Checkliste für Retrievalsysteme

Qualitätskriterien von Suchmaschinen

Suchmaschinen im World Wide Web wird nachgesagt, dass sie - insbesondere im Vergleich zur Retrievalsoftware kommerzieller Online-Archive - suboptimale Methoden und Werkzeuge einsetzen. Elaborierte befehlsorientierte Retrievalsysteme sind vom Laien gar nicht und vom Professional nur dann zu bedienen, wenn man stets damit arbeitet. Die Suchsysteme einiger "Independents", also isolierter Informationsproduzenten im Internet, zeichnen sich durch einen Minimalismus aus, der an den Befehlsumfang anfangs der 70er Jahre erinnert. Retrievalsoftware in Intranets, wenn sie denn überhaupt benutzt wird, setzt fast ausnahmslos auf automatische Methoden von Indexierung und Retrieval und ignoriert dabei nahezu vollständig dokumentarisches Know how. Suchmaschinen bzw. Retrievalsysteme - wir wollen beide Bezeichnungen synonym verwenden - bereiten demnach, egal wo sie vorkommen, Schwierigkeiten. An ihrer Qualität wird gezweifelt. Aber was heißt überhaupt: Qualität von Suchmaschinen? Was zeichnet ein gutes Retrievalsystem aus? Und was fehlt einem schlechten?

Wir wollen eine Liste von Kriterien entwickeln, die für gutes Suchen (und Finden!) wesentlich sind. Es geht also ausschließlich um Quantität und Qualität der Suchoptionen, nicht um weitere Leistungsindikatoren wie Geschwindigkeit oder ergonomische Benutzerschnittstellen. Stillschweigend vorausgesetzt wird jedoch der Abschied von ausschließlich befehlsorientierten Systemen, d.h. wir unterstellen Bildschirmgestaltungen, die die Befehle intuitiv einleuchtend darstellen. Unsere Checkliste enthält nur solche Optionen, die entweder (bei irgendwelchen Systemen) schon im Einsatz sind (und wiederholt damit zum Teil Altbekanntes) oder deren technische Realisierungsmöglichkeit bereits in experimentellen Umgebungen aufgezeigt worden ist. Insofern ist die Liste eine Minimalforderung an Retrievalsysteme, die durchaus erweiterungsfähig ist. Gegliedert wird der Kriterienkatalog nach (1.) den Basisfunktionen zur Suche singulärer Datensätze, (2.) den informetrischen Funktionen zur Charakterisierung gewisser Nachweismengen sowie (3.) den Kriterien zur Mächtigkeit automatischer Indexierung und natürlichsprachiger Suche.

Die entstehende Sammlung von Qualitätskriterien soll bei der Evaluation konkreter Systeme eingesetzt werden können. Wir bekommen - als Informationswissenschaftler - eine Vergleichsmöglichkeit zwischen unterschiedlichen Werkzeugen, - als Produzent - einen Vorschlagskatalog, was alles noch zu tun ist, und - als Knowledge Manager - eine Entscheidungshilfe, welche Software im Intranet eingesetzt werden kann.

Retrievalsoftware hat mehrere Einsatzmöglichkeiten:

- als Suchmaschine für das gesamte World Wide Web
- als Portalservice für einen Ausschnitt des Internet
- als Suchsystem bei (kommerziellen) Online-Archiven
- als Suchsystem bei singulären Angeboten im Web ("Independents")
- als Retrievalsystem für persönliche Informationssammlungen
- als Retrievalsystem bzw. Portalservice bei unternehmensweiten Intranets.

Die Suche geschieht unabhängig vom - durchaus zufälligen - Format der Dokumente, d.h. es werden alle Arten von Dokumenten gefunden, egal, ob im HTML-, PDF-, ASCII-Format oder was auch immer verwendet wird. "Dokumente" sind alle Typen medialer Darstellungen, also Texte, Bilder, Audio- oder Videosequenzen sowie beliebige Kombinationen daraus. Die Dokumente können sowohl dokumentarisch bearbeitet sein (beinhalten dann u.a. ein Abstract sowie Deskriptoren oder Notationen - der Standardfall in guten Informationssammlungen) als auch "nackte" Dokumente (ohne jeden informationellen Mehrwert beim Input ausgestattet).

Basisfunktionen: Suche nach der "Nadel im Heuhaufen"

Datenbankindex. Für den Fall, dass ein Intranet oder ein Online-Archiv über mehrere Datenbanken verfügt, ist es sinnvoll, mit einer Indexdatenbank zu arbeiten, die dem Nutzer eine Rangfolge von Datenbanken anbietet, mit deren Inhalt er sein Informationsproblem am günstigsten lösen kann. Im World Wide Web wäre eine Variante zu diskutieren, die diese Funktion bei den sog. "Portalservices" erfüllt: Der Index des Portals gibt bei Eingabe eines Sucharguments darüber Auskunft, bei welchem Independent oder auch bei welchem Host (in welcher Datenbank) die meisten Treffer zu erwarten sind.

Wenn man sich über seine Suchargumente zumindest grob im Klaren ist, lässt man im Gesamtindex des Datenbankanbieters oder in der gewählten Datenbankgruppe alle Datensätze durchsuchen. Resultat ist eine Liste der Datenbanken, die man nach der Auftretenshäufigkeit unseres Suchargumente sortieren lassen kann. Anhand der so entstehenden "Hitparade" entscheidet man,

mit welchen konkreten Datenbanken weitergearbeitet werden soll.

Bei der Formulierung der Suchargumente in der Indexdatenbank ist Vorsicht geboten. Wir können hier nicht von genau einem Vokabular ausgehen, z.B. einem Thesaurus, und auch nicht von genau einer natürlichen Sprache, insofern Intranets oder Hosts Datenbanken mit unterschiedlichen Dokumentationsmethoden und in unterschiedlichen natürlichen Sprachen nebeneinander aufliegen haben. Hier genau *einen* Dachthesaurus oder *ein* alles umfassendes Klassifikationssystem zu fordern, liegt zwar nahe, dürfte aber praktisch nur sehr schwer zu erstellen sein.

Datenbankaufruf. Hat sich der Rechercheur für eine oder mehrere verwandte Datenbanken entschieden, so sind diese über einen entsprechenden Button aufzurufen. Wünscht man, in gewissen Teilmengen einer Datenbank zu arbeiten, so kann man dies in einigen Systemen bereits beim Aufruf markieren (etwa für den Aufruf des Segmentes der letzten zwei Jahre einer bestimmten Datenbank).

Feldspezifische Suche - Suche im Basic Index. Eine gezielte Suche läuft normalerweise über Felder. Unterschieden werden durchweg zwei Feldtypen. Alphanumerische Felder erlauben die Suche mit Buchstaben, Zahlen und einigen Sonderzeichen, numerische Felder gestatten sowohl algebraische Operationen als auch (manchmal) Berechnungen. Bei alphanumerischen Feldern sucht man entweder nach einzelnen Wörtern oder nach kompletten Phrasen wie AUTOR=Marx, Karl. Ein Phrasenindex muss notwendigerweise für alle Felder vorliegen, die solche Mehrwortausdrücke beinhalten, müsste man doch sonst - um im Beispiel zu bleiben - AUTOR=Marx UND AUTOR=Karl suchen. Numerische Felder lassen neben der Gleichheit weitere algebraische Suchen zu, also nach "größer als", "kleiner als" bzw. nach Intervallen. Rechenoperationen benötigt man vor allem beim ökonometrischen Retrieval.

Gibt man bei der Suche kein Feldkürzel an, so verläuft die Suche automatisch im sog. "Basic Index". Was jeweils hierunter fällt, sollte im Intranet frei definierbar sein. Es muss dem Nutzer in einer Erläuterung zum Feldinhalt mitgeteilt werden.

Grammatische Varianten. Normalerweise wird nicht zwischen Klein- und Großbuchstaben unterschieden. Für bestimmte Suchen kann es aber sinnvoll sein, Großbuchstaben gezielt anzugeben. Lexis-Nexis bietet drei Varianten an: (1) allcaps (findet z.B. AIDS), (2) nocaps (findet aid), (3) caps (findet Aid). *AltaVista* sucht bei Kleinschreibung nach allen Varianten, bei einem Großbuchstaben jedoch nur solche Formen, die genau diesen Großbuchstaben enthalten.

Sucht man in Volltextdatenbanken, so muss die Suche nach einem Wort sowohl dessen Plural- als auch Singularformen umfassen. Für die englische Sprache bieten einige Systeme eine automatische Pluralbildung für den regelmäßigen Plural an. Möchte man seine Suche auf genau eine Form beschränken, so ist dies explizit anzugeben. So gilt bei *Lexis-Nexis:* plural (job) findet jobs, singular (job) findet job, ansonsten: job findet jobs sowie job. Diese Methode kann bei Eigennamen, die auf "-s" enden, nützlich sein. Das Retrievalsystem erkennt solche Eigennamen nicht und wendet seine Regeln auch - fälschlicherweise - darauf an. Eine Suche nach dem Namen des Gründers von *Apple Computer*, Steven Jobs, ist ein passendes Beispiel. Um nicht Ballast (über das Wort "job") zu bekommen, muss die Suche über die Pluralform gebildet werden: steven w/3 plural (job) - w/3 ist ein Abstandsoperator (s.u.!).

