1970) oder Yu (1974a, 1974b, 1975) mit einer benutzerabhängigen Pseudoklassifikation an /10/ und aus ähnlichen Gründen wurden auch die Methoden des Relevanzfeedbacks (spez. für die Klassifikationsanwendung dann die Clusterfeedback-Methoden) entwickelt /11/. Die Diskussion des Relevanzbegriffs hat ihren Niederschlag z.B. in den probabilistischen Theorien bzw. Modellen der IR-Systeme gefunden. Unter

probabilistischen Modellen sind dabei ausschließlich die Ansätze gemeint, die auf der Grundlage des sog. "Probability Ranking Principle" basieren /12/, und nicht die Methoden, die nur auf irgendeine Art mit wahrscheinlichkeitstheoretischen Begriffen operieren /13/.

Die signifikanteste methodologische Ausprägung dieser probabilistischen Theorien sind die sog. Relevanzfeedback-Strategien. Ein anderer Versuch, die Relevanzproblematik zu bewältigen, sind die sog. Fuzzy-Mengen bzw. die auf ihnen basierenden Modelle des Information Retrievals, die diese Problematik in die Unschärfe der natürlich-sprachlichen Ausdrücke verlagert haben und die auch als eine Reaktion auf die zunehmend logisierende Sprachbetrachtung gedeutet werden können /14/. Die Fuzzy-Modelle operieren, ähnlich wie die klassischen Retrievalmodelle, nur mit dem Begriff "Relevanzgrad" und lassen den Schlüssellbegriff der probabilistischen Modelle "Relevanzwahrscheinlichkeit" außer Acht /15/. Die unkritische Übernahme von solchen Ansätzen führt jedoch zu Problemen, die sich in einer unkritischen Majorisierung eines oder anderes Aspekts der Erschließung, der Wiedergewinnung oder auch der Bewertung der Informationen in einem IR-System niederschlagen. U.a. wird aus diesen Gründen in der vorliegenden Arbeit auch die Relevanz- bzw. die Evaluierungsproblematik eingehend behandelt.

Alle Informationserschließungsverfahren, die die automatische Klassifikation nicht anwenden, führen, unabhängig davon, wie man sie tatsächlich bezeichnet, zu einer strengen Suchlogik, d.h. bei der Recherche werden nur Dokumente gefunden bzw. angeboten, die entweder explizit einige der Begriffe der Suchfrage enthalten oder die explizit mit solchen Suchbegriffen indexiert wurden /16/. Ein intellektuell erstellter Thesaurus kann in übersichtlichen thematisch homogenen Bereichen zwar den durch solche Suchlogiken verursachten Informationsverlust mindern, jedoch nicht beseitigen. In thematisch heterogenen Dokumentenbeständen kann schwerlich ein solcher Thesaurus zur Verfügung stehen. Schon aus diesen Gründen wird angenommen, daß eine automatische Klassifikation eine sinnvolle Ergänzung der Indexierungsmethoden

repräsentiert. Durch die Anwendung der automatischen Klassifikation wird anhand der kontextuellen Assoziationen zwischen Dokumenten bzw. Deskriptoren eine Dokumentbeschreibung (d.h. das Ergebnis der automatischen oder intellektuellen Indexierung) implizit um Deskriptoren erweitert, zu deren Findung (und Zuteilung) keine, wie auch mächtige, Intuition beitragen kann.

Der Wunsch, Dokumentbeschreibungen erweitern zu können, muß sich in der Vorgehensweise einer maschinellen Klassifikation widerspiegeln. Die Deskriptoren (Terme) treten in einer Dokumentenkollektion meist in verschiedenen Kontexten auf. Daraus läßt sich die Notwendigkeit von überlappenden Clustern (d.h. Term- und/oder Dokumentengruppierungen) ableiten. Obwohl für die Erweiterung der Dokumentbeschreibungen sich primär die Termklassifikation anbietet, soll auch die (weniger umstrittene) Dokumentenklassifikation im Merkmalsraum durchgeführt werden. Die die Termklassifikation begleitenden Umgruppierungen im Merkmalsraum (d.h. die Bildung der Dokumentencluster) können dann u.a. zur Bildung von homogeneren Termclustern und somit zu einer "schärferen" Abgrenzung der einzelnen (durch Terme und Dokumente beschriebenen) Kontexte beitragen. Die beiden Clustertypen (d.h. Term- und Dokumentencluster) müssen miteinander eindeutig korrespondieren.

Ein automatisches Klassifikationsverfahren, das in einem IR-System angewendet werden soll, muß notwendigerweise Veränderungen und Zuwachs des Dokumentenbestandes wie auch des Wortschatzes in einem IR-System reflektieren können, d.h. das Verfahren muß diese Veränderungsmöglichkeiten schon bei seiner Konzeption berücksichtigen.

Im weiteren wird die Begründung für die algorithmische Gliederung des (dem o.g. Zweck entsprechenden) Verfahrens STEINADLER kurz aufgeführt.

Die spezifischen (d.h. in der Kollektion die seltenen) Terme erhöhen die Genauigkeit der Suche, führen jedoch zu einem Informationsverlust. Die allgemeineren Terme (d.h. in der Kollektion häufigeren) vervollständigen zwar die Suche, führen jedoch zu

einem ballastbehafteten Retrieval. Unabhängig davon können allerdings die allgemeineren Deskriptoren zu einer einfachen Abgrenzung von verschiedenen Fachgebieten führen (wie z.B. Term BEFUND in der Medizin oder Term EDV in der Informatik), während seltene Terme die Bildung zufälliger Cluster begünstigen können, da die entsprechenden Merkmalsvektoren nicht genügend besetzt sind. Dies war auch die Motivation für die Einteilung der Terme vor dem eigentlichen Klassifikationsprozeß in etwaige Wichtigkeitsklassen. Da jedoch die Wichtigkeit eines Terms für die Wiedergewinnung nicht eindeutig zu bestimmen bzw. zu definieren ist (z.B. auch deswegen, weil die Genauigkeit der Suche sich umgekehrt proportional zu ihrer Vollständigkeit verhält /18/), wurde in der Arbeit eine neutrale Bezeichnung "Prioritätsklasse" gewählt. Eine solche Einteilung ermöglicht die Durchführung der Clusteranalyse zwischen statistisch ungefähr gleich bedeutenden Termen und hält auch die Anzahl der Umgruppierungen im Merkmalsraum im Rahmen. Die einzelnen Prioritätsklassen (d.h. die Hierarchieebenen der Klassifikation) sind in bezug auf die in ihnen enthaltenen Terme disjunkt. Ausgehend von der Hypothese, daß die Bildung der Termcluster in den hinsichtlich der Genauigkeit und der Vollständigkeit ausgewogenen "mittleren" Prioritätsklassen am exaktesten wird, kann in einem der Clusteranalyse nachfolgenden Schritt eine Anpassung bzw. Korrektur der zufälligen Clusterbildung in den Randklassen stattfinden.

Durch die Verknüpfung der benachbarten Prioritätsklassen entsteht dann eine hierarchische Klassifikation, d.h. eine für Änderung und Ergänzung der Klassifikation (beim Zuwachs der Dokumenten und Terme) vorteilhafte Baumstruktur.

Durch diese Vorgehensweise bei der Clusterbildung sind die entstandenen Cluster im Prinzip "übersichtlicher" als die, welche durch irgendeine klassische automatische Klassifikationsmethode entstehen. Daher kann in der entsprechenden Suchstrategie eine Erweiterung vorgenommen werden, die einem Anwender die Möglichkeit bietet, über die für seinen Bedarf nützliche oder nutzlose Termkombinationen interaktiv zu entscheiden. Diese interaktive Suchstrategie (als Präzisierung bezeichnet) kann den Relevanzfeedback-Verfahren zugeordnet werden.

Die o.g. Einteilung des Vorgangs des Klassifikationsprozesses erwies sich in Retrievalexperimenten und nach einem umfassenen Studium der Theorie und Praxis der IR-Systeme als sinnvoll und dem Verwendungszweck adäquat. Der Verfasser ist sich jedoch im klaren, daß die Verwendung der automatischen Klassifikation nicht jeder Anwendung in IR-Systemen angemessen ist. Es muß von Fall zu Fall entschieden werden, wann und wie die Verfahren der automatischen Klassifikation im Prozeß der Informationserschließung und -wiedergewinnung eingesetzt werden können. Daher muß in der vorliegenden Arbeit auch ein genereller Überblick über die gesamte Problematik der automatischen Klassifikation, einschließlich der damit verwandten Fragestellungen der IR-Systeme (wie Evaluierung, Suchstrategie etc.), vermittelt werden.

