Automatische und intellektuelle Indexierung

Das “Big Systems” - Syndrom
Eine Antwort auf Robert Fugmann

Robert Fugmanns Buch zur Inhaltserschließung und sein Artikel in "Password" Nr. 10/1999 verneinen die Möglichkeit zufriedenstellender automatischer Indexierung. Es lässt sich jedoch zeigen, dass Fugmanns theoretische Basis der totalen Ablehnung natürlichsprachiger Systeme kaum haltbar ist. Zudem ist seine Einschätzung falsch, dass höchstens bei kleinen Systemen automatische Indexierung laufen könnte. Es sind im Gegenteil die großen Systeme, die automatische Indexierung zwingend benötigen. Und hier "läuft" automatisches Information Indexing und Retrieval erfolgreich, theoretisch verankert durch Informationslinguistik und -statistik sowie durch ordnungstheoretische Vorgaben und technisch abgesichert durch Patente. Natürlich sind Verbesserungen möglich und nötig; die Basisarbeit ist - für das Englische - bereits geleistet. Ganz sicher richtig ist Fugmanns Ansicht, dass ein Optimum an Retrievalqualität durch einen kombinierten Einsatz beider - also intellektueller und automatischer - Methoden gewährleistet werden kann, wobei Fugmann die automatische Indexierung jedoch nur als "Ergänzung" zur intellektuellen versteht. Im Gegensatz dazu müssen wir jedoch angesichts der großen Systeme und des daraus folgenden "Big-Systems' Syndrom" betonen: Es geht nur mithilfe der Kombination beider Ansätze der Inhaltserschließung.

Robert Fugmann vermutet in seinem Artikel (in Password 10/1999), dass die "Erfolge" der automatischen Indexierung ausschließlich unter Bedingungen experimenteller, relativ kleiner Informationssysteme scheinbar eintreten können. "Je kleiner die experimentellen Datenbanken sind und je oberflächlicher mit ihnen gearbeitet wird, insbesondere bei der Ermittlung von Informationsverlust, desto erfolgreicher und aussichtsreicher erscheinen paradoxerweise solche praktischen Beispiele" (S. 6). Automatische Indexierung geht nach Fugmann nämlich gar nicht, da der Prozeß der Inhaltsabbildung prinzipiell **indeterminiert** sei, und solche indeterminierten Prozesse sind weder formalisierbar noch programmierbar. Zur Begründung der Indeterminiertheit bemüht Fugmann die Hermeneutik, die Lehre des (richtigen) Verstehens und Interpretierens. "In dem Verzicht auf Interpretation liegt der Ursprung einer Reihe von weiteren Mängeln bei der mechanisierten Textverarbeitung" (ebd.). Die "Mängel" betreffen u.a. die so nicht mögliche Zuteilung recherchenützlicher Schlagwörter oder Deskriptoren, die Lexikalisierung von Paraphrasen und die zuverlässige Erkennung der "Essenz" eines Dokuments.

Ich möchte mich hier auf zwei Aspekte von Fugmanns Kritik konzentrieren: auf die theoretischen Voraussetzungen von Fugmanns irriger Ansicht und auf das "Small systems' Syndrom". Bei letzterem zeigt sich nämlich die gravierende Fehleinschätzung Fugmanns. Wir sind in der Informationswirtschaft mitnichten mit den "small systems" der 60er Jahre konfrontiert, sondern mit riesigen Systemen. Man denke hierbei nur z.B. an die Menge der Sites im World Wide Web, an große Pressedatenbanken (die Reuters-Datenbank wächst um knapp 30.000 Volltexte täglich) oder an die riesigen Volltextdatenbanken zu News, Gerichtsurteilen oder Patenten beispielsweise bei Lexis-Nexis.

Brauchen wir Informationshermeneutik beim Indexieren?