Eine Verallgemeinerung der automatischen Singular-Plural-Erkennung ist die Synonymverknüpfung aller Flexionsformen eines Wortstammes. Egal, welche Flexionsform man eingibt, man bekommt die Treffer zu allen Flexionsformen des gemeinsamen Wortstammes bzw. Lemmas.

Blättern im Wörterbuch. Über den Basic Index und auch über die Indices der Felder werden Wörterbücher geführt und dem Benutzer angeboten. Die Wörterbücher sind sowohl wortinvertiert (Mehrwortausdrücke sind zerlegt: "Marx" unter "M", "Karl" unter "K") als auch phrasen-invertiert ("Marx, Karl" nur unter "M") angelegt. Normalerweise wird die alphabetische Umgebung des Wörterbucheintrags angezeigt. Weitergearbeitet wird durch Markieren der gewünschten Wörter.

Präsentation von Begriffsordnungen. Setzt eine Datenbank eine Dokumentationssprache, also Klassifikation oder Thesaurus, ein, so ist für die entsprechenden Felder das "Blättern" zum Stöbern in Begriffsnetzen zu verfeinern, die entweder verbal oder - besser - graphisch angegeben werden. Zu unterscheiden ist zwischen paradigmatischen und syntagmatischen Beziehungen. Erstere sind die "fest verdrahteten" Relationen in Thesaurus und Klassifikationssystem (Synonymie-, Hierarchie-, Assoziationsrelation). Letztere sind diejenigen Relationen, die sich durch das gemeinsame Auftreten in Indexaten (bei gleichordnendem Indexieren) bzw. in thematischen Ketten (bei syntaktischem Indexieren) bilden. Da bei Begriffen, die mit vielen anderen Begriffen verbunden sind, eine große Menge an syntagmatisch verknüpften Begriffen vorhanden ist, empfiehlt sich eine Sortierung dieser nach der Anzahl des gemeinsamen Auftretens. (Dass hinter jedem Begriff bzw. jedem Begriffspaar die Anzahl der aktuellen Treffer angezeigt wird, dürfte selbstverständlich sein.)

Gewisse Relationsangaben erfordern eine Benutzeraktion. Diese sollten vom System unterstützt werden. Der Verweis von einem Nichtdeskriptor auf eine Kombination (z.B. Bibliotheksstatistik BK Bibliothek, Statistik) muss in die Eingabe Bibliothek UND Statistik (bei syntaktischem Indexieren Bibliothek UND.S Statistik, s.u.!) münden. Natürlich muss - automatisch, aber angezeigt - der Verweis vom Nichtdeskriptor zum Deskriptor ausgeführt werden. Syntagmatisch verknüpfte Begriffspaare erfordern bei der Suche eine UND bzw. UND.S-Verknüpfung.

Browsen in Begriffsordnungen. Ein großer Vorteil systematisch aufgestellter Bibliotheksbestände ist das Stöbern in Nachbarbereichen meines Such-themas. Links und rechts neben dem exakten Treffer stehen Bücher, die dem Suchthema ähnlich sein können. In elektronischen Umgebungen kann ich nicht nach links und rechts schauen, also sollten wir dies simulieren. Voraussetzung ist der Einsatz eines Klassifikationssystems. Bei der Präsentation der Begriffsordnung (im vorigen Absatz) kann man ausschließlich über die Begriffsleitern navigieren, etwa von DK 659.13 nach unten zu DK 659.131, DK 659.132 usw. sowie nach oben zu DK 659.1. Beim hier geforderten Browsen gelangen wir zu den nächsten Nachbarn in der Begriffsreihe (also zu DK 659.12, DK 659.14 usw.) sowie auch über Leitern- und Reihengrenzen hinweg (z.B. zu DK 659.115).

Synonymwörterbuch. Gewisse Systeme bieten zur Umschreibung von Suchargumenten sogenannte "Thesauri" an. Hier ist zu beachten, dass der Begriff "Thesaurus" zwei Bedeutungen hat. Einmal ist die Dokumentationssprache gemeint, zum andern geht es um Synonymwörterbücher. Dieser zweite Aspekt ist hier gemeint. Zwei Varianten zur Erstellung von Thesauri sind möglich. Zum einen werden publizierte Wörterbücher (z.B. das Synonymwörterbuch des "Dudens") ins System eingegeben. Zum andern werden im Rahmen eines statistischen Thesaurus "Synonyme" mittels automatisierter Verfahren kreiert. "Statistisch" ist das Verfahren, insofern Terme, die häufig im gleichen Kontext (etwa innerhalb eines Satzes) vorkommen, als "ähnlich" markiert werden.

Homonymwörterbuch. Verwendet ein Benutzer ein Suchargument, das mehreren Begriffen entspricht, so muss das System nachfragen, welche der Varianten gemeint ist. Nehmen wir an, wir gäben "Pirat" ein. Da das Wör-

terbuch über drei Einträge verfügt, wird eine Liste zur Auswahl angeboten: Pirat (Seeräuber) - Pirat (Seeräuberschiff) - Pirat (Software-Pirat). Durch Anklicken der gewünschten Varianten kann insbesondere in großen Volltextdatenbanken schon bei der Suchformulierung der Ballast minimiert werden. Der Erfolg von Systemen, die diese Option beherrschen (wie z.B. *Excalibur),* hängt wesentlich von den verwendeten Wörterbüchern ab.

Fragmentierung. Suchargumente können entweder komplett oder auch fragmentiert angegeben werden. Bei kontrolliertem Vokabular, etwa eines Thesaurus (im Sinne einer Dokumentationsmethode), empfiehlt sich die komplette Bezeichnung, ansonsten kann es in gewissen Fällen sinnvoll sein, mit sog. "Jokerzeichen" zu arbeiten, die die Fragmentierung (Truncation) herstellen. Eine große Bedeutung hat die Fragmentierung bei der hierarchischen Suche mithilfe eines Klassifikationssystems.

Bei Suchen nach Wörtern in frei formulierten Texten (Titel, Abstract, Volltext) sind wir stets mit den Varianten der grammatischen Flexionsformen konfrontiert, die mit geschickten Fragmentierungen zusammengefasst werden können. Ebenso sind unterschiedliche Schreibweisen (Meier oder Meyer) gemeinsam suchbar. Aber auch bei numerischen Feldern kann Truncation hilfreich sein. Denken wir z.B. an eine Suche nach Unternehmen im Bereich Berlin, die wir anhand der Postleitzahlen identifizieren wollen. Hier müssen wir nach der "1" fragmentieren.

Es gibt Fragmentierungszeichen für eine unbegrenzte Anzahl von Zeichen, d.h. hinter dem Suchargument können null oder beliebig viele Zeichen vorkommen. So findet Unternehm$ die Terme Unternehmen, Unternehmung, Unternehmer, Unternehmensphilosophie usw. Das Analogon zur gerade gezeigten Rechtsfragmentierung ist die Linksfragmentierung, eingesetzt etwa zur Suche nach Unternehmensformen ($unternehmen), und findet Bauunternehmen, Chemieunternehmen, Stahlunternehmen usw. (Technisch bedeutet die Linksfragmentierung kein Problem. Sie kann als einfache Rechtsfragmentierung gelöst werden, angewandt auf rückläufig sortierte Wörter.)

Darüber hinaus sind Fragmentierungen mit einer bestimmten Anzahl von Zeichen nutzbringend zur Rechts-, Links- sowie zur Binnenfragmentierung einsetzbar. Nehmen wir an, der Joker * steht für ein beliebiges Zeichen. Dann findet Me*er die Worte Meier, Meyer, Mejer usw. Die begrenzte Rechtsfragmentierung sei an einem weiteren Beispiel verdeutlicht: bank*** findet (im Englischen) u.a. bank, banker und banking, aber nicht bankruptcy.

Mengentheoretische Operatoren. Bei der Verwendung mehrerer Suchargumente müssen diese durch Operatoren miteinander verbunden werden. Im einfachsten Fall geschieht dies durch die mengentheoretischen Operatoren UND (Schnittmenge), ODER (Vereinigungsmenge) und NICHT (Exklusionsmenge). In Erinnerung an den Mathematiker und Logiker George Boole (1815 - 1864) werden die Operatoren auch "Boolesche Operatoren" genannt.

