economag.
Zitierquelle:
http://www.economag.de/magazin/2008/2/65+In+den+Tiefen+des+Webs+das+Richtige+finden

Betriebswirtschaftslehre

In den Tiefen des Webs das Richtige finden

Information Retrieval: Grundlagenforschung für Suchmaschinen

Suchmaschinen wie Google, Yahoo! oder Windows Live sind aus dem Alltag eines jeden kaum noch wegzudenken. Sie erschließen rund zehn bis zwanzig Milliarden Dokumente. Und das ist nur die Spitze des Eisberges. Im "Deep Web", das von Google & Co. nur in geringem Maße erfasst ist, schlummern weitere Milliarden Dokumente mit teilweise fachlich hoch spezialisiertem Wissen. Web-Suchmaschinen, Deep Web-Informationsdienste und firmeninterne Suchsysteme haben eines gemein: Sie alle benötigen den praktischen Einsatz von so genanntem Information Retrieval. Welche wissenschaftliche Disziplin und welche Aufgaben verbergen sich dahinter?

Universitätsprofessor Dr. Wolfgang G. Stock, Heinrich-Heine-Universität Düsseldorf

Retrievalsysteme helfen beim Suchen und Finden von Dokumenten (Stock, 2007). Je nach Dokumentform unterscheiden wir nach Systemen zum Textretrieval und solchen zum Retrieval nicht-textueller digitaler Dokumente. Letztere bearbeiten Audio-, Bild- und Filmdokumente; sie sind bei weitem noch nicht so gut entwickelt wie die Retrievalsysteme im Textbereich. Bei den Textdokumenten können wir nach strukturierten, schwach strukturierten und nicht-strukturierten Texten differenzieren. Die logische Organisation der einzelnen Daten erfolgt bei strukturierten Texten in Tabellenform (beim relationalen Datenbankmodell). Schwach strukturierte Texte sind alle Arten der publizierten und nicht-publizierten textuellen Dokumente von der privaten E-Mail über einen internen Bericht, einer Webseite und der wissenschaftlichen Veröffentlichung bis zur Patentschrift, die alle jeweils eine gewisse Struktur aufzeigen.

Im Rahmen der Wissensrepräsentation (Stock & Stock, 2008) können den Texten strukturierte Daten, so genannte Metadaten, hinzugefügt werden, etwa kontrollierte Terme zur Inhaltsrepräsentation und Autorenangaben mit Normeinträgen. Nicht-strukturierte Texte zeigen - wie das Wort bereits vermuten lässt - überhaupt keine Struktur; in der Realität kommen sie kaum vor. Das Gegenstandsgebiet des Information Retrieval wendet sich zwar

prinzipiell allen Dokumentformen zu, bevorzugt werden jedoch eindeutig schwach strukturierte Texte.

Die Retrievalforschung ist eine interdisziplinäre Angelegenheit, die sich vorrangig zwischen Informationswissenschaft (mit Bezug auf den Content/Inhalt der Dokumente) und Informatik (mit Bezug auf die technische Realisierung der Retrievalsysteme) abspielt (Baeza-Yates & Ribeiro-Neto, 1999; Ferber, 2003; Frakes & Baeza-Yates, Hg. 1992).

Information Retrieval hat zwei Hauptaufgaben: Erstens werden Terme in den Dokumenten sowie in den Suchanfragen derart mittels informationslinguistischer Verfahren bearbeitet, dass wir die zur Query passenden Dokumente überhaupt auffinden. Zweitens gilt es, mit Hilfe von Retrievalmodellen die gefundenen Dokumente in eine nach Relevanz sortierte Reihenfolge zu bringen.

Julia Roberts und das Plural-S

Zur Bearbeitung der in den Dokumenten und in der Anfrage vorgefundenen Terme setzen Retrievalsysteme Natural Language Processing oder Informationslinguistik ein (Abbildung 1). Nach der Erkennung der Schrift der Zeichen (arabisch, lateinisch, kyrillisch usw.) wird die Sprache des Textes identifiziert. Haben wir es mit WWW-Seiten zu tun, erfolgt eine Trennung des Textes von Layoutelementen und Navigationslinks, wobei alle vorhandenen Strukturinformationen zu identifizieren sind.

Nun stehen wir an einem Kreuzungspunkt, einer Grundsatzentscheidung für das weitere Vorgehen: Wir können erstens den Text in n-Gramme (Zeichenfolgen mit jeweils n Zeichen) oder zweitens in Worte zerlegen ("parsen"). Eine Variante der n-Gramm-Methode verfolgt bei Sprachen mit stark unregelmäßig flektierenden Wortformen zunächst den Weg über die Worte bis zu den Grundformen, um danach die n-Gramme zu separieren. Worte werden durch Trennzeichen (unter anderem Leer- und Satzzeichen) erkannt.