0.2 Kurze Zusammenfassung

Die Arbeit beschäftigt sich mit der Anwendung der Verfahren der automatischen Klassifikation bei der Informationserschließung und der Interpretation ihrer Ergebnisse bei der Wiedergewinnung in IR-Systemen. Besonders wird das vom Verfasser entwickelte Verfahren zur automatischen Klassifikation der Deskriptoren und Dokumente STEINADLER und die dazu kreierte Suchstrategie behandelt. Die Arbeit soll einen umfassenden Einblick in die gesamte Problematik der automatischen Klassifikation in IR-Systemen vermitteln.

Im folgenden wird eine kurze Übersicht über den Inhalt der Arbeit angeboten. Detailliertere Zusammenfassungen sind am Anfang jedes Kapitels aufgestellt.

Im Kapitel 1 wird zunächst der fachliche Rahmen der Arbeit abgesteckt. Insbesondere werden die Aufgaben der Informationswissenschaft (auch in Beziehung zur "künstlichen" Intelligenz) und der Informationslinguistik (Computerlinguistik) angesprochen, sowie die Terminologie der IR-Systeme erläutert. Besondere Abschnitte werden der Relevanzproblematik und den Informationsgenerierungsprozessen gewidmet. Zum Schluß des Kapitels wird ein Überblick über die Forschungsprojekte und Labormodelle der IR-Systeme vermittelt.

Das Kapitel 2 bietet ein umfassenden Einblick in die Theorie und Methoden der automatischen Klassifikation. Diese Übersicht ist durch die Anwendung von solchen Methoden in einem IR-System motiviert. Die Methoden, die speziell für IR-Systeme entwickelt wurden, sind zum Schluß des Kapitels aufgeführt. Dabei werden auch die Evaluierungsmöglichkeiten der Klassifikationsergebnisse angesprochen.

Das Kapitel 3 beschäftigt sich mit der Motivation und mit dem Grundgedanken und bringt auch eine schematische Beschreibung des Ablaufs des Verfahrens STEINADLER.

Der Thesaurusbegriff wird im Kapitel 4 allgemein eingeführt. Anschließend wird der in dieser Arbeit verwendete Thesaurusbegriff abgeleitet und die Organisation bzw. der Aufbau des Thesaurus erklärt, die dann die Klassifikation mit STEINADLER verwendet.

Die Kapitel 5, 6 und 7 beschreiben die einzelnen Phasen der Klassifikation mit STEINADLER, d.h. die Bildung der Prioritätsklassen, die eigentliche Clusteranalyse innerhalb der einzelnen Prioritätsklassen und die Verknüpfung und gegenseitige Anpassung der Cluster benachbarter Prioritätsklassen (d.h. der Hierarchieebenen der Klassifikation), begleitet jeweils durch die entsprechenden Aufwandschätzungen.

Mit den Updatingmöglichkeiten der Ergebnisse der automatischen Klassifikation generell beschäftigt sich das Kapitel 8. Der Lösungsansatz für die Korrektur, Ergänzung bzw. Erweiterung und das Löschen von Objekten in der durch STEINADLER generierten Klassifikationsstruktur wird detailliert beschrieben.

Einleitend zu der im System CONDOR verwendeten Suchstrategie auf der Basis der STEINADLER-Ergebnisse wird im Kapitel 9 ein Überblick über die traditionellen Recherchearten und Suchtechniken sowie über die Technik der Clustersuche angeboten. Die Rankingalgorithmen und die Diskussion der weiteren Recherche schließen das Kapitel.

Die interaktive Komponente der Recherche auf der Basis von STEINADLER-Ergebnissen (d.h. die Präzisierung) wird im Kapitel 10 beschrieben. Da diese Komponente den Relevanzfeedback-Strategien zugeordnet werden kann, werden auch die Motivation und das Grundprinzip dieser Strategien (einschließlich des sog. Clusterfeedbacks) erklärt.

Eine kritische Betrachtung der Bewertungsproblematik und eine Übersicht über die wichtigsten Bewertungsmaße einschließlich ihrer Interpretation werden im Kapitel 11 angeboten. Speziell werden noch die Bewertungsansätze bei der Anwendung der Clusteranalyse behandelt.

Die Arbeit schließt (im Kapitel 12) mit einer generellen Betrachtung der sinnvollen und nichtsinnvollen Wechselwirkungen der automatischen Klassifikation mit anderen möglichen Erschließungsmethoden oder -instrumenten. Dabei werden auch die Anwendungsmöglichkeiten der automatischen Klassifikation besprochen. Vorerst wird geklärt, was unter der "Verbesserung der Effektivität durch die automatische Klassifikation" zu verstehen ist, d.h. es wird das wünschenswerte Effektivitätskriterium formuliert.

Die einzelnen Kapiteltexte werden durch eine Reihe von Anmerkungen (nach den einzelnen Kapiteln gegliedert) ergänzt.

In den Anhängen werden noch einige statistische Daten und Beispiele des Retrievalablaufs sowie der erzeugten Term- und Dokumentencluster aufgeführt.

Am Schluß der Arbeit sind die im Text verwendeten mathematischen und auch sonstigen Symbole aufgelistet.

KAPITEL 1

EINIGE GRUNDASPEKTE DER INFORMATIONSVERARBEITUNG

Die Grundgedanken und Methoden bzw. Vorgehensweisen, die in der Arbeit beschrieben werden, können zu informationswissenschaftlicher Forschung zugeordnet werden. Aus diesem Grund werden anfangs (im Abschnitt 1.1) die Aufgaben und Ziele der Informationswissenschaft kurz aufgelistet. Ausgehend aus der Unabhängigkeit der drei wissenschaftlichen Disziplinen, die sich in irgendeiner Form mit Sprach- bzw. Informationsverarbeitung beschäftigen, d.h. der Informationswissenschaft, Computerlinguistik und "künstlichen Intelligenz", wird ihre Beziehung zueinander angesprochen.

Da sich die Arbeit primär mit der Anwendung von automatischen Klassifikationsverfahren in Information Retrieval Systemen (IR-Systemen) beschäftigt, wird im Abschnitt 1.2 eine Übersicht über die Terminologie der IR-Systeme bzw. Definitionen der wesentlichen dort verwenden Begriffe angeboten und eine Abgrenzung des Begriffs IR-System durchgeführt.

Der Begriff der Relevanz gehört zu den Schlüsselbegriffen der Theorie der IR-Systeme, zu der diese Arbeit einen Beitrag leisten will. In Abschnitt 1.3 wird daher eine Einführung in die Relevanzproblematik vorgenommen.

In Abschnitt 1.4 wird ein Informationsprozeß definiert, der die Problematik von IR-Systemen widerspiegelt, d.h. der die Wiedergewinnung und auch Informationserschließung als einen approximativen Prozeß darstellt und und die mögliche Wirkung des Benutzers in Betracht zieht.

Abschnitt 1.5 stellt einen kurzen Exkurs in die Computerlinguistik dar. Der Verfasser ist sich im klaren, daß dieser Einblick etwas subjektiv dargeboten wird. Es war auch nicht der Zweck des Abschnitts einen umfassenden Überblick zu liefern, sondern lediglich die Stellung der quantitativen und heuristischen Methoden in der Linguistik zu umschreiben. Da die automatische Klassifikation der Texte (Dokumente) und der Terme (Deskriptoren) in die Nähe der linguistischen Verfahren gestellt wird, war jedoch ein kurzer Exkurs hinsichtlich einer solchen Zuordnung nötig.

Im Abschnitt 1.6 wird ein Überblick über Forschungs- und Labormodelle der IR-Systeme vermittelt, die meist auch die automatische Klassifikation als Mittel zur Erschließung von Informationen konzeptionell enthalten.