Mit Klammersetzung kann man gewünschte Bindungen herstellen. Möchte man etwa die Vereinigungsmenge zweier Argumente A und B mit der Menge C schneiden, so ist mit Klammern wie folgt zu formulieren: (A oder B) und C.

Abstandsoperatoren. Insbesondere bei langen Texten ist die Verwendung des mengentheoretischen UND-Operators kritisch, können doch Datensätze gefunden werden, wo die durch UND verknüpften Argumente in völlig unterschiedlichen Kontexten vorkommen. Lösungsmöglichkeiten erschließen sich durch Abstandsoperatoren, die alle das Boolesche UND verschärfen. Datenbanken, die mittels syntaktischem Indexieren durch Kettenbildung arbeiten, können so ihre Syntax abfragbar machen.

Der engste Abstand zwischen zwei Suchargumenten liegt vor, wenn die Bezeichnungen im selben Feld direkt nebeneinander stehen. Der hierzu passende Abstandsoperator hat zwei Ausprägungen, insofern er einmal die Reihenfolge beachtet, das andere Mal nicht. Unter Berücksichtigung der Reihenfolge langt es bei einigen Systemen, die Suchphrase (ggf. durch Anführungszeichen "..." gekennzeichnet) hinzuschreiben, also etwa "Julia Roberts", oder durch einen eigenen Operator, z.B. "adj" (adjacency), zu verknüpfen: Julia adj Roberts. Eine Phrase der Art "Roberts, Julia", z.B. ein Wörterbucheintrag, wird so nicht gefunden. Hier muss man mit einem Operator arbeiten, der von der Reihenfolge der Suchargumente absieht, etwa der Operator (n): Julia (n) Roberts.

Nicht nur der Abstand von genau einem Wort ist nützlich, sondern auch größere Intervalle. Angeboten wird bei einigen Online-Archiven ein allgemeiner Abstandsoperator W/n, wobei n eine ganze Zahl zwischen 1 und - je nach Software - einigen Hundert ist. Der Abstand wird ausschließlich innerhalb eines Feldes ausgezählt. Das Suchargument "Julia Roberts" w/5 "Lyle Lovett" findet bei Lexis-Nexis alle Datensätze, in denen die beiden Namen, egal in welcher Reihenfolge, im gleichen Feld im Abstand von höchstens fünf Wörtern stehen.

Der allgemeine Abstandsoperator ist ein reiner Zählalgorithmus. Es gibt auch grammatische Varianten von Abstandsoperatoren. Bezugspunkte sind der gleiche Satz, der gleiche Absatz, das gleiche Feld, eine besondere Stellung im Satz (etwa am Satzanfang), und dies unterteilt in positive wie negative Fälle, wie es z.B. *TRIP* anbietet.

Bei der Anwendung syntaktischen Indexierens zur Inhaltsabbildung werden die Deskriptoren in mehreren Teilmengen abgelegt. Ein Indexat habe die beiden thematischen Ketten

- Italien - Wein - Export.
- Italien - Kfz - Import.

Das Retrieval erfordert in diesem Fall die Eingabe eines Abstandsoperators und nicht des Booleschen UND. Wenn wir die thematischen Teilmengen als Analogon zu einem Satz sehen, müsste entsprechend mit dem Operator UND.S gearbeitet werden. Die Suche nach dem italienischen Weinexport ergibt sich beim Suchargument Wein UND.S Italien UND.S Export. Eine Suche nach Wein UND.S Italien UND.S Import findet unser Beispielindexat korrekterweise jedoch nicht. Erst durch Abstandsoperatoren wird ein Retrieval im Rahmen der syntaktischen Indexierung überhaupt praktikabel.

Häufigkeitsoperator. In Volltextdatenbanken kann es sinnvoll sein, gewisse Terme nach ihrer Auftretenshäufigkeit zu suchen. Eine solche Suche wird von der - nicht immer zutreffenden - Voraussetzung begründet, dass ein Term um so wichtiger ist je öfter er in einem Text

vorkommt. Anwendbar sind Häufigkeits-operatoren vor allem dann, wenn jede Indexierung fehlt und wenn sehr viele Datensätze zum Suchargument vorliegen.

Hierarchische Suche. Bei Klassifikationssystemen erhalten wir durch geschickte Rechtsfragmentierung eine einfach durchzuführende Form hierarchischen Retrievals. Eine unbegrenzte Rechtsfragmentierung findet die Datensätze zu einem Begriff sowie zu allen Unterbegriffen. Eine begrenzte Fragmentierung findet soviele Ebenen wie Fragmentierungszeichen verwendet werden. Folgende Beispiele mögen dies verdeutlichen:

- DDC=382$ (findet alles zu 382 nebst allen Unterbegriffen)
- DDC=382* (findet alles zu 382 und zu den Unterbegriffen der nächsten Ebene)
- DDC=382** (findet alles zu 382 und zu den Unterbegriffen der nächsten zwei Ebenen).

Nach unserer Verabredung, keine unerklärten Befehle zu verwenden, muss das Retrievalsystem sowohl auf "382" als auch auf "*" bzw. "$" verzichten und statt dessen umgangssprachliche Klassenbenennungen sowie ankreuzbare Alternativen anbieten. Die hierarchische Suche arbeitet mit den Optionen der Präsentation von und des Browsen in Begriffsordnungen zusammen.

Thesaurusrelationen sind nicht wie bei den Klassifikationen über Fragmentierungen abfragbar. Vielmehr benötigen wir hier eigene Operatoren. Bei der Software GRIPS gibt es den hierarchischen Operator "Down", der einen Deskriptor und alle seine Unterbegriffe gemeinsam sucht. Bei der WISO-CD-ROM können die Unter- bzw. Oberbegriffe sowie die verwandten Begriffen jeweils ersten Grades auf Knopfdruck in die Suche übernommen werden. Sinnvoll erscheinen u.a. Abfragebefehle der Art:

- Suche nach einem Deskriptor samt der Unterbegriffe der nächsten Ebene
- Suche nach einem Deskriptor samt der Unterbegriffe der nächsten n Ebenen (n > 1)
- Suche nach einem Deskriptor samt aller Unterbegriffe
- Suche nach einem Deskriptor samt der Oberbegriffe der nächsten Ebene
- Suche nach einem Deskriptor samt aller verwandter Begriffe.

Gewichtetes Retrieval. Nicht jedes Suchargument ist für den Suchenden gleich wichtig. Es kann sein, dass ein bestimmtes Wort äußerst relevant, die anderen Suchargumente aber peripher sind. Hier muss das System Gewichtungswerte als Eingaben zulassen (etwa Term "A" - Faktor 10; "B" - Faktor 3; "C" Faktor 1). Insbesondere bei syntaktischem Indexieren (etwa nach der Textwortmethode N.*Henrichs*) ist es möglich, einen dokumentspezifischen Gewichtungswert für jedes Textwort zu errechnen. Man kann so Dokumente selektieren, die den Suchterm wichtig besprechen, und andere ignorieren, wo der Term eher am Rande thematisiert wird.

Datenbankübergreifende Suche. Sucht man in mehreren Datenbanken gleichzeitig, so ist immer damit zu rechnen, dass man Dubletten erhält. Das Umgehen mit den Dubletten erfordert einige Retrievalbefehle. Zunächst sind die Dubletten zu identifizieren. Die entsprechenden Identifikationsprogramme vergleichen ausgewählte Inhalte gewisser Felder. Insbesondere Schreibvarianten in den unterschiedlichen Datenbanken machen das Verfahren nicht völlig zuverlässig. Wenn beispielsweise eine Datenbank einen Autorennamen mit "Marx, Karl" ansetzt und eine andere mit "Marx, K." und das Identifikationsprogramm beim Autorenfeld zehn Stellen vergleicht, so kann eine entsprechende Dublette nicht erkannt werden.

Der zweite Schritt ist das Löschen der Dubletten. In der Regel werden die Datenbanken in der Reihenfolge ihres Aufrufs abgearbeitet. Haben wir beispielsweise die Datenbanken 15, 16 und 17 (in dieser Reihenfolge) aufgerufen, so bleiben alle Datensätze aus 15 erhalten. In 16 werden die Datensätze entfernt, die als Dublette zu einem Datensatz aus 15 erkannt worden sind, usw. Gibt es Qualitäts- oder Preisunterschiede zwischen den Datenbanken, so muss man dies durch die daran orientierte Reihenfolge beim Aufruf berücksichtigen, also zuerst die "beste" oder billigste Datenbank nennen, dann die zweitbeste usw.

