Information Retrieval

Reginald Ferber

GMD-IPSI

Ein gutes Buch über... einen Artikel zu... ein Foto von ... – nach einem bestimmten Inhalt zu suchen, war schon immer schwierig, besonders wenn man gar nicht genau weiß, ob es das, was gesucht wird, auch tatsächlich gibt; oder wenn man gar nicht genau weiß, was man eigentlich sucht. Im überschaubaren Rahmen hilft eine gute Buchhändlerin, erfahrene Bibliotheksbedienstete oder eine Expertin im jeweiligen Sachgebiet weiter. Schwieriger wird es, wenn das Gebiet, in dem gesucht werden muß, zu groß oder zu unübersichtlich ist, oder wenn keine hilfreiche Seele zur Verfügung steht.

Systematiken, wie in solchen Fällen vorgegangen werden kann, gibt es schon lange. Jede Bibliothek hat ein System, nach dem die Bücher aufgestellt bzw. archiviert werden. Außerdem gibt es Kataloge und Bibliographien, die sie nach verschiedenen Kriterien – wie Autor oder Sachgebiet – erschließen. Diese Verfahren sind zwar zunehmend flexibel (ein Buch kann nur an einer Stelle stehen, ein Katalog kann es unter verschiedenen Einträgen aufführen und eine Bibliographie kann ein komplexes Klassifikations- und Verweissystem verwenden), sie bleiben jedoch, auch wenn sie auf Rechner verlagert werden, aufwendig, und ihre effektive Nutzung setzt entsprechende Kenntnisse voraus. Vor allem aber sind sie statisch: Ist ein Objekt einmal klassifiziert, wird diese Klassifizierung im allgemeinen nicht mehr geändert werden. Der Zugriff auf Objekte ist auf die Einordnung des Inhalts zum Zeitpunkt der Klassifizierung beschränkt, auch wenn sich die Sicht der Dinge im Laufe der Zeit ändern sollte.

Boolesches Retrieval

Schon seit der Frühzeit der elektronischen Datenverarbeitung ist deshalb versucht worden, Inhalte in elektronischer Form zu speichern und über entsprechende Programme flexibel zugreifbar zu machen. Solche Informationssysteme oder auch Information Retrieval (IR) Systeme waren zunächst auf wissenschaftliche Veröffentlichungen beschränkt und bauten auf Darstellungsformen auf, wie sie in Bibliographien verwendet wurden: Angaben wie Titel, Autor und Erscheinungsform, aber auch Stichwörter und Zusammenfassungen (Abstracts) wurden als Textdokumente gespeichert. Der Zugriff auf diese Dokumente erfolgte nach der Methode des sog. Booleschen Retrieval: Recherchierende geben an, welche Wörter oder "Terme" in den gesuchten Dokumenten vorkommen sollen und welche nicht vorkommen dürfen. Das System liefert dann alle Dokumente, die diese Bedingungen erfüllen. Das Verfahren läßt sich verfeinern, indem die Struktur der Dokumente genutzt wird. So kann man angeben, in welchen Teilen eines Dokuments ein Term vorkommen soll, also ob beispielsweise nur im Titel oder nur bei den Autorennamen danach gesucht werden soll. Weiter können Verknüpfungen mit logischem "AND" und "OR" vorgenommen werden: sind zwei Terme mit AND verknüpft, so müssen *beide* im Dokument vorkommen, sind sie mit OR verknüpft, reicht es, wenn der eine *oder* der andere auftritt, damit das Dokument ausgegeben wird.

Bei größeren Dokumentsammlungen kann natürlich nicht bei jeder Anfrage in jedem Dokument nachgesehen werden, ob es die angefragten Terme enthält. Deshalb wird vorher eine Liste angelegt, in der - wie in einem ausführlichen Stichwortverzeichnis - zu jedem Term aufgelistet ist, in welchen Dokumenten der Sammlung er auftritt. Um eine Anfrage zu bearbeiten, kann dann in dieser sog. invertierten Liste für jeden Term der Anfrage abgelesen werden, in welchen Dokumenten er auftritt. Dieses Verfahren spart viel Rechenzeit, es kostet dafür aber auch viel Speicherplatz: Eine invertierte Liste kann genauso viel Speicherplatz einnehmen wie die Dokumente, über die sie Auskunft gibt.

Boolesches Retrieval hat den Vorteil, daß es leicht verständlich ist: es ist klar, warum ein Dokument gefunden wurde. Es hat den Nachteil, daß es eine ungeordnete Menge von Dokumenten liefert, die bei allgemeineren Anfragen sehr schnell sehr groß und unübersichtlich werden kann. Trotzdem ist es auch heute noch die mit Abstand am weitesten verbreitete Methode.