1.1 Informationswissenschaft

Während im Bereich der informationswissenschaftlichen Forschung meist relativ unbefangen auf das Gebiet der künstlichen Intelligenz verwiesen wird, fehlt umgekehrt in Arbeiten über künstliche Intelligenz und ihre Prinzipien meist jeglicher Hinweis auf die Informationswissenschaft. Beide Gebiete erheben auch einen gewissen Anspruch auf eine integrative Funktion in bezug auf die zahlreichen wissenschaftlichen Disziplinen, die zu den spezifischen Lösungen im beanspruchten Wirkungskreis beitragen. Nilsson (1971) schreibt: "My view is that artificial intelligence (...) an engineering discipline since its primary goal is to build things (...) Instead of central theory, there are many theoretical subjects that are relevant (...) in artificial intelligence. Some of these include mathematical logic, computational linguistic, theory of computation, information structures, control theory, statistical classification theory, graph theory, and theory of heuristic search" /1/. Demgegen sagen Steinmüller/ Tittlbach (1980): "Es ist anzustreben, daß die Informationswissenschaft zugleich als Metatheorie über den gesamten Informationsbereich unter Einschluß der informationellen Aspekte der genannten Disziplinen entsteht", wie z.B. "Informatik, Dokumentations- und Bibliothekswissenschaft, Klassifikationstheorie, Informationstheorie, Nachrichtentheorie, Planungs- und Entscheidungstheorie, Kommunikationswissenschaft, Sprachwissenschaften, Linguistik, Wissenschaftstheorie, formale (Logik und Mathematik) Methoden der Information" /2/. Walker (1981) versucht eine Differenzierung zu berücksichtigen, indem er über 3 selbständige Forschungsgebiete spricht, setzt jedoch ihre Wechselwirkungen und Überschneidungen voraus:

- Informationswissenschaft;
- Computerlinguistik;
- künstliche Intelligenz ("artificial intelligence" - AI).

Während die Informationswissenschaft sich auf "formalising the processes of knowledge formulation, organization, codification, retrieval, dissemination, and aquisition" konzentriert /3/ (d.h.

auch auf Wissenstransfer), soll die Computerlinguistik ("computational linguistics") auf die folgenden zwei Arten relevant sein: "First, it can provide some practical techniques for interacting with and controlling the operations of computer through natural language (...) second, it addresses the general issue of communication in natural language; recent research has made it clear how complex the processes associated with human understanding really are and how more there is that need know" /4/. "Research in artificial intelligence can be characterized by the kind of activity or area of behavior studied or by basic concepts and techniques that reflect underlying mechanism. In the first case, it is appropriate to refer to vision and image analysis, language and speech understanding, robotics, knowledge based systems, automatic programming and program synthesis, distributed data management, and game playing. (...) Considered in relation to concepts and techniques, AI is concerned with issue of representation and modeling sense reasoning, knowledge acquisition and use, heuristic search procedures, and system control structure" /5/.

Diese Gliederung steht in keinem Widerspruch zu der Auflistung der zugrundeliegenden Prinzipien der AI von Nilsson (1971), ist aber relativ kontrastiv zu Winstons (1977) Versuch, ein "around" System aufzubauen. Nach Walker (1981) kann dieser Kontrast so umgeschrieben werden: "Research on concepts and techniques is often motivated by the desire to formalize abstract principles. However, as noted above, these principles are almost always developed and tested in the context of a system implementation" /6/.

Die (allgemeine) Informationswissenschaft kann nach Zimmermann (1984) als die "wissenschaftliche Behandlung des Transfers von Wissen" betrachtet werden, sei es nun 'künstlichem' (AI, computerintern) oder menschlichem (psychologisch motiviert). Unter dem 'Wissen' wird dabei keine philosophische Kategorie verstanden, sondern "es schließt auch 'Meinen' und 'Glauben' ebenso ein wie theoretisches, praktisches (Handlungs-)Wissen oder in materiellen Gegenständen ('Technologien', 'Werkzeugen') gleichsam kondensiertes Wissen" /7/. Somit wird die Information auch als

Prozeß bzw. als Ergebnis des Prozesses der Vermittlung von Wissen verstanden (vgl. auch Fugmann (1981) bzw. Abschnitt 1.4).

"Wissenstransfer setzt einen Produzenten, einen Rezipienten und ein Vermittlungssystem (Informationssystem) voraus. Das Vermittlungssystem kann auch als Kommunikationssystem betrachtet werden, da die Interaktion zwischen Produzent und Rezipient in der Regel wechselseitig erfolgt" (vgl. Zimmermann (1984)).

So aufgefaßt, kann die Wissenschaft vom Transfer von Wissen in der Fachkommunikation, in Büro und in der Verwaltung (z.B. Fach-Informationswissenschaft) auch als "spezielle" Informationswissenschaft bezeichnet werden. Darunter werden auch Themen dieser Arbeit fallen.

Nach Zimmermann (1984) wird dann "insbesondere nützlich, vor diesem Hintergrund auch einige etablierte Wissenschaften zu beleuchten: So könnte die Linguistik (oder Sprachwissenschaft) unter den vorgestellten Aspekten durchaus auch als eine spezielle Informationswissenschaft aufgefaßt werden. Betrachtet man nämlich die Sprache als ein Kommunikationssystem, das dazu dient, sprachlich kodiertes Wissen zu vermitteln, so sind im Grunde alle entsprechenden Rahmenbedingungen der (allgemeinen) Informationswissenschaft erfüllt" /8/. Mit einigen Abstrichen gilt dies auch z.B. für die Erziehungswissenschaft, Journalistik oder Genetik.

"Auch aus einem anderen Blickwinkel erscheint ein derartig allgemeiner Ansatz zur Definition von Informationswissenschaft nützlich: verbindet er doch verschiedene Disziplinen methodisch und inhaltlich", d.h. als eine Metatheorie (vgl. auch Steinmüller/Tittlbach (1980)), "auch wenn sich in der Spezialisierung vielfältige Unterschiede ergeben" /9/.

Diese Arbeit beschäftigt sich vorrangig mit einer Problematik der IR-Systeme. Auch wenn die sog. AI-Systeme (d.h. Frage-Antwort- oder auch Expertensysteme) als eine evolutionäre Weiterentwicklung der IR-Systeme betrachtet werden können, werden sie diese nicht ablösen können. Beide werden vielmehr entweder ne-

beneinander oder miteinander gleichzeitig und als gegenseitige Ergänzung existieren. Addis (1982) schreibt "Expert Systems An Evolution in Information Retrieval" und Walker (1981) sagt: "The IR-Systems of information science and (...) expert systems of Artificial Intelligence can be viewed as constituting two ends of a continuum of facilities relevant for knowledge sythesis and interpretation. (...) both represent static states, the content of information retrieval systems providing the raw materials from which people derive information relevant for their needs; the expert systems embodying digested knowledge consensually validated as relevant for some area of inguiry" /10/.

Kuhlen (1984) resumiert (ausgehend von einer Autonomie der Informationswissenschaft und der Künstlichen Intelligenz (KI)): "Keineswegs ist die KI eine umfassende Meta-Disziplin für die Informationswissenschaft. So leisteten im Zusammenhang von Information-Retrieval-Problemen KI-Methoden nur einen (...) Beitrag /von mehreren/. (...) Auf der anderen Seite erzwingen die Notwendigkeiten des Informationsmarktes (internationaler Konkurenzdruck) mehr ingenieurmäßige /?/ Entwicklungsarbeit, um z.B. die konzeptionell schon etwa 15 Jahre alten Information-Retrieval-Systeme den heutigen Hard- und Software-Möglichkeiten anzupassen. Der augenblickliche Stand der KI wird den Anforderungen der Informationspraxis von Materialumfang wie von der Komplexität her möglicherweise nicht gerecht. (...) Viele Probleme, für die KI-Methoden relevant sind, sind sehr schwierig und nur längerfristig lösbar, z.Z. noch im Stadium der Grundlagenforschung oder eng begrenzten Experimente (z.B. Mini-Welten). (...) Die Informationspraxis kann nicht auf die Ergebnisse der Grundlagenforschung in der KI warten. (...) Es ist eine gefährliche Fehleinschätzung, wenn in industriellen KI-Projekten davon ausgegangen wird, daß sämtlichen formalen, konzeptuellen und softwaretechnischen KI-Grundlagen bereits erarbeitet wurden und für ehrgeizige Anwendungen nur noch ingenieurmäßig umgesetzt werden müssen. (...) Durch eine verstärkte Zusammenarbeit von KI und Informationswissenschaft, auch und gerade im Bereich der grundlagenorientierten Forschung, ist die Chance gegeben, neuartige Informationssysteme zu konzipieren und zu entwickeln, welche den technischen Poten-

tialen gerecht werden, welche durch die geplante, intensive Förderung der Informationstechnik für die nächsten Jahre zu erwarten sind und welche den direkten Zugang von gelegentlichen und naiven Benutzern zu Informationssystemen ermöglichen sollten. Der zu erwartende Bedarf an Systemen, die eine hohe Informativität bei einer einfachen, benutzergerechten Kommunikationsschnittstelle gewährleisten, wird nur durch eine fruchtbare Zusammenarbeit beider Disziplinen gedeckt werden können" /11/.