Umformulierung von Suchergebnissen zu Suchargumenten. Manchmal werden Suchfragen nur durchgeführt, um zu einer aussagekräftigen Menge von Suchargumenten zu kommen. Dies kann u.a. dann der Fall sein, wenn wir in einer Synonymdatenbank (etwa zu chemischen Bezeichnungen) alle Varianten zu einem Suchterm eruieren wollen. Nehmen wir an, wir suchten nach "Aspirin". Die Suche in der Datenbank der chemischen Bezeichnungen führt zu mehreren Dutzend Synonymen. Diese interessieren uns jedoch im Einzelnen nicht; wir benötigen sie vielmehr als Suchargumente in einer anderen Datenbank. Der entsprechende Befehl bei *DIALOG* lautet "Map". Hierbei werden die Suchergebnisse eines Feldes auf Suchargumente desselben Feldes abgebildet.

Das "Mapping" ist auch über Feldgrenzen hinaus sinnvoll. Nehmen wir nun an, wir suchten alle die Patente, die Patente eines bestimmten Unternehmens zitieren. Zunächst müssen wir alle Patente finden, die unser Unternehmen als Patentanmelder nennen. An den gefundenen Datensätzen sind im Sinne unserer weiteren Recherche nur die Patentnummern (Feldkürzel pn) interessant. Über einen Ausgabebefehl erhalten wir eine Liste der Form pn=aaa bis pn=zzz. Im nächsten Schritt müssen die Feldkürzel pn durch die Kürzel ct (für cited patents) umbenannt und alle mit dem logischen Oder verbunden werden. Über die anschließende Suche nach den zitierten Patenten löst sich unser Informationsproblem.

Ökonometrische Befehle. Wirtschaftsstatistische Datenbanken berichten in numerischer Form über Objekte und zugehörige Merkmale. Werden Daten auch nach dem Merkmal "Zeit" erhoben, so haben wir Zeitreihen vor uns. Bei Datenbanken mit statistischen Informationen sind zwei Arten von Befehlen zu unterscheiden, erstens die Suchbefehle, die Zeitreihen auffinden, und zweitens Befehle, die bei den gefundenen Zeitreihen ansetzen.

Die Suchbefehle unterscheiden sich nicht prinzipiell von den "normalen" Booleschen Operatoren. Der große Unterschied zu nicht-statistischen Datenbanken liegt in den Befehlen für die Weiterverarbeitung von Zeitreihen. Angeboten werden einfache Berechnungen, etwa Summenbildungen über verschiedene Zeitreihen hinweg (z.B. die deutsche Ausfuhr nach Frankreich *und* Italien, dargestellt als *eine* Zeitreihe) oder die Änderung der Währungsbasis (z.B. statt DM in US-$). Ökonometrische Befehle reichen von Glättungen, Saisonbereinigungen, Abweichungen von der Vorperiode, Korrelationen zwischen Zeitreihen, Regressionsrechnungen, Umrechnungen zwischen Absolut- und Indexwerten bis hin zum Durchspielen makroökonomischer Modelle.

Anzeigen von Suchergebnissen. Nehmen wir an, wir haben nach einigen Suchschritten ein Ergebnis gefunden, von dem wir meinen, dass es "passen" könnte. Es gilt demnach, die Datensätze ganz oder teilweise am Bildschirm anzuzeigen. Der Anzeigebefehl verlangt in der Regel nach drei Argumenten, dem Suchschritt, den gewünschten Datensätzen im Suchschritt und dem Ausgabeformat. Weiß der Recherchierende nicht mehr - nach langem Suchen -, welche Nummer der letzte Suchschritt hat oder wünscht er einen Überblick über seine bisherige Strategie, so gibt es hierfür einen eigenen Anzeigebefehl.

Bei der Ergebnisanzeige werden entweder vorformulierte oder selbstdefinierte Formate benutzt. Vorformuliert sind beispielsweise ein Freiformat (zum ersten Sichten eines Ergebnisses, z.B. eine einfache Titelliste) oder ein Vollformat (mit Informationen aus allen Fel-

dern). Man definiert eigene Formate, indem man diejenigen Felder angibt, deren Informationen man bei der Ausgabe sehen möchte. Wichtig ist ein Befehl zum "Ankreuzen" eines Datensatzes. Bei einer ersten Durchsicht (etwa nur der Titel) oder auch nach der Vollanzeige entscheidet der Nutzer, ob er den Datensatz behalten möchte oder nicht. Es sollte die Möglichkeit eingeplant sein, alle gefundenen Datensätze auf einmal zu markieren.

Unter gewissen Umständen kann es sinnvoll sein, nur wenige Informationen eines umfassenderen Datensatzes auszugeben, diese aber sortiert und in einem Tabellenformat aufbereitet. Wenn wir etwa nach Unternehmen gesucht haben, können wir Firmen nach Umsatzzahlen sortieren und in einem Reportformat unter Angabe von Namen, Umsatz, Stadt o.ä. als Tabelle ausgeben.

Viele Datenbankanbieter liefern ihre Daten ausschließlich im ASCII-Format. Das hat den Vorteil, dass man sie weiterverarbeiten kann. Es hat den Nachteil, dass "gewohnte" Formate wie der Satzspiegel einer Zeitung verloren gehen. Entsprechend bieten einige Online-Archive zusätzlich zur ASCII-Ausgabe ein Quasi-Faksimile an, das den Satzspiegel simuliert.

Bestellen von Volltexten. Suchergebnisse aus bibliographischen Datenbanken sind Hinweise auf Literaturstellen, Suchergebnisse der ASCII-Volltextdatenbanken sind Texte ohne Graphiken. In allen anderen Datenbanken - etwa in Intranets oder im Internet - sollte der Volltext im Originalformat bereits vorliegen.

Muss ein Volltext extern beschafft werden, so gibt es mehrere (elektronische) Wege, den Text zu ordern. Viele Hosts bieten einen entsprechenden Bestellbefehl an. Die Bestellung wird an die gewünschte externe Bibliothek oder auch an die hauseigene Bibliothek weitergeleitet und dort bearbeitet. Zudem kann man kommerzielle "Document Delivery Services" konsultieren. Solche Dienstleister arbeiten mit Bibliotheken und Verlagen zusammen und halten große Mengen wissenschaftlicher Zeitschriften zur jederzeitigen Benutzung bereit. Ideal wäre es, wenn ein Link vom bibliographischen Nachweis direkt zur elektronischen Version

Datenbankindex
Datenbankaufruf
--- genau eine Datenbank
--- Aufruf von Datenbanksegmenten
--- datenbankübergreifender Aufruf
Feldspezifische Suche
--- Suche innerhalb von Feldern (Phrasenindex)
--- Suche im Basic Index
--- algebraische Operationen bei numerischen Feldern
Grammatische Varianten
--- Groß- / Kleinschreibung
--- Singular- / Pluraleinstellung
--- Wortstamm
Blättern im Wörterbuch
Präsentation von Begriffsordnungen
--- verbal
--- graphisch
--- paradigmatische Relationen
---syntagmatische Relationen
--- Vorbereiten von nötigen Aktionen (bei BS, BK und syntagmatischen Relationen)
Browsen in Begriffsordnungen
Synonymwörterbuch
--- Thesaurus (im sprachwissenschaftlichen Sinn)
--- "statistischer" Thesaurus
Homonymwörterbuch
--- Dialog zur Abklärung homonymer Bezeichnungen
Fragmentierung
--- Links-, Mitte-, Rechtsfragmentierung
--- Anzahl der zu ersetzenden Zeichen (genau n Zeichen; beliebig viele Zeichen)
Mengentheoretische Operatoren
--- Boolesche Operatoren
--- Klammersetzung
Abstandsoperatoren
--- direkt benachbart (mit und ohne Beachtung der Reihenfolge)
--- Zähloperator
--- grammatische Operatoren
Häufigkeitsoperator
Hierarchische Suche
--- Suche nach einem Deskriptor samt der Unterbegriffe der nächsten n Ebenen
--- Suche nach einem Deskriptor samt der Oberbegriffe der nächsten n Ebenen
--- Suche nach einem Deskriptor samt aller verwandter Begriffe
Gewichtetes Retrieval
Datenbankübergreifende Suche
--- Dublettenerkennung
--- Dublettenelimination
Umformulierung von Suchergebnissen zu Suchargumenten
--- Mapping im selben Feld
--- Mapping in ein anderes Feld
Ökonometrische Befehle
Anzeige / Ausgabe
--- Titelliste
--- Markieren von Datensätzen
--- Reportausgabe in Tabellenformat
--- Ausgabe des Datensatzes in frei wählbarem Format
--- Faksimile-Simulation
Bestellen von Volltexten
--- Vorhalten in Originalformat
--- Link zur elektronischen Version
--- Link zu Dokumentlieferdiensten
Verwalten von Suchprofilen
--- Push-Service (E-Mail)
--- Portfolio

Tabelle 1: Grundfunktionen

des betreffenden Artikels führen würde (wie etwa in der Kooperation zwischen *Web of Science* und *Link* vom *Springer-Verlag* realisiert).