Repräsentation von Inhalten

Nimmt man den Anspruch, nach Inhalten von Dokumenten zu suchen, ernst, ist das Boolesche Retrieval natürlich wenig überzeugend. Ein Dokument wird dabei im wesentlichen wie ein Sack voll Wörter behandelt, der vielleicht noch ein paar extra Taschen für

Titel, Autor und Stichwörter hat. Dabei gelten verschiedene Formen eines Wortes bereits als verschiedene Terme, Synonyme oder Unterbegriffe werden nicht erfaßt. Man weiß nur, welche Wörter im Dokument bzw. in seinen Teilen vorkommen, nicht aber, was für eine syntaktische oder semantische Rolle sie darin spielen.

Damit ist auch schon eines der Kernprobleme des Information Retrieval benannt: *Wie lassen sich Inhalte im Rechner so darstellen, daß sie automatisch verglichen werden können?*

Während das Boolesche Retrieval Wörter nur als Zeichenketten betrachtet, versucht die Forschung zur künstlichen Intelligenz (KI), Sprache automatisch zu "verstehen", d. h. zum Beispiel aus einem geschriebenen Satz durch einen Rechner die Schlüsse zu ziehen, die auch Menschen beim Lesen des Satzes ziehen würden. Die KI-Forschung zeigt, daß solche Systeme auch schon für kleine Themengebiete und einzelne Sätze extrem aufwendig sind: Es muß sehr viel "Alltagswissen" eingebaut werden, damit einfache Schlußfolgerungen gezogen werden können, wie z. B. zu erkennen, worauf sich ein "es" "er" oder "dieses" in einem Satz bezieht. Das zeigt, daß Sprache – insbesondere Alltagssprache – im allgemeinen vieldeutig ist und sich der "Sinn" eines Satzes häufig erst aus dem Kontext ergibt, in dem er verwendet wird.

Diese Beobachtung verweist auf ein weiteres Problem des IR: *Der Informationsbedarf von Anfragenden ist häufig vage. Muß er in einer formalen Anfragesprache ausgedrückt werden, verstärkt sich diese Unsicherheit noch.*

In den letzten 40 Jahren sind eine Reihe von pragmatischen Ansätzen entwickelt worden, um IR-Verfahren zu verbessern:

Wortstammreduktion

Um Wörter nicht nur als Zeichenketten zu vergleichen, sondern nach ihrer "Bedeutung", kann man als ersten Schritt nicht die spezifische Form, in der sie auftreten, verwenden, sondern ihre Wortstämme. Im Englischen ist es verhältnismäßig einfach, mit einer Reihe von Regeln, die Endungen entfernen, Wörter automatisch auf ihren Stamm zurückzuführen. Im Deutschen mit seinen vielen Wortformen, die auch den Stamm verändern, ist das erheblich aufwendiger, bzw. weniger erfolgreich.

Thesauri

Eine andere Möglichkeit, Wörter nach Bedeutungen zu strukturieren, ist die Verwendung von Thesauri, also Zusammenstellungen von Begriffen, die nach ihrer Bedeutung als Oberbegriffe, Spezialisierungen oder verwandte Begriffe geordnet sind. Terme aus diesem fest vorgegebenen und strukturierten Vokabular können Dokumenten zur Charakterisierung ihres Inhalts als sog. Indexterme zugeordnet werden (Indexierung). Eine Anfrage kann breiter gemacht werden, indem die Dokumente statt nach einem Thesaurusbegriff nach dessen Oberbegriff und *allen* seinen Unterbegriffen durchsucht werden. Sie kann eingeschränkt werden, indem ein spezifischerer Begriff gewählt wird.

Die manuelle Konstruktion und Pflege eines Thesaurus ist aufwendig; auch die korrekte Nutzung zur Indexierung von Dokumenten und zur Formulierung von Anfragen setzt genaue Kenntnisse sowohl des Thesaurus als auch des Fachgebietes voraus. Die Nutzung bleibt damit im wesentlichen auf spezialisierte und professionell gepflegte Informationsdienste beschränkt.

Man kann einen Thesaurus auch automatisch berechnen, indem man in großen Dokumentsammlungen aus dem gemeinsamen Auftreten von Wörtern in Dokumenten assoziative Beziehungen zwischen ihnen ableitet. Solche Thesauri weisen nicht die strenge Struktur auf, die "von Hand" konstruierte Thesauri auszeichnen, sie sind aber einfacher und schneller zu erzeugen. Daher können sie aktueller sein und auch für spezifische Fachgebiete bereitgestellt werden.