Zunächst müssen wir dem grundsätzlichen Bedenken Fugmanns begegnen, dass wegen der hermeneutischen Aspekte automatische Indexierung unmöglich sei. Fugmann setzt stillschweigend voraus, dass (1.) eine Zuteilung von Begriffen nötig sein kann und (2.), dass ein Indexer stets interpretierend an die Arbeit geht. Schon vor 30 Jahren wurde diesem widersprochen. Norbert Henrichs betont 1970 die Faktizität eines gegebenen Textes und spricht den Texten - für die Zwecke der Indexierung - einen "latenten Sinngehalt" ab. Ein Indexer, der nach der Henrichs´schen Textwortmethode arbeitet, wird demnach stets auf den konkreten Sprachgebrauch in einem Text zurückkommen müssen, zuteilende Indexierung ist hier untersagt. Interpretation ist gemäß Henrichs nicht nur nicht notwendig, sondern schädlich. Und automatische Indexierung ist stets interpretationsfrei - im Sinne Henrichs also ein großer Vorteil.

Zuteilende Indexierung - automatisch

Nun verlassen wir die Textwortmethode und wenden uns Dokumentationssprachen, also Thesauri und Klassifikationssystemen, zu. Hier ist zuteilende Indexierung in der Tat manchmal nötig. Aber warum soll dies nicht automatisch geschehen können? Vorausgesetzt, unser Thesaurus verfügt über eine elaborierte Fassung der Nichtdeskriptor-Deskriptor-Beziehung, dann ist es beim Auftreten eines Nichtdeskriptors (ab einem gewissen Schwellenwert in der informationsstatistischen Betrachtung) doch ohne Weiteres möglich, den zugehörigen Deskriptor zuzuteilen. Dies "gilt" nicht nur in der Theorie, sondern wird beim

Online-Archiv Profound mit seinem "Info-Sort"-Thesaurus praktisch vorgeführt (vgl. Password 10/1998).

Auch in Deutschland haben wir ein Referenzobjekt. In der Regel arbeiten Systeme automatischer Indexierung um so besser, je mehr Textmaterial vorhanden ist. Insofern ist der Volltext die ideale Basis automatisierter Ansätze. Aber es geht auch mit (weitaus) weniger Text. Dies konnte die diesjährige Preisträgerin der "Infobox", die Universitätsbibliothek Düsseldorf mit ihrem Projekt MILOS (Maschinelle Indexierung auf linguistischer Basis für OPAC-Systeme) zeigen (vgl. Password Nr. 6/1999, 12). MILOS setzt auf die Sachtitel von Büchern auf und teilt Schlagwörter zu.

Fugmanns stillschweigend unterstellte Voraussetzungen sind nicht haltbar und dies gleichermaßen aus theoretischen Erwägungen wie durch praktische Erfahrungen.

Big Systems und ihre Retrievalprobleme

Das "Big Systems'-Syndrom" begegnet uns in mehreren Varianten.

(1) Datenbanken mit mehreren Millionen Datensätzen sind heute keine Seltenheit. Auch wenn die Texte mittels normierten Vokabulars erschlossen sind, können die Treffermengen immer noch so groß sein, dass sie für den Nutzer - allein aus Massengründen - unbrauchbar werden. Hier hilft die automatische Indexierung weiter, und zwar vor allem in der Form des Relevance Ranking. Die Menge der Suchergebnisse wird derart sortiert, dass die zur Suchfrage bestpassenden Treffer in der Liste oben erscheinen und die Texte von oben nach unten an Relevanz verlieren.

(2) Im World Wide Web wird derzeit überhaupt nicht nach einheitlichen Kriterien intellektuell inhaltlich erschlossen. Suchmaschinen sind notwendig auf Verfahren automatischer Indexierung angewiesen. Michael Burrows, der mehrere Patente zur automatischen Indexierung hält, bemerkt: "In the prior art, it has been well known that computer systems can be used to index information stored in records of databases. Many techniques are known to parse and index databases. However, indexing extremly large databases presents special problems". Seine Patente zeigen, dass automatische Indexierung theoretisch wie technisch möglich ist, das System, das auf diesen (und weiteren) Patenten aufbaut ("AltaVista"), ist eine der "Top"-Suchmaschinen im World Wide Web.