Verwalten von Suchprofilen. Arbeitet man ein Suchargument genau einmal ab, so spricht man von einer "retrospektiven Recherche". Wenn man sein Suchargument als "Suchprofil" abspeichert, kann man zu unterschiedlichen Zeitpunkten dasselbe Suchargument benutzen. Eine solche selektive Informationsvermittlung ist immer dann sinnvoll, wenn der Informationsbedarf eines Nutzers oder einer Nutzergruppe über eine längere Zeit stabil bleibt.

Denken wir z.B. an ein Forscherteam, das an einem Thema arbeitet. Nach der ersten Recherche ist das Team daran interessiert, stets die neuesten Informationen zu seinem Thema zu erhalten. Gelöst wird dies, indem das Suchprofil immer dann abgefragt wird, wenn die entsprechende Datenbank aktualisiert worden ist. Das Suchargument läuft dann freilich nicht über den Gesamtbestand der Datenbank, sondern nur über den jeweiligen Zuwachs. Nach dem gleichen Verfahren ist ein Pressespiegel (mehr oder minder automatisch) herstellbar.

Ein Retrievalsystem muss in der Lage sein, Suchprofile zu verwalten. Auf Ergebnisse wird als Push-Service (in der Regel via E-Mail) aktiv vom System hingewiesen.

Profildienste zu Daten, die ständig in Bewegung sind (etwa Börsenkurse oder Meldungen von Presseagenturen), werden verwaltet, indem die entsprechenden Angaben realtime aus den operativen Systemen abgezogen werden. In seinem "Portfolio" kann der Nutzer jederzeit die jeweils aktuellen Angaben abrufen. Einige Datenbankanbieter, darunter *Profound*, bieten Portfolios in Kombination von numerischen Informationen (Aktienkursen) und Literaturinformationen (Broker-Reports und Pressemeldungen zu den überwachten Unternehmen) an.

Data Mining durch Informetrie: Ein Heuhaufen als Ganzes

Wir verlassen nunmehr die Suche nach den "Nadeln" im Heuhaufen und wenden uns größeren Einheiten zu. Bislang haben wir Retrievalbefehle besprochen, die möglichst zielgenau Datensätze finden. Nunmehr wenden wir uns gewissen Mengen von Nachweisen zu, die als Ganzes qualifiziert werden sollen. Hierbei kann es sich beispielsweise um alle Patente eines Unternehmens handeln oder um alle bisherigen Schriften eines Wissenschaftlers.

Bei der Suche nach Datensätzen erhalten wir als Suchergebnis einen Eintrag, der - als Dokumentationseinheit - so in das System eingegeben wurde wie wir ihn nun herausbekommen. Bei der informetrischen Suche kreieren wir neue Informationen, also solche, die nie explizit in die Datenbank eingegeben wurden und die sich erst durch den informetrischen Suchvorgang selbst zeigen. Informetrische Recherchen setzen Methoden ein, die experimentell Inhalte elektronischer Datenbanken analysieren.

Informetrische Analysen verfolgen zwei Ziele. Zum einen geht es darum, im Sinne von Data Mining neue Informationen zu erstellen. Zum andern benutzen wir die informetrischen Ergebnisse als Zwischenschritt, um das Retrieval nach Dokumentationseinheiten zu verfeinern.

Rangordnungen im Dienste des Data Mining. Informetrische Verteilungen sind in der Regel linksschief, d.h. recht wenige Items vereinigen große Mengen an Werten auf sich. Diese Konzentration gilt in der Informetrie als Gesetzmäßigkeit, die in unterschiedlichen Kontexten u.a. von G.K.*Zipf*, S.C.*Bradford*, A.J.*Lotka* und E.*Garfield* empirisch bestätigt worden ist.

Die Suche nach Rangordnungen nutzt das informetrische Gesetz aus. Diese Suche verläuft stets in zwei Schritten. Im ersten Schritt ist diejenige Menge an Datensätzen zu bestimmen, die auszuwerten ist. Suchen wir die patentaktivsten Unternehmen in einem bestimmten Technikfeld, so müssen wir alle Patente zum Technikfeld suchen. Brauchen wir einen neuen wissenschaftlichen Mitarbeiter in einem Spezialgebiet, so öffnen wir eine einschlägige Datenbank und suchen alle Artikel, die das Spezialgebiet thematisieren. Die Datensätze werden nun aber nicht ausgegeben, sondern gewisse Feldinhalte im zweiten Schritt in eine gewünschte Rangordnung gebracht. Argumente der Reportausgabe sind das Feld, ggf. eine Beschränkung der Zeichenfolge im Feld (z.B. nur die IPK-Viersteller), die Sortierrichtung (absteigend, aufsteigend) sowie die Anzahl der Werte (etwa: die ersten Zehn). Wichtig bei der Software ist die obere Grenze an verarbeitbaren Datensätzen. Wollen wir sinnvoll Patent- und FuE-Statistik (etwa von Ländern) betreiben, so müssen durchaus mehrere Millionen Datensätze bearbeitet werden (und nicht nur wenige zehntausend, die einige Online-Archive derzeit zur Bearbeitung anbieten).

Rangordnungen zur Verfeinerung einer Retrievalstrategie. Rangordnungen können auch zur Abstimmung einer Retrievalstrategie dienen. Nehmen wir an, wir seien mit einem unbekannten Sachverhalt konfrontiert und finden keinen passenden Deskriptor. Wir geben im Basic Index Terme ein, die unseren Sachverhalt mehr oder minder gut beschreiben. Das Ergebnis wird informetrisch im Sinne einer Rangordnung bearbeitet. Wir erstellen eine Rangfolge der Deskriptoren, absteigend nach Häufigkeit (oder Auftretenswahrscheinlichkeit in %) sortiert. An der Spitze dieser Rangordnung stehen diejenigen Deskriptoren, die die Indexer zu den (mehr oder minder gut passenden) Texten vergeben haben. Mit den häufigsten zwei bis drei Deskriptoren müssten wir gut weiterarbeiten können.

Informetrische Zeitreihen. Die Fragestellung bei informetrischen Zeitreihen ist: Wie entwickelt sich ein Thema im Laufe der Zeit? Im Gegensatz zu Zeitreihendatenbanken, in denen die Zeitreihen abrufbar gespeichert vorliegen, erstellen wir im Rahmen des Data Mining die gewünschte Zeitreihe aktiv durch das informetrische Retrieval. Zunächst werden alle Datensätze zum Thema selektiert. Ausgegeben wird ausschließlich der Inhalt des Feldes "Jahrgang" mit der zugehörigen Anzahl der Dokumentationseinheiten. Eine sinnvolle Aufbereitung wäre die Darstellung der zeitlichen Entwicklung als Graphik.

Semantische Netze im Data Mining. Suchbare Aspekte wie u.a. Autoren, Deskriptoren, Notationen, Zitationen, Wörter in Abstracts oder im Volltext kommen in den Dokumentationseinheiten nicht isoliert voneinander vor, sondern stehen in irgendeiner Beziehung zueinander. Die Methode, solche Beziehungen aufzuspüren, ist die Clusteranalyse, eine sinnvolle Darstellung der Cluster ist die Form der semantischen Netze. Eine (von mehreren möglichen) Berechnungsformeln für Ähnlichkeit zwischen gewissen Items ist der Jaccard-Sneath-Koeffizient. Gezählt wird die Häufigkeit (a) des Auftretens von Item A (etwa einem Deskriptor in einem Datensatz) in einer Datenmenge, die Häufigkeit (b) eines Items B (einem zweiten Deskriptor) in der selben Menge sowie die Häufigkeit (g) des gemeinsamen Auftretens von A und B im selben Datensatz. Der Jaccard-Sneath-Koeffizient von A und B errechnet sich nach der Formel g / a + b - g. Die Ähnlichkeit von A und B ist Null, wenn g gleich Null ist; die Ähnlichkeit von A und B ist 1, wenn a gleich b gleich g (wobei g ungleich Null) ist, d.h. wenn A und B stets gemeinsam auftreten. Der Ähnlichkeitskoeffizient wird für alle infragekommenden Paare einer Menge berechnet.

Semantische Netze, wie wir sie beim informetrischen Data Mining benutzen wollen, sind ungerichtete Graphen, an deren Knoten Elemente von Datensätzen (Autorennamen, Deskriptoren, Textwörter usw.) und an deren Pfaden Bindungsstärken zwischen den Elementen (wie der Ähnlichkeitskoeffizient von Jaccard und Sneath) abgetragen werden. Bei beiden Aspekten wird mit Schwellenwerten gearbeitet. Es kommen erstens nur solche Datensatzelemente in ein semantisches Netz, die bei der Rangordnung einen zu definierenden Wert überschreiten. Und zweitens werden nur solche Bindungsstärken eingezeichnet, die einen Schwellenwert bei der Ähnlichkeit erreichen.