1.2 Abgrenzung des Begriffs: Information Retrieval System (IR-System)

Im Mittelpunkt der Arbeit steht ein besonderes Verfahren zur automatischen Klassifikation von Dokumenten und Deskriptoren, das zur Anwendung in einem IR-System entwickelt wurde. Ehe nun das Verfahren beschrieben und in Spektrum der IR-Verfahren eingeordnet wird, sollen einige wesentliche Begriffe der IR-Systeme aufgeführt (bzw. erläutert) werden.

Unter einem Informationssystem versteht Neveling/Wersig (1975) "ein Kommunikationssystem, das durch den Austausch von Daten Informationen bewirken soll" /12/. Nach Gebhardt (1981) wird unter einem System "eine Menge von Objekten, die untereinander in Beziehungen mit wohldefinierten Eigenschaften stehen" verstanden. Wenn diese Beziehungen vornehmlich in der Aufnahme, Verarbeitung und Weitergabe von Informationen bestehen, nennt man das System ein Informationssystem" /13/. Diese Definitionen schließen in sich im Grunde jedes informationsverarbeitende oder verwaltende System mit ein, d.h. auch ein beliebiges natürliches System (wie z.B. das menschliche Denken) oder ein künstliches, wie z.B. die Bürokratie /14/. Eine wesentliche Einschränkung von IR-Systemen als spezielle Informationssysteme formuliert noch Lancaster (1968). Er spezifiziert IR-Systeme wie folgt /15/:
"An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. It merely informs him on the existence (or nonexistence) and whereabouts of documents relating to his request."
Heute werden solche spezielle IR-Systeme als Referenz-Informa-

tionssysteme bezeichnet. Dabei ist es notwendig zu betonen, daß die Definition von Lancaster heute nicht mehr so "eng" gefaßt werden kann. Die Einschränkung bezieht sich bei Lancaster v.a. auf das Retrieval in Literatur-Informationsbanken (sog. bibliographischen Datenbanken, vergleichbar mit den "gedruckten" Bibliographien und Referateblättern). Die neuere Wissenschaft bezieht auch Fakten-Informationssysteme mit ein: so werden auch Systeme, die Operationen über die ermittelten Fakten ausführen (z.B. Zeitreihen-Rechnungen zu statistischen Daten), zu den IR-Systemen gerechnet. Auch durch die Möglichkeit der automatischen Klassifikation und dadurch ermöglichter Präzisierung der Suchfrage auf der Basis des Vorzeigens der automatisch erstellten Ordnungsstruktur des Datenbestandes, bekommt der Benutzer wohl, wenn auch häufig nicht sehr präzise, einen groben Überblick über die Inhalte der Dateneinheiten. Bei Relevanzfeedback-Verfahren kommt es ebenfalls schon zu einer Art des Wissensaustausches zwischen dem Benutzer und einem System, allerdings waren diese Ansätze 1968 mehr oder minder unbekannt. Lancaster (1968) wollte offenbar lediglich das "klassische" IR-System (z.B. Literatur-Informationssystem) gegenüber dem Frage-Antwort-System oder dem Fakten-Retrieval abgrenzen.

Unter "Retrieval" wird allgemein "die Methode verstanden, in einem bestimmten Datenbestand Suchvorgänge durchzuführen" /16/. Datenbestand (bezogen auf ein System) ist dann "die Menge der gespeicherten Dateneinheiten, auf die innerhalb des Systems zugegriffen werden kann" /17/. Ein Retrievalsystem wird demnach beschreibbar als "die Gesamtheit der methodologischen Grundlagen, technischen Verfahren und technischen Einrichtungen, die das Retrieval ermöglichen" /18/. Im weiteren wird unter einem IR-System eine solche "Gesamtheit" verstanden, wobei neben die Informationswiedergewinnung auch die Informationserschließung einbezogen wird. Erst eine konzeptionell (und praktisch) gelöste Informationserschließung (Indexierung im weiteren Sinne) bildet nämlich die notwendige Voraussetzung für die Wiedergewinnung (Retrieval). Bei dem Faktenretrieval ("Datenretrieval") werden "als Antwort auf eine Suchfrage Daten selektiert, die geeignet sind, die Suchfrage direkt zu beantworten" /19/. Die Grenzen zwischen einem IR-

System und einem Faktenretrieval-System werden heute - wie erwähnt - nicht mehr so eng gezogen, wie dies in den 60-er (und auch früheren 70-er) Jahren der Fall war. Komplexere Erschließungsmethoden ermöglichen auch eine komplexere Information bzw. Antwort für den Benutzer bzw. Suchfragesteller. Ein Faktenretrievalsystem (oder ein Frage-Antwort-System) kann somit u.U. als eine evolutionäre Weiterentwicklung des IR-System betrachtet werden (vgl. auch Addis (1982), Walker (1981), van Rijsbergen (1979, 1983), Panyr/Lehmann (1983)).

Die Informationsquellen werden in der IR-Praxis (und Theorie, spez. in der IuD-Praxis /20/) meist als Dokumente bezeichnet. Auch hier existieren z.T. erhebliche Unterschiede in der Auffassung des Dokument-Begriffs. Wersig/Mayer-Uhlenried (1970) beziehen ihre Definition lediglich auf die Dokumentationssysteme. Sie betrachten Dokumente dabei als Träger der Dokumentationsdaten, d.h. "ein Dokument ist die materielle Einheit eines Trägers dokumentarischer Daten", d.h. der Daten, die in der Lage sind, die Rezipienten (Benutzer) eines Dokumentationssystems in bezug auf ihre Anfragen zu "informieren" /21/. Der Begriff "Dokument" wird in dieser Arbeit in weitgehend ähnlichem Kontext benutzt bzw. verstanden. Da sich aber die automatische Klassifikation nicht nur auf solche dokumentarische Daten zu beschränken braucht, wird in der Arbeit auch eine allgemeinere Definition des Dokument-Begriffs verwendet. Der Verwendungszweck bleibt aber dem bei Neveling/Wersig (1975) ähnlich. Demnach ist ein Dokument eine Informationsquelle, die durch eine Menge materiell fixierter und geordneter "Daten oder Informationen" gebildet ist, die eine beliebige (physikalische) Form und charakteristische Merkmale haben können /22/. Es ist offensichtlich, daß darunter auch die Dokumente im Sinne von Neveling/Wersig (1975) einschließlich ihrer akustischen Form (z.B. Tonbandaufzeichnung) und zusätzlich beliebige klassifizierbare potentielle Informationsquellen, wie z.B. DV-Programme, Bilder oder Geräusche, anfallen.

Anstelle des Terminus "Dokument" wird in der Arbeit synonym (wenn auch vergröbend) auch "Informationsobjekt", "Informationsquelle" oder "dokumentarische Bezugseinheit" benutzt.