Das Vektorraummodell

Neben diesen Versuchen, von Wörtern zu Begriffen oder "Konzepten" überzugehen, kann man zur genaueren Spezifizierung von Informationsinhalten die Terme, die einen Informationsinhalt charakterisieren, gewichten. Das sog. Vektorraummodell stellt jedes Dokument als eine Liste von gewichteten Termen dar. Die Terme und ihre Gewichte können dem Dokument von Hand oder automatisch aufgrund von Häufigkeitsüberlegungen zugewiesen werden. Bei automatischer Gewichtung werden zum einen Terme, die in einem Dokument häufig vorkommen, stärker gewichtet; zum anderen werden Terme, die in vielen Dokumenten auftreten, schwächer gewichtet: Man geht davon aus, daß diese Terme wenig geeignet sind, einen spezifischen Inhalt zu beschreiben. Um eine Anfrage an ein Vektorraumsystem zu stellen, werden Suchterme eingegeben, die auch wieder gewichtet werden können. Das Retrieval-System berechnet für jedes Dokument einen Ähnlichkeitswert zur Anfrage, indem es für jeden Suchterm, der auch im Dokument auftritt, die beiden Gewichte des Terms miteinander multipliziert und diese Werte für alle Suchterme aufaddiert. Die Dokumente werden dann nach diesem Ähnlichkeitswert geordnet ausgegeben, so daß die ähnlichsten (und damit hoffentlich wichtigsten) den Nutzenden zuerst präsentiert werden.

Relevance Feedback

Um den Informationsbedarf von Nutzenden genauer zu bestimmen, sind Relevance-Feedback-Verfahren entwickelt worden. Sie gehen zum einen davon aus, daß Nutzende am besten anhand von konkreten Dokumenten entscheiden können, was sie suchen; zum anderen davon, daß Dokumente mit ähnlichem Inhalt auch ähnlich repräsentiert sind. Deshalb wird Nutzenden zunächst eine Auswahl von Dokumenten präsentiert, aus denen sie die für sie interessantesten auswählen. Das System kann dann weitere Dokumente suchen, die den positiv bewerteten ähnlich sind. Dieses interaktive Verfahren wird fortgesetzt, bis die Nutzenden es beenden. Es kann zu einem Dialogverfahren ausgebaut werden, das die Entwicklung der Suche bei der Interaktion mit den Nutzenden berücksichtigt.

Multimedia Objekte

Zunehmend handelt es sich bei den Dokumenten in Sammlungen nicht mehr nur um Textdokumente, sondern um multimediale Dokumente, die auch Bilder, Grafiken, Ton oder Video enthalten können. Für diese nichttextuellen Medien gibt es bisher kaum universell verwendbare automatisierte Verfahren, mit denen Inhalte erschlossen oder verglichen werden können. Für spezielle Probleme, z. B. bei der Automatisierung, bei Fingerabdrücken oder teilweise auch bei Paßfotos, gibt es Bilderkennungsverfahren, die aber nur unter spezifischen Bedingungen eingesetzt werden können. Methoden, mit denen automatisch Inhalte von digitalen Bildern oder Tondokumenten erschlossen werden können, werden einen wichtigen Forschungsschwerpunkt der nächsten Jahre bilden.

Verteilte und heterogene Dokumentmengen

Bisher waren IR-Systeme im wesentlichen auf in sich geschlossene, zentral verwaltete und einheitlich strukturierte Dokumentsammlungen beschränkt, die von Fachleuten benutzt wurden. In den letzten Jahren hat sich mit wachsender weltweiter Vernetzung der Zugang zu digitalen Dokumenten enorm erweitert. Im World-Wide-Web z. B. lassen sich die unterschiedlichsten Dokumente zu den unterschiedlichsten Themen finden. Sie ändern sich häufig. Ansätze, um in dieser vielfältigen, dynamischen und heterogenen Datenmenge suchen zu können, gibt es in Form von Indexierungsprogrammen, die das Netz regelmäßig absuchen, aber nur wenig effektive Suchmöglichkeiten in invertierten Listen anbieten, oder in experimentellen Systemen für “verteilte digitale Bibliotheken” (distributed digital libraries), bei denen in verschiedenen Forschungseinrichtungen elektronische Dokumente zu spezifischen Themengebieten unter den selben Zugriffsformaten und –mechanismen gesucht und abgerufen werden können.

Insgesamt zeigt sich, daß die Effektivität von IR-Verfahren davon abhängt, auf welche Datensammlungen sie angewendet werden. Je strukturierter und homogener Sammlungen sind und je besser diese Strukturen von spezifischen IR-Systemen genutzt werden können, desto effektiver können IR-Verfahren sein. Den “GPS” (General Problem Solver) gibt es auch im Information Retrieval nicht.