Beim Data Mining zeigt uns die Clusteranalyse - insbesondere bei großen und deshalb ggf. unübersichtlichen Mengen - eine Strukturierung von Informationen. Hat ein Unternehmen beispielsweise mehrere Forschungsschwerpunkte, so dürften diese im semantischen Netz differenziert aufscheinen.

Semantische Netze als Basis der Einengung von Suchfragen. Clusteranalyse bzw. semantische Netze haben eine nicht zu unterschätzende Bedeutung beim Information Retrieval. Gesetzt den Fall, wir erhalten bei einer Recherche eine viel zu große Treffermenge, die als Ganzes nicht auswertbar ist. Über das semantische Netz erfährt der Rechercheur, dass unterschiedliche abgrenzbare Teilmengen existieren (unter der Voraussetzung, dass es sie tatsächlich gibt), so kann er seine Suche darauf abstimmen. Eine solche Clusteranalyse wird beispielsweise von Verity angeboten.

Es sind zwei Strategien möglich. Zum einen kann man positiv zu einem bestimmten Cluster weitersuchen. Man engt seine Suche - unter Vernachlässigung aller anderen Themen - auf dieses eine Thema ein. Die zweite Strategie arbeitet negativ, indem gewisse Cluster ausgeschlossen werden. Gesucht wird nunmehr gezielt unter Ausschluss dieser Cluster. Beide Strategien garantieren, dass die Treffermenge reduziert wird.

Informationsflussgraphen. Analysen von Informationsflüssen beantworten Fragen der Art: Fließen Informationen von A nach B? Informationsflüsse können dann nachgezeichnet werden, wenn über sie berichtet wird. In systematischer Form haben sich drei Arten solcher Berichtssysteme etabliert: in der wissenschaftlich-technischen Literatur, in der Patentliteratur, bei Urteilen. Ein wissenschaftlicher Artikel, ein Patent oder ein Gerichtsurteil deutet in einer Zitation an, woher gewisse Informationen stammen. Immer dann, wenn Datenbanken Zitationen nachweisen, können Informationsflüsse im Retrieval rekonstruiert werden. Im Gegensatz zu den oben beschriebenen semantischen Netzen sind Informationsflussdarstellungen gerichtete Graphen, wobei der Pfad die Richtung der Informationsübermittlung von Sender zum Empfänger anzeigt.

Anwendungsbeispiel sei die Stellung eines Unternehmens im Informationsfluss. Den Input an Informationen erhalten wir, wenn wir die wissenschaftlich-technische Literatur sowie die Patente des Unternehmens nach Zitationen durchsehen. Der Output ergibt sich, indem wir nach Zitierungen dieser Literatur und der Patente recherchieren. Informationsflussgraphen zeigen die Stellung eines Unternehmens (oder auch eines einzelnen Forschers) im wissenschaftlich-technischen Informationsfluss. Wenig Informationsinput, aber viel Output deuten auf die Rolle eines Technologieführers hin; umgekehrt wenig Output bei viel Input auf einen Nachahmer.

Retrieval bei Zitationsdatenbanken. Neben der Bedeutung der Informationsflussanalysen für das Data Mining besteht auch hier eine Funktion für das "normale" Information Retrieval, gibt es doch spezielle Retrievalstrategien, die an Zitationsdatenbanken durchführbar sind: Informationsflüsse "nach hinten", Informationsflüsse "nach vorne", Vergleich von Zitationsapparaten. Voraussetzung für das er-folgreiche Retrieval bei Zitationsdatenbanken ist ein Ausgangsdokument, das für ein Informationsproblem einschlägig ist. Die Informationsflüsse "nach hinten" zu verfolgen, heißt, alle zitierte Literatur zu suchen. Dies ist kein großes Problem, das auch ohne elektronisches Retrieval möglich wäre. Ganz anders ist dies bei Suchen "nach vorne". Hier recherchieren wir alle die Artikel, Patente usw., die unser Ausgangsdokument zitieren. Solch ein Retrieval ist ausschließlich bei Zitationsdatenbanken durchführbar. Als Beispiele können *Web of Science* für wissenschaftliche Literatur und *Derwent Innovation Index* für Patente dienen.

Automatische Indexierung und natürlichsprachiges Retrieval

In diesem Abschnitt können wir keine Befehlslisten aufführen. Auf der Nutzerseite gibt es nur wenige: etwa das Suchen mit (irgendeinem) natürlichsprachigen Ausdruck oder mit einem Musterdokument. Aber wir können Algorithmen nennen, die für das automatische Indexing bzw. Retrieval wesentlich sind. Systeme sind nämlich um so leistungsfähiger, je elaborierter die Menge dieser Algorithmen ist.

Informationslinguistik. Informationslinguistische Methoden, die in Retrievalsystemen Einsatz finden (bzw. finden sollten), haben vorwiegend die Aufgabe, die Bildung von Sucheinstiegen in einen Text vorzubereiten. Die Methoden werden sowohl auf die Datensätze als auch auf die Suchfragen angewandt.

Triviale Voraussetzung aller weiterer Schritte ist die Isolation einzelner Wörter bzw. von Zeichenfolgen mit n Elementen. Die erste Variante arbeitet mit Wörtern, d.h. Zeichenfolgen, die zwischen zwei Leerzeichen, Satzzeichen o.ä. stehen. Die zweite Variante zerlegt Zeichenfolgen in sog. "N-Gramme", in Tupel, zu jeweils n Zeichen.

Während mengentheoretisch vorgehende Retrievalsysteme mit rund zehn Stoppwörtern auskommen, erhöhen natürlichsprachige Systeme deren Anzahl auf bis zu 500. Eingeschlossen sind Wörter, von denen erwartet wird, dass sie normalerweise keine eigenständigen thematischen Passagen ausdrücken. Stoppwortkombinationen sind nicht zu eliminieren, sondern müssen als Phrase suchbar bleiben. Die Terme "be", "not", or" und "to" sind jeweils Stoppwörter und damit einzeln nicht suchbar. Die Phrase "to be or not to be" dagegen sollte durchaus abfragbar bleiben.

Grammatikalische Flexionsformen werden auf deren Wortstamm reduziert. In der englischen Sprache reicht es häufig aus, bei Beachtung einiger Regeln die Endungen (über eine vorgegebene Suffixliste) zu tilgen. Zu den Suffixlisten treten gewisse Regeln. Ist z.B. "-ing" in der Liste enthalten, wird das Wort "ringing" auf den Stamm "ring" reduziert, aber durch eine an der Länge orientierten

Regel nicht auf "r". Alle Varianten eines Wortstammes gelten dabei als dasselbe Wort. In der deutschen Sprache (aber auch für unregelmäßige Formen des Englischen) ist das Nutzen der Suffixlisten nicht zielführend. Hier arbeitet man mit Lemmatisierung, d.h. mit der in der Regel an Listen orientierten Rückführung auf das Lemma ("Kühen" auf "Kuh").

Phrasen müssen automatisch als solche erkannt werden. Eine Phrase ist ein Ausdruck, der aus mehreren einzelnen Wörtern besteht. Hier gilt nicht das einzelne Wort (oder dessen Wortstamm) als Schlagwort, sondern die Phrase als Ganzes. Das System muss beispielsweise in der Lage sein, "Information Retrieval", den Körperschaftsnamen "Institute for Information Economics" oder den Eigennamen "Willi Bredemeier" als eine Einheit zu erkennen. Phrasen werden identifiziert, wenn sie bereits in einer Liste eingetragen sind. Für die Erkennung neuer Phrasen gibt es zwei Methoden, die sich gegenseitig ergänzen. Die erste Methode arbeitet mit "Indikatorbegriffen". Indikatorbegriffe für Personennamen sind Vornamen, für Unternehmen einschlägige Abkürzungen ("Ltd.", "AG", "GmbH"). Die zweite Methode zerlegt den Text in "Textklumpen". Zu diesem Zweck wird die Stoppwortliste massiv erweitert, sie enthält alle Adverbien, alle Hilfsverben sowie weitere Verben. Betrachtet werden die Wörter, die zwischen den Stoppwörtern übrigbleiben. Stehen hier Einwortbegriffe, so werden diese außer acht gelassen. Bleiben jedoch mehrere nebeneinanderstehende Wörter übrig, so sind dies Kandidaten für Phrasen. Eine neue Phrase wird in das Phrasenwörterbuch aufgenommen, wenn ihre Auftrittshäufigkeit einen Schwellenwert überschreitet.

In der deutschen Sprache sind Sonderformen zu beachten. So muss ein System die Bindestrichergänzung (aus "Film- und Fernsehwirtschaft" den Begriff "Filmwirtschaft" ableiten) genauso beherrschen wie die Kompositazerlegung (in "Unternehmensfusion" die beiden Begriffe "Unternehmen" und "Fusion" erkennen).