Im weiteren wird nicht mehr zwischen Dokument- (d.h. selektierte Einheit ist das Dokument selbst) und Nachweisretrieval (d.h. selektierte Einheit sind Dokumentbeschreibungen) unterschieden /23/. "Eine Dokumentbeschreibung ist die Darstellung formaler und/oder inhaltlicher Charakteristika eines Dokuments in einer für Information und Dokumentation verwendbaren Form" /24/. Unter einer formalen Dokumentbeschreibung wird dann "die Beschreibung formaler Charakteristika eines Dokuments" /25/ verstanden. Auf diese "formale Charakteristika" wird (zunächst) nicht näher eingegangen. "Eine inhaltliche Dokumentbeschreibung ist die Beschreibung inhaltlicher Charakteristika eines Dokuments" /26/. Die Bezeichnung "Dokumentbeschreibung" wird im Sinne der (system-)internen Dokumentrepräsentation nach dem Prozeß der inhaltlichen Dokumenterschließung benutzt. Diese Repräsentation kann dann ein Dokument selbst oder das Ergebnis seiner Indexierung werden, evtl. einschließlich der Informationen, die aufgrund der automatischen Klassifikation gewonnen werden, oder auch bibliographische Angaben. Die inhaltliche Dokumenterschliessung wird als eine der funktionalen Komponenten eines IR-Systems betrachtet.

Die inhaltliche Dokumenterschließung wird auch als Indexierung bezeichnet. Laut DIN 31623 wird unter der Indexierung "die Gesamtheit der Methoden und Verfahren sowie deren Anwendungen, die zur Zuordnung von Deskriptoren oder Notationen zu Dokumenten führen mit dem Ziel der inhaltlichen Erschließung sowie gezielten Wiederauffindung" /27/ verstanden.

Eine der Indexierungsmethoden wird dann die automatische Indexierung, "nach der zu einem Dokument Deskriptoren oder Notationen von einer Datenverarbeitungsanlage ermittelt werden" /28/. Synonym zum "Deskriptor" wird in weiterem auch die Bezeichnung "Term" verwendet. Unter einem Term wird dabei ein Deskriptor oder eine Deskriptorgruppe (z.B. Cluster) einschließlich einiger Inhaltskennzeichnungen (wie z.B. die Verknüpfungen zwischen Deskriptoren oder Clustern) verstanden. Diese Kennzeichnungen sind nicht notwendigerweise natürlich-sprachlich, sondern formal-sprachlich (u.a. auch Notation irgendwelches vorgegebenen Klassifikationssystems). Im Angloamerikanischen entspricht es etwa dem Ausdruck "index term".

In DIN 31623 wird hauptsächlich die sog. gleichordnende Indexierung behandelt (d.h. ohne Syntax). Unter der gleichordnenden Indexierung versteht man die "Indexierungsmethode, bei der die Deskriptoren oder Notationen unabhängig von ihrem hierarchischen Niveau und ihren dokumentspezifischen Zusammenhängen gleichrangig nebeneinander gestellt und dem betreffenden Dokument zugeordnet werden" /29/.

Dagegen wird unter sog. syntaktischer Indexierung eine Indexierungsmethode verstanden, bei der neben der Angabe der Deskriptoren deren Gewichtung, Rolle und/oder unterschiedlich enge Verknüpfungen im betreffenden Dokument in besonderer Weise kenntlich gemacht werden. Laut Jansen (1978) ist syntaktische Indexierung "eine Methode, die auf dem semantischen Fundament aufbaut, das durch die gleichordnende Indexierung gelegt wird. Zusätzlich zu den semantischen Einheiten, repräsentiert durch Deskriptoren, treten syntaktische Angaben auf, etwa in Form von Verknüpfungs-, Rollen- oder Gewichtungsindikatoren" /30/.

Bei automatischer Indexierung werden die Deskriptoren überwiegend nach dem sog. Extraktionsprinzip gewonnen, d.h. dem Dokument selbst entnommen. Dies wird besonders dann (gilt auch für intellektuelle Indexierung) sinnvoll, wenn es sich um Gebiete handelt, "in denen die Bedeutung der in Fachsprache verwendeten Termini weitgehend unklar ist und von verschiedenen Autoren und Fragestellern höchst unterschiedlich interpretiert wird (...) Ähnlich liegen die Verhältnisse bei ausgeprägten terminologischen Studien, bei denen es sich ebenfalls um das Auffinden bestimmter Wörter oder Wortgruppen in Publikationen handelt, und bei denen erst aus dem Kontext die Bedeutung ermittelt werden soll" oder kann (Fugmann (1976), /31/). Dies wird wahrscheinlich aber bei jedem heterogenen Dokumentenbestand der Fall, da graphematisch gleiche Wörter verschiedene Bedeutungen haben können.

Prinzipiell ist beim Extraktionsprinzip nicht von Bedeutung, ob die Deskriptoren den Dokumenten mit oder ohne Hilfe eines vorgegebenen Thesaurus zugeteilt werden.

Dem Extraktionsprinzip entgegen steht laut DIN 31623 die sog. Additionsmethode, "nach der einem Dokument Deskriptoren zugeteilt werden, die im Dokument nicht verbal enthalten sein müssen" /32/. Man kann auch beide Prinzipien kombinieren.

Obwohl ein Vergleich zwischen den Ergebnissen der automatischen und manuellen Indexierung den Schluß zuläßt, daß die Ergebnisse gleich gut sein können (vgl. Sparck Jones/Kay (1976a), Salton (1971) und (1975), Gebhardt (1981)), sind bisher bei keiner Institution Anwendungen mit umfangreichen Dokumentenbeständen bekannt, die automatisch indexiert wurden (vgl. Gebhardt (1981)). Als einer der Gründe wird z.B. der hohe Erfassungsaufwand der meist nicht maschinell lesbaren Texte (z.B. mit Hilfe eines Klarschriftlesers - vgl. Supper (1978), Hermann (1973)) genannt. Das automatische Indexieren wird zwar mancherorts für begrenzte Zwecke (z.B. bei Titeln) praktiziert, aber von Dokumentaristen (speziell Indexierern) mit Mißtrauen betrachtet (vgl. Sparck Jones/Kay (1976a)). Für die automatische Indexierung spricht u.a. die auf jeden Fall bessere Konsistenz der Indexierungsergebnisse. Es ist allgemein bekannt, daß manuelle Indexierungsresultate verschiedener Indexierer auch unterschiedlich ausfallen /33/.

Häufiger wird schon die sog. maschinenunterstützte Indexierung angewandt. Anstelle des Begriffs "maschinenunterstützte" wird verschiedentlich der Ausdruck "semiautomatische Indexierung" verwendet. Darunter wird das Indexierungsprinzip verstanden, das auf der Basis einer automatischen Indexierung arbeitet, aber offen für einen beliebigen intellektuellen Einfluß bleibt. Darunter fällt u.a. auch das Darmstädter Projekt für automatische Indexierung AIR (vgl. Lustig (1979), Knorz (1981, 1983)), da dort die Existenz einer repräsentativen intellektuell indexierten Dokumentenmenge vorausgesetzt wird, oder das Saarbrücker Indexierungsprojekt CTX, das mit intellektuell erstellten Lexika bzw. Thesauri arbeitet /34/. Die semiautomatische Indexierung wird als eine Erweiterung der automatischen Indexierung betrachtet.

Die automatische Indexierung arbeitet meist auf der Basis der Extraktionsmethode. Der Effekt der Additionsmethode beim auto-

matischen Indexieren kann z.T. durch die automatische Klassifikation erreicht werden (s.u.). Sie kann auch eine Rolle bei der selektiven Zuteilung der Deskriptoren zu Dokumenten einnehmen.

Die Indexierung wird - wie in DIN 31623 vermerkt - zum Zweck der "inhaltlichen Erschließung sowie gezielten Wiederauffindung" durchgeführt. Der Wiederauffindungsvorgang wird allgemein als Recherche bezeichnet. Unter der Recherche verstehen Neveling/ Wersig (1975) "die Methode, aus verschiedenen Datenbeständen diejenigen Dateneinheiten (Dokumente oder ihre Teile) herauszusuchen, die mit einem gesuchten Sachverhalt und/oder gesuchten Eigenschaften übereinstimmen" /35/. Je nach der Tiefe wird noch eine weitere Teilung vorgenommen:

- als Grobrecherche (GR) wird die Methode bezeichnet, die aus Datenbeständen die Dateneinheiten ermittelt, die mit einem Suchauftrag "grob" übereinstimmen;

- als Feinrecherche (FR) wird dann die Methode bezeichnet, die eine bei der Grobrecherche selektierte Dokumentenmenge in einer weiteren Analyse auf genauere Übereinstimmung mit der Fragestellung überprüft.