(2a) Auch ist die Möglichkeit zu bedenken, dass ein Informationsproduzent außerhalb des Web nicht die Mittel hat, eine intellektuelle Inhaltserschließung durchführen zu lassen. Für Unternehmen wie beispielsweise das Handelsblatt oder die F.A.Z. ist es lohnend, einen informationellen Mehrwert über diverse Klassifikationssysteme zu erstellen, lassen sich doch die so entstehenden elektronischen Versionen gewinnbringend verkaufen. Für kleinere Zeitungen, denken wir etwa an den "Trierischen Volksfreund", lohnt dies nicht. Für den Nutzer wäre aber ein Zugriff via Relevance Ranking neben dem "normalen" Volltextretrieval sicherlich sinnvoll - zumindest besser als gar keine Erschließung (wie derzeit).

(3) Große Systeme sprechen weite Nutzerschichten an. Im World Wide Web recherchieren nicht nur Information Professionals, sondern auch "Laien". Hier müssen Methoden angeboten werden, mit denen jeder Laie klarkommt. Eine wichtige Rolle spielen dabei die natürlichsprachig formulierte Suchfrage und eine fehlertolerante Abarbeitung des Suchausdrucks - beides Aspekte von automatischem Indexing bzw. Retrieval.

Boolesches und natürlichsprachiges Retrieval

Ein Beispiel sei gestattet. Beim Online-Archiv Lexis-Nexis ist die automatische Indexierung ("Freestyle") seit Jahren neben dem Booleschen Retrieval installiert (vgl. Password 11/1998). Durch die Verbindung beider Retrievalarten lassen sich interessante Suchstrategien abarbeiten. In einem ersten Schritt grenzt man durch Freestyle eine umfangreiche Menge von Datensätzen auf die relevantesten Dokumente ein. Im zweiten Schritt arbeitet man in dieser Teilmenge mit den gewohnten Booleschen Operatoren. Der umgekehrte Weg ist

ebenfalls möglich. Zunächst wird mengentheoretisch gesucht, danach mittels Relevance Ranking sortiert. Geschickt eingesetzt, läßt sich die Präzision einer Treffermenge so enorm erhöhen.

Ähnliche Kombinationen aus Boolescher Suche und Relevance Ranking bieten inzwischen diverse Online-Archive und Suchmaschinen im World Wide Web an.

Kleine Systeme können zur Not auf automatische Indexierung verzichten; hier arbeiten auch intellektuelle Verfahren erfolgreich; große Systeme können dies nicht; "Zur Not" bei kleinen Systemen, weil auch diese, etwa Intranets von Unternehmen, eigentlich natürlichsprachige Interfaces bräuchten. Warum muten wir unseren Nutzern eigentlich zu, in Booleschen Operatoren oder Abstandsoperatoren zu denken, anstatt sie ihre Fragen so stellen zu lassen, wie sie es gewöhnt sind: in sprachlich einigermaßen korrektem Deutsch oder Englisch?

Warum soll ein Nutzer denn so etwas Kryptisches wie "Waschen+7" - Fugmanns Vorschlag zufolge (Fugmann 1999, 89) - eingeben, wenn er auch "Ich brauche Neuigkeiten über Waschmaschinen" oder schlicht "Waschmaschinen?" schreiben kann?

Relevance Ranking und Indexieren mit normiertem Vokabular in Kombination

Der Schlüssel für zufriedenstellendes Retrieval in großen Systemen liegt in der Kombination mehrerer Methoden. Es ist nichts dagegen einzuwenden, dass auch intellektuell indexiertes Material vorliegt, aber dies ist nicht zwingend notwendig. Wichtig ist vielmehr, dass es überhaupt einen Zugang über normiertes Vokabular gibt und dass Relevance Ranking möglich ist. Wir können das Thema in dieser kurzen Replik nicht vertiefen. Nur soviel soll angedeutet werden (vgl. auch Password Nr. 2 / 1999): Ein Optimum an Retrievalqualität bzl. der Indexierung in großen Systemen ist an folgende Voraussetzungen geknüpft:

- Die Informationen liegen im Volltext vor (möglichst mit einer Ausgabe im Originallayout).
- Das Recherchesystem läßt sowohl exaktes Suchen als auch natürlichsprachig eingegebene Suchargumente zu.
- Die Volltexte sind nicht nur durch mengentheoretische (Boolesche) Operatoren suchbar, sondern darüber hinaus durch eine Vielzahl von Abstandsoperatoren.
- Die Volltexte sind durch Deskriptoren bzw. durch Notationen inhaltlich erschlossen - ob dies nun Resultat intellektueller menschlicher Arbeit oder automatischen Indexierens ist, ist für den Nutzer belanglos.
- Über Algorithmen der Informationslinguistik und -statistik sind Rangordnungen der Volltexte relativ zum Suchargument möglich (Relevance Ranking), da der Nutzer bei großen Systemen ansonsten von einer unstrukturierten Treffermenge schlicht "erschlagen" würde.
- Die Nutzeroberflächen sind so beschaffen, dass auch Laien damit zurechtkommen.

Fugmanns Anschauungen haben sich gewandelt. Im Buch hieß es: "Die automatische Indexierung ist in all ihren Varianten im Vergleich überwiegend negativ zu beurteilen. Dies gilt auch für die Aussicht auf künftige nachhaltige Qualitätsverbesserungen" (Fugmann 1999a, 133). In Password 10 / 1999 wird diese doch sehr starke Behauptung relativiert: "In meinem Buch ist (auf Seite 133) der Nutzen von automatischer Indexierung allzu kurz gekommen ... Als Ergänzung zum intellektuellen Indexieren ist automatisches Indexieren interessant, besonders dann, wenn intellektuelles Indexieren nur oberflächlich betrieben wird oder nur betrieben werden kann" (S. 8). Dies deckt sich zu weiten Teilen mit der Theorie und Praxis der Informationswirtschaft, wie ich sie sehe.

Online-Archive und Information Professionals, die das "Big-Systems'-Syndrom" nicht berücksichtigen, können durchaus "auf das berufliche Abstellgleis geschoben werden" (Fugmann in Password 10 / 1999, S. 8, Anm. 6). Wir leben nicht mehr in der Zeit von Cranfield, sondern von In-STAR-TREC-8. ■

Wolfgang G. Stock

Literatur

Michael Burrows: *Method For Indexing Information of a Database*. - Patent Nr. US 5,745,899 vom 28.4.1998.

Robert Fugmann: The Five-Axiom Theory of Indexing and Information Supply. - In: *Journal of the American Society for Information Science* 36 (1985), 116-129.

Robert Fugmann: *Theoretische Grundlagen der Indexierungspraxis*. - Würzburg: Ergon, 1992.

Robert Fugmann: *Inhaltserschließung durch Indexieren: Prinzipien und Praxis*. - Frankfurt: Deutsche Gesellschaft für Dokumentation, 1999a.

Robert Fugmann: *Mechanisierte Indexierung kann intellektuelle nicht ersetzen*. - In: Password Nr. 10 (1999), 6-8.

Norbert Henrichs: Philosophische Dokumentation. Literatur-Dokumentation ohne strukturierten Thesaurus. - In: *Nachrichten für Dokumentation* 21 (1970), 20-25.

Wolfgang G. Stock: Robert Fugmann: Inhaltserschließung durch Indexieren. Intellektuelles Indexieren für Buchregister und Inhouse-Datenbanken. - In: *Password* Nr. 7-8 / 1999, 26-27.

ALLECO

Expansion in den News-Bereich

Der Wirtschaftsinformationsdienst ALLECO der Deutschen Telekom AG und der ECOFIS Wirtschaftsinformationen GmbH (Password 12/1999) expandiert in den News-Bereich. Zielgruppe sind die kleinen und mittleren Unternehmen. Die Nachrichten werden von dem Deutschen Industrie- und Handelstag (DIHT) zur Verfügung gestellt. Auf längere Sicht strebt ALLECO eine Führungsposition im bislang unterrepräsentierten Online-Markt für kleine und mittlere Unternehmen an.

Auch die Konjunkturstudien von DIHT und Creditreform sind bei ALLECO online gegangen. Hinzukommt allwöchentlich eine Reportage beispielsweise zum Zeitpunkt der Abfassung dieser Meldung über die gängigsten Fehler von Online-Shops.