Rangordnung
Zeitreihe
Semantisches Netz
Informationsflussgraph

Tabelle 2: Informetrische Funktionen

Das Problem der Synonyme und Homonyme stellt sich hier genauso wie bei den Basisfunktionen. Synonyme sind zusammenzufassen; bei Homonymen wird beim Benutzer nachgefragt.

Wichtig ist bei mehrsprachigen Informationssammlungen (darunter das gesamte World Wide Web) ein sprachunabhängiger Zugang zu den Dokumenten. Der Nutzer gibt seine Suchformulierung in seiner Landessprache ein, und das System übersetzt in alle gewünschte Sprachen. Solch ein multilingualer Zugang wird durch (intern verwaltete) Wörterbücher unterstützt.

Die sinntragenden Wortstämme sowie die Phrasen werden, wie wir gleich sehen werden, als Zählbasis für statistische Berechnungen verwendet. Es ist demnach die Markierung jedes Vorkommens in einem Text nötig. Pronomina, die an die Stelle ihrer Nomen rücken, müssen dabei entsprechend beachtet werden. (Deshalb dürfen Pronomina erst nach diesem Arbeitsschritt als Stoppwörter eliminiert werden.) Betrachten wir den Satz: "The president has a girl friend, but he doesn't love her", so muss das "he" dem Wort "president" und "her" der Phrase "girl friend" zugeordnet werden. Die Auftretenshäufigkeit von "president" und von "girl friend" im Beispielsatz ist demnach jeweils gleich zwei.

Suchfragen können Schreibfehler beinhalten. Retrievalsysteme müssen deshalb in der Lage sein, mit Fehlern seitens der Benutzer fertig zu werden. Eine Lösungsmöglichkeit wurde mit dem phonetisch orientierten "Soundex"-Verfahren beim Projekt *Okapi* ausgearbeitet.

Informationsstatistik zählt Wörter (bzw. Lemmata und Phrasen sowie deren Pronomina) sowie Datensätze und setzt die ausgezählten Werte in bestimmte Relationen. Ziel informationsstatistischer Methoden ist die Rangordnung von Dokumenten nach deren Relevanz relativ zur Suchfrage.

Die Primitivvariante von Informationsstatistik zählt Wörter in einem Datensatz - je häufiger desto wichtiger. Obwohl durchaus bei Suchmaschinen eingesetzt, ist diese Methode nahezu völlig untauglich, da der Schluss von der Worthäufigkeit auf Relevanz nicht zwingend ist.

Ein erster Algorithmus ist die dokumentspezifische Wortgewichtung (within-document frequency weight) WDF, d.i. die relative Häufigkeit von Textwörtern (Quotient aus der Häufigkeit des Wortes und der Gesamtmenge der Wörter im Text, normalerweise logarithmisch ausgedrückt). WDF arbeitet niemals alleine, sondern nur im Verbund mit mindestens einem weiteren Wert, der "inversen Dokumenthäufigkeit" (inverse document frequency weight) IDF. Während ein Textwort im Kontext eines Textes um so wichtiger ist, je häufiger es vorkommt, gilt in Beziehung zur Gesamtdatenbank die umgekehrte Proportion. Ein Wort ist demnach um so wichtiger, je weniger Dokumente dazu in der Datenbank vorhanden sind. Um die Spannweite der Gewichtungswerte nicht allzu groß werden zu lassen, arbeitet man auch bei der Berechnung der inversen Dokumenthäufigkeit mit logarithmischen Werten. Die klassische Berechnungsformel von Karen Sparck Jones lautet: IDF(i) = (log2 N / n) + 1, wobei IDF(i) die inverse Dokumenthäufigkeit des Wortes i, N die Gesamtanzahl der Datensätze in der Datenbank und n die Anzahl der Datensätze ist, in denen i vorkommt. Der Wert des IDF ist - für jedes Wort - in ständiger Bewegung, insofern die Datenbank wächst und so auf jedem Fall N ändert. WDF / IDF werden verfeinert, indem die Stellung des Wortes in Textteilen Berücksichtigung findet. Wenn man annimmt, dass Terme am Textanfang wichtiger sind als solche in der Mitte, so wird dies durch geeignete Gewichtungsfaktoren berücksichtigt. Eine Variante dieses Verfahrens arbeitet mit den dokumentarischen Beschreibungen der Texte. Die Termgewichtung orientiert sich hier am Auftreten in bestimmten Feldern. Ein Vorkommen im Sachtitel oder in den Meta-Tags (bei HTML-Dokumenten) wird z.B. höher gewichtet als ein Vorkommen im Fließtext.

Für jedes Wort eines Textes ergibt sich durch Multiplikation der dokumentspezifischen Wortgewichtung (mit oder ohne Berücksichtigung der Positionen im Text) mit seiner inversen Dokument- häufigkeit ein Gewichtungswert. Betrachtet man mehrere Wörter, so addieren sich die einzelnen Gewichtungswerte pro Dokument. Diese Summe bildet die Basis für Sortierungen nach Relevanz.

Bei Suchen mit mehr als einem Suchwort existiert eine weitere Variante. Hier geht es um den Abstand zwischen den

Suchargumenten. Stehen die einzelnen Wörter enger zusammen (im Idealfall nebeneinander oder innerhalb eines Satzes), so ist der entsprechende Text wahrscheinlich wichtiger für einen Nutzer als wenn die Suchwörter zwar alle vorkommen, doch bezuglos an unterschiedlichen Textstellen aufscheinen. Der Gewichtungswert für einen Text errechnet sich aus der Anzahl der Wörter, die zwischen den einzelnen Suchargumenten stehen. Der Wert für den Abstand gegebener Suchwörter in einem Text ist um so größer, je kleiner der Wortabstand ist.

Ergänzend anzumerken gilt es informationsstatistische Verfahren, die ausschließlich im World Wide Web anwendbar sind und die auch bei Suchmaschinen bereits Eingang finden. Ein Gewichtungsfaktor ist die "Popularität" eines Links, d.h. ein HTML-Dokument ist um so wichtiger, je häufiger darauf verwiesen wird. Experimentiert wird mit dem Gewichtungsfaktor "Zugriff auf Sites". Hier wird eine WWW-Site um so wichtiger, je öfter bereits auf sie zugegriffen wurde.

Ranking. Aus der informationslinguistischen Analyse der Suchfrage ergibt sich eine Liste von Suchargumenten. Sie beinhaltet nicht nur die vom Benutzer explizit eingegebenen Terme (in lemmatisierter Form), sondern auch deren Synonyme; Phrasen sind erkannt. Einige Wörter (Stoppwörter) fallen weg. Eine erste "rohe" Ergebnismenge sind solche Datensätze, die mit mindestens einem Suchargument übereinstimmen. An dieser Stelle setzen unterschiedliche Verfahren zum Relevance Ranking an: (1) Bei nur einem Suchbegriff ist nur das Verfahren über IDF, WDF bzw. der Position im Text möglich. Nun unterstellen wir, dass mehrere Suchwörter vorhanden sind. Verfahren (2) sortiert nach gemeinsamem Vorkommen; d.h. zuerst kommen die Dokumente, die die meisten Suchwörter enthalten, dann die, die eines weniger haben, usw. Bei der gleichen Anzahl Suchwörter kann (Variante 2a) wie in Verfahren (1) weitergearbeitet werden oder es wird (2b) nach dem Wortabstand sortiert. Sortierverfahren (3) arbeitet mit der Summe der einzelnen Werte für das Produkt aus IDF, WDF und Position. Hier ist es durchaus möglich, dass auch beim relevantesten Treffer nicht alle Suchargumente vorkommen. Bei allen Varianten können (4) weitere Gewichtungsfaktoren (wie z.B. die Linkpopularität) hinzukommen. Nach dem Relevance Ranking dürften die der Suchfrage ähnlichsten Texte oben in der Trefferliste stehen.

Ordnungstheorie. Im Rahmen der Anwendung der Ordnungstheorie bei der automatischen Indexierung geht es um die Zuordnung von kontrolliertem Vokabular (Deskriptoren, Notationen) bzw. von Schlagwörtern zu einem Datensatz. Deskriptoren kann man nur dann verwenden, wenn ein Thesaurus vorliegt, Notationen nur im Rahmen eines Klassifikationssystems. Die Schlagwortvergabe arbeitet ausschließlich mit dem in den Texten vorhandenen Termmaterial.

Eine Dokumentationssprache (Thesaurus oder Klassifikationssystem), die zur automatischen Indexierung herangezogen wird, muss über besonders ausgefeilte Synonymie-Relationen verfügen. Nur so kann das textspezifische Vokabular in das Vokabular der hinterlegten Dokumentationssprache übertragen werden. Über eine große Menge an Nicht-Deskriptoren soll erreicht werden, dass möglichst wenig Termmaterial "übersehen" wird.