Diese Strategie - sie ist weitgehend analysestrategisch, d.h. effizienzbezogen - erhält bei der Anwendung der automatischen Klassifikation in einem IR-System besondere Bedeutung. Beide Recherchearten werden häufig nicht unterschieden, d.h. entweder wird nur ein Recherchetyp in einem System möglich sein, oder sowohl die Grobrecherche als auch die Feinrecherche werden als ein Suchvorgang realisiert.

Zu den Methoden des Recherchieren gehört auch die sog. Freitextrecherche. Hierbei wird in Dokumenten oder in invertierten Dateien nach einem bestimmten Wortlaut oder einer bestimmten Begriffskombination im Text (meist satzbezogen - vgl. z.B. Wieland (1979)) gesucht. Unter einer invertierten Datei wird vereinfacht eine Deskriptorenliste mit Verweisen auf die Dokumente verstanden, die den Deskriptor selbst (d.h. Listendeskriptor) beinhalten

(evtl. mit der Angabe der Position des Deskriptors im Dokument, d.h. Absatz oder Satz und des Ranges im Satz). Wenn die aufgenommen Dokumente als

$$D_i = \{T_j \mid T_j \in D_i\}$$

bezeichnet werden, können dann die Einträge in der invertierten Datei als

$$T_j = \{D_i \mid T_j \in D_i\}$$

bezeichnet werden. Nach Neveling/Wersig (1975) ist die Benennung "invertierte Datei" allerdings "fehlinterpretierbar, denn sie bezieht sich auf die Umordnung der Dateneinheiten gegenüber der Dateneingabe. Baut man aus den eingegebenen Zitierungen eines Dokuments eine invertierte Datei auf, die die Angaben enthält, welches Dokument wo zitiert worden ist, so kann man diese Datei nur noch bei Kenntnis des Aufbauvorgangs als invertiert erkennen" /36/. Neveling/Wersig benutzen stattdessen die Bezeichnung "merkmalbezogene Datei". Zu einem Deskriptor sind dann meist die Kennzeichnungen (z.B. Dokumentnummer) aller Dokumente (oder Dokumentationseinheiten) zugeordnet, die mit diesem Deskriptor indexiert worden sind /37/.

Unter einem IR-System wird im weiteren ein Quintupel

$$IR = (I, \mathcal{D}, Q, R, E)$$

verstanden. $\mathcal{D}$ repräsentiert dabei die Menge aller möglichen Dokumentbeschreibungen (speziell die Dokumente selbst), d.h. den aktuellen Dokumentenbestand. I steht für die Informationserschließungsfunktion, speziell also für einen Indexierungsvorgang evtl. mit einer automatischen linguistischen Analyse und einer automatischen Klassifikation. Q ist die Menge aller möglichen Suchfragen (Suchaufträge, Bedarfsmeldungen oder Anwendungen), die an das IR-System gestellt werden können; R bezeichnet schließlich die Retrievalfunktion. Je nach dem Erschließungsvorgang I kann sich hinter der Menge $\mathcal{D}$ auch eine implizierte

Dokumentbeschreibung verbergen, wie z.B. eine Term- und/oder Dokumentenklassifikation. Die Funktion R kann als eine Folge aller möglichen Retrievalvorgänge dargestellt werden (z.B. GR, FR, Relevanzfeedback, Präzisierung etc.). Sie vergleicht u.a. die Dokumentbeschreibungen (z.B. die Deskriptorenmengen der einzelnen Dokumente) mit der Suchfrage (oder Anwenderbedarfsmeldung) $f \in Q$ (z.B. die Menge der aktuellen Suchargumente). Die Menge E stellt schließlich die Menge der Rechercheergebnisse (d.h. Systemvorschläge) dar. Es gilt für $e \in E$ und $f \in Q$, daß

$$e = R(f, \mathfrak{D}), \text{ wobei } R(f, \mathfrak{D}) \subset \mathfrak{D}$$
$$\text{oder } R(f, \mathfrak{D}) = \emptyset,$$

speziell dann $R: \mathfrak{D} \times Q \dashrightarrow E$.

Dabei ist zunächst ohne Interesse, ob die Ergebnismengen $e \in E$ dem Anwender des IR-System gleichgeordnet oder in einer von der Relevanz der Dokumente bezüglich des Suchauftrags $f \in Q$ abhängigen Reihenfolge angeboten werden.

Offensichtlich kann es sich bei den Funktionen I und R um sehr komplexe Funktionen handeln. Die Struktur der Menge $\mathfrak{D}$ ist zunächst bei einer solchen formalen Beschreibung ohne Interesse. Die Suchaufträge aus Q können entweder natürlich-sprachige Ausdrücke, Ausdrücke in einer formalen Sprache oder eine Kombination von beiden sein. Zu der Beziehung der beiden Funktionen I und R sagt z.B. Fugmann (1976) (vgl. auch Robertson et al. (1982) im ähnlichen Kontext): "Es hängt auch sehr von der gestellten Aufgabe ab, ob der mit dem Wiederauffinden von Informationen verbundene Aufwand bevorzugt beim Indexieren oder verstärkt beim Retrieval geleistet werden sollte, ob man weiterhin auf ein stärker ballastbehaftetes, jedoch informationsverlustarmes Retrieval Wert legt, ob die Verhältnisse umgekehrt gelagert sind, oder ob man in einem Indexierungs- und Retrievalsystem über Wahlmöglichkeiten verfügen sollte."/38/

Als Parameter für die Güte des Retrievals sind meist die Maße "Recall" und "Precision" verwendet. Unter einem Recall (d.h.

Vollständigkeit der Suche) versteht man das Verhältnis der Anzahl der wiedergewonnenen relevanten Dokumente im Systemvorschlag zu der Anzahl aller bzgl. der gestellten Suchfrage relevanten Dokumente im Dokumentenbestand. Die Precision (d.h. die Genauigkeit der Suche) wird dann durch den Anteil der relevanten wiedergewonnenen Dokumente an allen wiedergewonnenen Dokumenten im Systemvorschlag ausgedrückt. Diese Werte sagen jedoch für sich allein wenig. Sie werden deswegen häufig zu sog. Recall-Precision-Graphen kombiniert (vgl. näher Kapitel 11).

1.3 Relevanzproblematik

Maron/Kuhns (1960) bezeichnen in ihrem einflußreichen Aufsatz die Relevanz als den Schlüsselbegriff der Theorie der IR-Systeme /39/. Ähnlich schreibt z.B. auch Bookstein (1979): "The concept of relevance is fundamental to information retrieval" /40/. Nach Saracevic (1975) ist Relevanz ein Maß der Übereinstimmung zwischen Dokument und Suchfrage aus der Sicht eines "Schiedsrichters" /41/. Bei Saracevic ist die Relevanz auch aus anderen Perspektiven definiert, so z.B. in der Kommunikationstheorie als ein Maß der Effektivität der Verbindung zwischen Sender und Empfänger ("source" und "destination" /42/) in einem Kommunikationsprozeß. Ein Retrievalprozeß kann freilich auch als ein Kommunikationsprozeß aufgefaßt werden /43/. Es ist auch offensichtlich, daß ein Kommunikationssprozeß dann am effektivsten ablaufen wird, wenn die Relevanzsysteme (und freilich auch die Interpretationsschemata) von Sender und Empfänger (oder z.B. auch von Indexierern und IR-Anwendern) identisch oder zumindest ähnlich werden (vgl. Schütz (1971), (1981)).

Die sog. probabilistischen Modelle von IR-Systemen betrachten die Relevanz als eine Zufallsgröße. Für einen solchen stochastischen Ansatz gibt es vor allem zwei Quellen /44/:

- die Relevanzbewertung eines Dokuments bezüglich einer Suchfrage unterliegt Zufallsschwankungen;

- der Zusammenhang zwischen Relevanzbewertung eines Dokuments und (meßbaren) Eigenschaften des Dokuments (z.B. Zahl der darin enthaltenen Suchargumente, Ähnlichkeit zur Suchfrage) unterliegt Zufallsschwankungen, die nicht nur von den Zufallsschwankungen der Relevanzbewertung herrühren.