Vorauszusetzen sind für die ordnungstheoretische Bearbeitung von Texten die beiden Stufen der Informationslinguistik und Informationsstatistik, d.h. wir verfügen für jeden Text über eine nach Relevanz geordnete Liste der Textwörter bzw. Phrasen. Die Gewichtungswerte aller dem Text zugeordneten Wörter werden mit einem Schwellenwert verglichen. Für die verbleibenden wichtigsten Terme geschieht ein Abgleich mit der Liste der De-skriptoren und ihrer Synonyme. Ein Deskriptor des verwendeten Thesaurus wird dann zugeteilt, wenn er (bzw. sein Wortstamm) selber oder wenn einer seiner Synonyme vorkommt. Analog kann bei Klassifikationen vorgegangen werden. Hierbei werden Notationen dann vergeben, wenn die entsprechenden natürlichsprachigen Bezeichnungen der Notationen im Text vorkommen. Diese Variante der automatischen Indexierung (eingesetzt z.B. beim "InfoSort"-Algorithmus von *Profound)* fundiert ein "normales" Boolesches Retrieval mit kontrolliertem Vokabular.

Wir nehmen nun an, dass das System automatischer Indexierung über keine Dokumentationssprache verfügt, trotzdem aber einzelne Terme als Sucheinstiege in einen Text auszeichnen möchte. Solche "auszuzeichnenden" Terme bezeichnen wir als "Schlagwörter". Alle nach den informationslinguistischen und -statistischen Methoden markierten Textwörter sind Kandidaten für Schlagwörter. Gemäß Gerald Salton kann man jedes Schlagwort als Dimension verstehen. Eine Datenbank mit n verschiedenen Schlagwörtern spannt demnach einen n-dimensionalen Raum auf. Die Repräsentation eines Dokuments ist nach *Salton* derjenige Vektor, der durch alle jeweils vergebenen Schlagworte verläuft. Nach diesem Grundgedanken wurde *Saltons* berühmtes System SMART, Salton's Magical Automated Retrieval Technique, kreiert. Da die Terme durch die statistischen Verfahren mit einem dokumentspezifischen numerischen Wert versehen sind, bekommt unser Vektor auf jeder Dimension einen konkreten Wert zugeordnet. Es ist hierbei möglich, mit einem Schwellenwert zu arbeiten. Ein Vektor läuft nur dann durch eine Dimension, wenn der Schwellenwert überschritten wird. Durch die Analyse der Vektoren lässt sich ein "statistischer Thesaurus" aufbauen. Dimensionen, die von mehreren Vektoren gemeinsam berührt werden, gelten dabei als "verwandt" oder "ähnlich". Hierbei wird in der Regel mit mehreren Schwellenwerten gearbeitet. Gefordert ist ein hoher Gewichtungswert der einzelnen Textwörter sowie eine möglichst große Zahl gemeinsamen Auftretens unterschiedlicher Schlagwörter innerhalb der Vektoren. Begriffe, zwischen denen solch eine "statistische Ähnlichkeit" festgestellt wurde, können in der Recherche wie Synonyme behandelt werden. Wenn also der Nutzer einen dieser Terme eingibt, wird automatisch oder nach Systemrückfrage auch nach den anderen "statistisch ähnlichen" Termen gesucht.

Nach dem Gesagten ist die Bestimmung der automatischen Indexierung bei der Variante über Schlagworte recht einfach. Sowohl die Texte in der Datenbank als auch die Suchanfragen werden durch das geschilderte Extraktionsverfahren indexiert. Die Suche ist ein Abgleich des Vektors, der die Suchanfrage repräsentiert, mit den Vektoren der Dokumente der gesamten Datenbank. Ausgegeben werden die Dokumente derjenigen Vektoren, die dem Suchvektor am ähnlichsten sind.

Informationslinguistik
--- Isolation einzelner Wörter
--- Stoppwortmarkierung
--- Wortstammanalyse (Stemming, Lemmatisierung)
--- Phrasenerkennung (Listenabgleich, Indikatorbegriffe, Textklumpen)
--- Synonyme / Homonyme
--- multilingualer Zugang
--- Pronomina
--- Fehlertolerante Frageformulierung
--- Bindestrichergänzung
--- Kompositazerlegung
Informationsstatistik
--- Zählung von Worthäufigkeiten
--- Dokumentspezifische Wortgewichtung (WDF)
--- Gewichtung nach Position im Text
--- Inverse Dokumenthäufigkeit (IDF)
--- (bei mehreren Suchwörtern) Wortabstand
--- (bei WWW-Dokumenten) Popularität eines Links
--- (bei WWW-Dokumenten) Zugriffshäufigkeit
Ranking
--- (1) nur ein Suchwort: IDF * WDF * Position
--- (2) mehrere Suchwörter: 1. Schritt: Anzahl gemeinsam vorkommender Wörter
--- (2a) 2. Schritt: IDF * WDF * Position
--- (2b) 2. Schritt: Wortabstand
--- (3) mehrere Suchwörter: IDF * WDF * Position
--- (4) zusätzlich weitere Gewichtungsfaktoren
Ordnungstheorie
--- Deskriptor- bzw. Notationsvergabe
--- Schlagwortvergabe
Assoziatives Retrieval (Suche mit Musterdokumenten)
--- Suche über Schlagwörter
--- Suche über Zitationen
--- Suche bei nicht-textlichen Datensätzen
Zusammenspiel aller Suchmöglichkeiten

Tabelle 3: Algorithmen automatischer Indexierung und natürlichsprachigen Retrievals

Assoziative Suche. Im Booleschen Retrieval benutzen wir exakte Suchbegriffe, bei der bisher geschilderten automatischen Indexierung natürlichsprachige Sätze. Die "assoziative Suche" arbeitet mit ganzen Dokumenten als Suchargumenten. Der Nutzer benötigt hierzu ein "Musterdokument" als Ausgangspunkt für den Recherchevorgang. Retrievalsysteme arbeiten mit drei Varianten der assoziativen Suche: (1) Die Suche über Schlagwörter arbeitet mit den oben bereits geschilderten natürlichsprachigen Verfahren der automatischen Indexierung. Das Musterdokument wird automatisch indexiert, d.h. aus dem Text werden die relevanten Schlagwörter extrahiert. Je nach systemintern eingestelltem Schwellenwert handelt es sich hierbei um einige wenige bis ca. ein Dutzend Schlagwörter. Die gefundenen Suchargumente werden in der Regel dem Nutzer angezeigt, so dass dieser seine Suchfrage bearbeiten kann, indem er unerwünschte Terme löscht oder weitere Wörter hinzufügt. Es schließt sich eine normale Suche im Rahmen der automatischen Indexierung an. Rechercheergebnis ist eine Menge von Dokumenten, die - nach Relevanz sortiert - dem Musterdokument jeweils mehr oder weniger entsprechen. (2) Eine zweite Variante assoziativen Retrievals ignoriert den gesamten Fließtext eines Dokuments und arbeitet ausschließlich mit den Zitationen. Das Verfahren geht von der Idee aus, dass zwei Texte, die viele andere Texte gemeinsam zitieren, miteinander verwandt sind. Operationalisiert wird die Idee durch das Auszählen des Vorkommens gemeinsamer Zitationen. Dem Musterdokument wird die Liste der Zitationen entnommen. Im assoziativen Retrieval werden alle anderen Dokumente mit dieser Liste verglichen. Rohmenge für die Ausgabe sind Texte, die mindestens eine Fußnote mit dem Musterdokument gemeinsam haben. Die Sortierung nach Relevanz wird über die Menge gemeinsamer Zitationen bestimmt, d.h. oben stehen solche Texte, die die meisten Zitationen mit dem Musterdokument vorweisen. (3) Bei Informationssammlungen mit nicht-textuellen Inhalten, z.B. Bildern oder Videosequenzen, findet man durch den Vergleich zwischen den Bildern ähnliche weitere Bilder. Die Bilder werden in Parzellen gerastert, und für die Teile werden Graustufen- oder Farbhistogramme erarbeitet. Zusätzlich können Strukturen oder Umrisse des Gesamtbildes Eingang in den Suchalgorithmus finden.

Wichtig für den Nutzer ist ein reibungsloses Zusammenspiel von Basisfunktionen, informetrischen Befehlen und natürlichsprachigen Aspekten, so etwa das Weiterbearbeiten einer im mengentheoretischen Retrieval gefundenen Treffermenge mittels Relevance Ranking oder die Boolesche Bearbeitung einer natürlichsprachig erhaltenen Treffermenge. ■

Wolfgang G. Stock

Weiterführende Literatur

F. Wilfrid Lancaster: Information Retrieval Systems. Characteristics, Testing, Evaluation. - New York: Wiley, 1979.

Eleonore Poetzsch: Information Retrieval. Einführung in Grundlagen und Methoden. - Potsdam: Verlag für Berlin-Brandenburg, 1998.

Wolfgang G. Stock: Informationswirtschaft. Management externen Wissens. - München; Wien: Oldenbourg, 2000.