In den probabilistischen Modellen von IR-Systemen wird dann versucht, diese "Relevanzschwankungen" durch die "Relevanzwahrscheinlichkeit" zu beschreiben.

Die Theorie der probabilistischen Retrievalmodelle basiert im Prinzip auf den Arbeiten von Robertson (1977a, 1977b, 1977c) und Cooper (1973). Einige Annahmen wurden allerdings schon früher, z.B. bei Maron/Kuhns (1960) oder speziell bei Robertson/Sparck Jones (1976), formuliert /45/. Von Cooper wurde dann zuerst das sog. probabilistische Rankingprinzip ("probabilistic ranking principle") ausgesprochen /46/: "If a reference retrieval system s response to each request is a ranking of the document in the collections in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then overall effectivness of the system to its users will be the best that is obtainable on the basis of that data." Also demnach soll das IR-System, um eine optimale Effektivität (für den Benutzer) zu gewährleisten, in der Lage sein, die angebotenen (d.h. wiedergewonnenen) Dokumente nach der Wahrscheinlichkeit des positiven Relevanzurteils seitens des Benutzers oder ihrer Nützlichkeit für das Benutzerproblem oder Informationsbedürfnis zu ordnen (vgl. auch Robertson (1977c)).

Robertson et al. (1982, 1983) schlagen deswegen ein Relevanzkonzept vor, das die Wirkung des Benutzers ("Patron") berücksichtigt. Demnach wird die Relevanz als eine Relation zwischen einem Dokument und einem Benutzer in bezug auf sein Informationsbedürfnis verstanden. Nach dieser Interpretation wird dann die Menge der relevanten Dokumente einfach der Menge derjenigen Dokumente gleichgesetzt, die das Informationsbedürfnis des Kunden-Anwenders ("Patron") zufriedenstellen ("Pertinenz").

Eine Relevanzbeurteilung der wiedergewonnenen Dokumente hinsichtlich des Suchauftrages kann prinzipiell von zwei Seiten durchgeführt werden:

- vom IR-System, das zu entscheiden hat, ob ein Dokument nachgewiesen werden soll (System-Relevanz);

- vom Schiedsrichter (bzw. Benutzer selbst), der die inhaltliche Übereinstimmung des formulierten Problems (Suchauftrag) mit wiedergewonnenen Dokumenten beurteilt.

Nach Sager (1975) können noch, falls der Schiedsrichter mit dem Benutzer identisch ist, zwei Beurteilungsarten unterschieden werden /47/:

- Relevanzgrad-Beurteilung, d.h. der Benutzer bewertet "möglichst objektiv" die Relevanz des Dokuments für sein Problem;

- Pertinenz-Beurteilung, d.h. der Benutzer bewertet "subjektiv" den Informationswert des Dokuments bezüglich seines Problems (seines Informationsbedürfnisses).

Diese Beurteilungen müssen nicht übereinstimmen, so z.B. kann ein höchst relevantes Dokument für den Benutzer keinen Informationswert haben, da er es nicht versteht, oder zu allgemein ist, oder er es schon kennt. Im zweiten Falle (Pertinenz-Beurteilung) dient die vorher erwähnte Veränderung des Wissens infolge eines Wissenstransfers (vom IR-System zum Benutzer) als das entscheidende Kriterium bzw. als ein Indikator der "Nützlichkeit" eines wiedergewonnenen Dokuments für das Informationsbedürfnis des Benutzers.

Aus der funktionalen Gesichtspunkt der sog. Prager linguistischen Schule ("Cercle Linguistique de Praga" /48/) kann auch eine Parallele zwischen der Beziehung Relevanz - Pertinenz (aus der Position des Benutzers) und der Beziehung Thema - Rhema (in englischen Topic - Comment) hergestellt werden. Vereinfacht kann das Thema als die schon gegebene, bekannte Information im Unterschied

zum Rhema als der neuen, noch unbekannten Information (im einem Satz) beschrieben werden. Der Begriff des "Funktionalen" im Sinne der Prager Schule ist jedoch sehr vielschichtig. "Aufgabe", "Zweck" oder "Absicht" sind wohl die zutreffendsten Interpretationen des Begriffs der Funktion im Sinne dieser Schule /49/.

Girke (1972) oder Danes (1973) erweitern die ursprünglichen satzspezifischen Definition von Thema-Rhema zu einer textspezifischen Definition. Girke (1972) spricht auch über Textthema bzw. über Textrhema und meint, daß "die Thema-Rhema-Gliederung noch stärker unter dem Aspekt Textrelevanz untersucht werden muß" /50/.

Die Pertinenz-Beurteilung ist auch die Ausgangsbasis für die probabilistischen Modelle des Informationsretrievals, speziell dann für die sog. nutztheoretischen Modelle der Indexierung ("Utility-theoretic-Indexing" /51/).

Das System bestimmt den System-Relevanzgrad nach einer formalen Übereinstimmung zwischen dem formulierten Problem (Suchauftrag) und den expliziten (d.h. Dokumentbeschreibung) oder impliziten (z.B. Ergebnis der automatischen Klassifikation) Identifikationsmerkmalen des Dokuments. Diese systemseitige Relevanzbeurteilung (Relevanzgradbestimmung) des gefundenen Dokuments bezüglich der Suchfrage geschieht mit Hilfe von sog. Rankingalgorithmen.

Die Relevanzbeurteilung seitens eines Benutzers und seine Auswertung seitens des Systems ist dann der Gegenstand der sog. Relevanzfeedback-Verfahren.

Im Prinzip kann freilich die Relevanzgrad-Beurteilung seitens des Benutzers auch als Basis für die Evaluierung eines Systems dienen. Die Pertinenz-Beurteilung dagegen hat überwiegend nur im Zusammenhang mit einem Relevanzfeedback-Verfahren einen Sinn. Sie bildet dann auch die Basis von probabilistischen Suchstrategien (vgl. auch Robertson (1977a)).

Nach Robertson/Belkin (1978) sind für die Notwendigkeit eines Ranking von Dokumenten zwei unterschiedliche Gründe angegeben:

- Relevanz selbst ist eine "multi-valued" oder kontinuierliche (fortlaufende) Variable;

- Retrieval ist ein prinzipiell approximativer Prozeß.

Diese zwei Gründe führen dann zu zwei verschiedenen Rankingprinzipien, wobei

- das eine auf einem Relevanzgrad,

- das andere auf dem Begriff der Relevanzwahrscheinlichkeit

basiert.

Während der Relevanzgrad (d.h. die Relevanzvariable) im Prinzip die Diskrepanz zwischen Dokumentbeschreibung und Informationsbedürfnis auszudrücken versucht, wird nach dem zweiten Prinzip (d.h. der Relevanzwahrscheinlichkeit) hauptsächlich die Unstimmigkeit zwischen der Suchfrage und dem eigentlichen Benutzerbedürfnis ausgedrückt /52/. Die beiden Prinzipien sollten eigentlich kombiniert werden. Als die geeignete Vorgehensweise, die auch der Vorstellung über den approximativen Charakter der Retrievalprozesse entspricht, können dabei primär die Relevanzfeedback-Algorithmen dienen.

1.4 Informationsgenerierungsprozeß

Nach Neveling/Wersig (1975) ist "ein Informationsprozeß ein Kommunikationsprozeß, der zur Verringerung von Ungewißheit führt (Information bewirkt)". Die "Ungewißheit" soll dann (nach Neveling/Wersig (1975)) "etwa verstanden werden als eine Situation, in der das Individuum eine Differenz zwischen internen Motivationsvorgaben und bestehender Situation nicht aufgrund bestehenden Wissen lösen kann, sondern weitere Daten einholen muß. Dies gilt nicht nur für aktuell eingetretene Situationen, sondern auch für zukünftige Situationen, auf die sich das Individuum ständig vorbereiten muß" /53/.

Nach Fugmann (1981) findet der Mensch häufig in Vielerlei keine Ungewißheit, und trotzdem können einschlägige Nachrichten (wie z.B. Tod des Freundes, eine Naturkatastrophe) für ihn von großem Informationswert sein. In einer fortgeschrittenen Definition wird demzufolge Ungewißheit durch Nichtwissen ersetzt, und Information ist dann die Vermittlung von Wissen oder das Ergebnis eines Wissensaustausches. Auf diese Problematik wurde schon im Abschnitt 1.1 näher eingegangen.

Moser (1977) spricht in diesem Zusammenhang über die Unbestimmtheitslage: "Information läßt sich als Ergebnis eines Prozesses (der Informationsgenerierung) verstehen, der durch Integration (Verarbeitung) räumlich und/oder zeitlich transmittierten Daten zur Veränderung einer gegebenen Unbestimmtheitslage führt" /54/. "Eine für die Informationsgenerierung wichtige Konsequenz, wie sie aus dem vorgetragenen Ansatz unmittelbar abgeleitet werden kann, ist schließlich, daß Information nur soweit entstehen kann, wie die Unbestimmtheit oder auch Unwissen vorherrscht und - evtl. in einem iterativen Verfahren - expliziert wurde. Art und Ausmaß der Informationsgenerierung hängen allerdings, wie erwähnt, zugleich von der Datenlage, d.h. von Art und Umfang der erreichbaren Daten ab" /55/.

Bei der automatischen Informationsgenerierung (bzw. Informationserschließung) wird im Endeffekt zu einem bestimmten Zweck, in einer bestimmten Zeit und evtl. auch für einen bestimmten Benutzer meist aus einer Wissensquelle eine Information generiert. Ein solcher Vorgang kann dann einem informationellen Prozeß gleichgesetzt werden.

Die Relevanzvorstellungen (in speziellen von Relevanzsystemen oder Relevanzkonzepten) des Benutzers und die des Systems (z.B. der automatischen Klassifikation) stimmen meistens nicht voll überein. Zum Zwecke der Anpassung solcher Relevanzsysteme aneinander bzw. zum Zwecke der Erläuterung des Interpretationsschema des Systems muß daher in einem informationellen Prozeß auch ein Kommunikationsprozeß integriert werden. Dieser Kommunikationsprozeß (z.B. der Präzisierungsdialog bzw. das Relevanzfeedback-

Verfahren - vgl. Kapitel 10) dient also entweder der Klärung des Interpretationsschemas (z.b. des Darstellungsformalismus auf sprachlicher Metaebene) des Systems oder der Annäherung des Relevanzkonzepts des Systems an das des Benutzers (und umgekehrt).

Die evtl. Bewertung der Ergebnisse der automatischen Informationsgenerierung (-erschließung) kann so mit der Möglichkeit der Annäherung beider Relevanzkonzepte verknüpft werden. So z.B. kann die Frage nach einer "guten" oder "falschen" Klassifikation erst nach einem solchen Prozeß beantwortet werden, z.B. durch die Evaluierung der Rechercheergebnisse. Der Prozeß der Informationsgenerierung in IR-Systemen (d.h. der Indexierungsprozeß oder auch die automatische Klassifikation) kann also nicht unabhängig von dem Recherchevorgang gesehen werden, er bildet mit dem Rechercheverfahren in einem IR-System eine Einheit.

In Abb. 1.1 wird versucht, einen derartigen Informationsgenerierungsprozeß schematisch darzustellen. Dabei wird angenommen, daß das Interpretationsschema des Systems (Is) und das des Benutzers (Im) übereinstimmen (d.h. Is = Im). Demzufolge reduziert sich der Zweck des Kommunikationsprozesses K (z.B. des Retrievalprozesses) auf die Annäherung des Relevanzkonzeptes des Systems (Rs) an das des Benutzers (Rm). Falls Is $\neq$ Im ist, dient der Kommunikationsprozeß K u.a. auch der Erhöhung der Systemstransparenz und somit auch der Anpassung des Interpretationsschemas Im an Is. Aus der Abb. 1.1 und o.g. Erläuterungen ist ersichtlich, daß hier das Relevanzkonzept des Systems als anpassungsfähig an das Relevanzkonzept des Benutzers betrachtet wird, wogegen das Interpretationsschema des Systems als etwas vorgegebenes vorliegt. Demzufolge muß nach dieser Konzeption der Benutzer sein Interpretationsschema anpassen.

Auch das Relevanzkonzept hängt eng mit der Relevanzdefinition zusammen. Sie wurde im vorherigen Abschnitt eingehend behandelt. Darunter fallen systemseits die Fragen des Rankings der Ergebnisse, der Gewichtung der Indexterme und Suchargumente /35/ und auch die evtl. Modifikation der Klassifikationsergebnisse (z.B. durch ein Relevanzfeedback-Verfahren). Das Interpretationsschema

kann u.U. bei einzelnen Anwendungen dem System vorgegeben werden, aber in der Regel geschieht dies nicht. Zu den Aspekten des Interpretationsschemas gehört z.B. die formale Art der Präsentation der Rechercheergebnisse, die Form der gewonnenen Klassifikationsstruktur oder evtl. auch die algorithmische Vorgehensweise während der Erschließung (G).

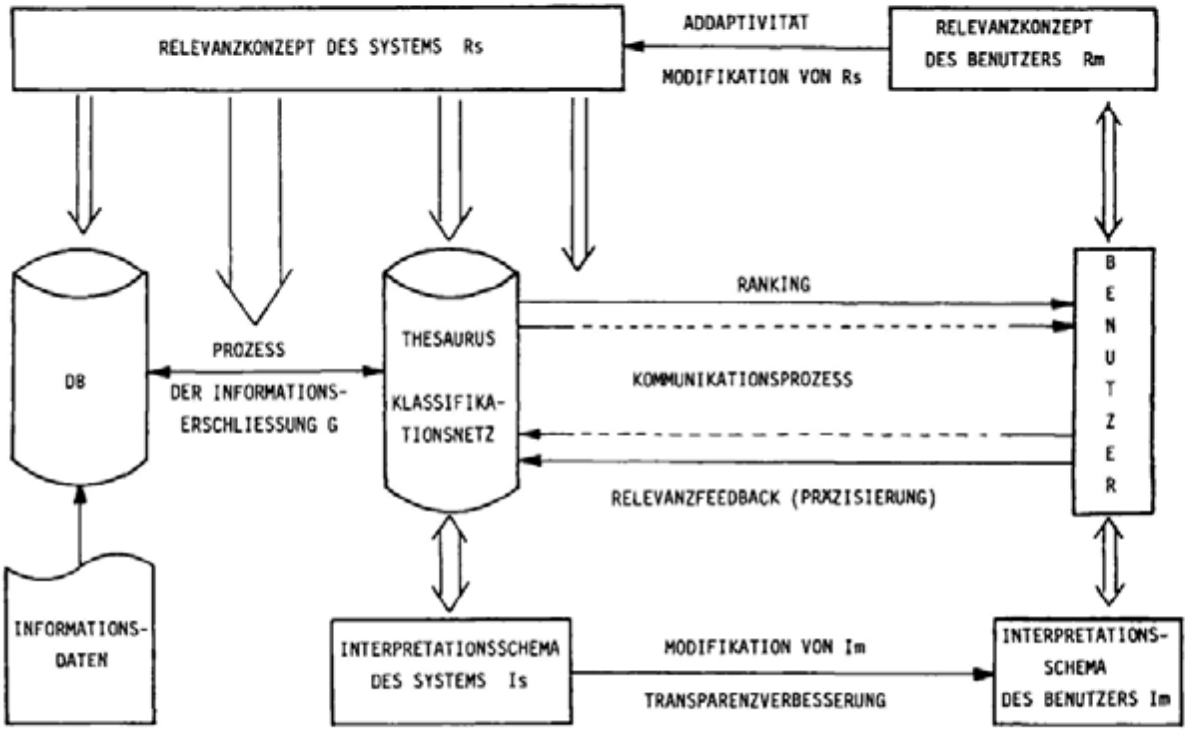

Abbildung 1.1

1.5 Bezug der Arbeit zur Informationslinguistik

1.5.1 Probleme der linguistischen Forschung

Die formalen Methoden der Sprachbeschreibung (wie sie z.B. die generative Linguistik und die mit ihr verwandte algebraische oder mathematische Linguistik entwickelt haben /57/) erreichten relativ schnell - trotz ihrer strikten Formalismen - eine hohe Akzeptanz in der theoretischen Linguistik. Man spricht in diesem Zusammenhang (wenn auch ungenau) von "qualitativer Linguistik" /58/.