Da Stoppworte selten zum Thema eines Dokumentes etwas beitragen, werden diese anhand einer vorgegebenen Liste markiert und - soweit vom Nutzer nicht ausdrücklich anders gewünscht - von der Suche ausgeschlossen. Es schließt sich eine Erkennung und (im Dialog mit dem Nutzer abzuarbeitende) Korrektur

etwaiger Eingabefehler an. Werden im Dokument Eigennamen (Julia Roberts - Heinrich-Heine-Universität Düsseldorf - Henkel KGaA) thematisiert, wird der Name als Ganzes indexiert und der betreffenden Person oder Institution zugeordnet. Dies geschieht vor der Grundformbildung, um Irrläufer auszuschließen (etwa das "s" bei Julia Roberts für ein Plural-S zu halten und abzuschneiden).

Java als Programmiersprache, als Insel oder als Kaffee

Eine zentrale Stellung in der Informationslinguistik nimmt die morphologische Analyse ein. Hier können wir - je nach Sprache - entweder Suffixe abtrennen und Stammformen bilden (zum Beispiel retrieval auf retriev) oder mit Hilfe von Wörterbüchern oder Regeln die jeweiligen Grundformen (Lexeme) ermitteln (und dann auf retrieve abbilden). Homonyme (Java als Programmiersprache, als Insel oder als Kaffee) werden getrennt und Synonyme (Samstag - Sonnabend) zusammengefasst. Komposita, also Worte mit mehreren Bestandteilen wie im Deutschen „juristische Person" oder im Englischen beispielsweise „soft ice" müssen als genau ein Begriff aufgefasst und indexiert werden. Im umgekehrten Fall sind bei der Kompositazerlegung (Wellensittichfutter auf Wellensittich, Sittich, Sittichfutter und Futter, aber nicht auf Welle) die sinnvollen Bestandteile zusätzlich zum zusammengesetzten Begriff einzeln in die invertierte Datei aufzunehmen.

Besonders verzwickt gestaltet sich dies bei Worten, die sich in unterschiedliche Begriffe zerlegen lassen (wie Staubecken: Staub-Ecken oder Stau-Becken?). Je nach Sprache können nach diesem Arbeitsschritt weitere sprachspezifische Aspekte behandelt werden (wie zum Beispiel im Deutschen die Bindestrichauflösung: bei Film- und Fernsehwirtschaft den Term Filmwirtschaft erkennen).

Im folgenden Schritt geht es in das Umfeld der bereits identifizierten Begriffe: einmal zum semantischen Umfeld (zum Beispiel zu Unter- und Oberbegriffen), zum andern zum Umfeld nach statistischer Ähnlichkeit (erhalten durch Analysen gemeinsamen Vorkommens zweier Terme in den Dokumenten). Für ein Unternehmen, das eine eigene Sprache entwickelt hat, ist es wünschenswert, diesen Unternehmensjargon auch bei der Suche verwenden zu können.

Die Sprache des Unternehmens

Hierzu bedarf es (zum Teil erheblicher) Vorarbeiten, die jeweilige hausspezifische Terminologie in einer Wissensordnung (einer Nomenklatur, einem Klassifikationssystem, einem Thesaurus oder einer Ontologie) abzubilden. Es ist im Zuge von Entwicklungen des "Web 2.0" oder "Unternehmen 2.0" derzeit in der Diskussion, die Mitarbeiter mittels einer Folksonomy Dokumente "taggen" zu lassen, um so den authentischen Sprachgebrauch einer Community zu berücksichtigen (Peters & Stock, 2007).

Multilinguale Recherche

Ist eine mehrsprachige Recherche gewünscht (zum Beispiel Suchargumente in deutsch, aber Treffer in deutsch, englisch und russisch), müssen Algorithmen multilingualen Retrievals (unter anderem automatische Übersetzung) vorliegen. Im zunächst letzten Arbeitsschritt gilt es, Anaphora (zum Beispiel Pronomen)

und Ellipsen (unvollständige Ausdrücke) aufzulösen. In den zwei Sätzen „Die Schmidt & Meyer GmbH produzierte Metallröhren. Sie steht vor dem Konkurs." muss das „sie" dem Firmennamen Schmidt & Meyer GmbH zugeordnet werden.

An einigen Stellen in unserem informationslinguistischen Ablaufplan wird es nötig sein, mit dem Nutzer in einen Dialog zu treten, so etwa bei den Eingabefehlern, bei den Homonymen, ggf. beim semantischen und statistischen Umfeld sowie bei der Übersetzung. Ohne eine solche Dialogkomponente dürfte ein optimales Retrieval nicht möglich sein, es sei denn, ein Nutzer identifiziert sich stets beim System, und dieses erfasst die Eigenschaften und Vorlieben des speziellen Nutzers. Wenn ein Nutzer zum Beispiel häufig nach Indonesien, Bali oder Jakarta gefragt hat und nun aktuell Java eingibt, ist die Wahrscheinlichkeit groß, dass die indonesische Insel und nicht die Programmiersprache gemeint ist.

Das finden, was gesucht wird - (k)eine Kunst

Ziel der Informationslinguistik ist, möglichst umfassend und zugleich genau diejenigen Texte zu finden, die der Intention des Nutzers entsprechen. Gemäß der Maxime " „Find what I mean, not what I say" " (Feldman, 2000) haben wir bewusst den Term "Intention" gewählt: Es geht nicht (nur) darum, welche Zeichen der Nutzer als Suchargument eingibt, es geht letztlich darum, das auszugeben, was er wirklich braucht.

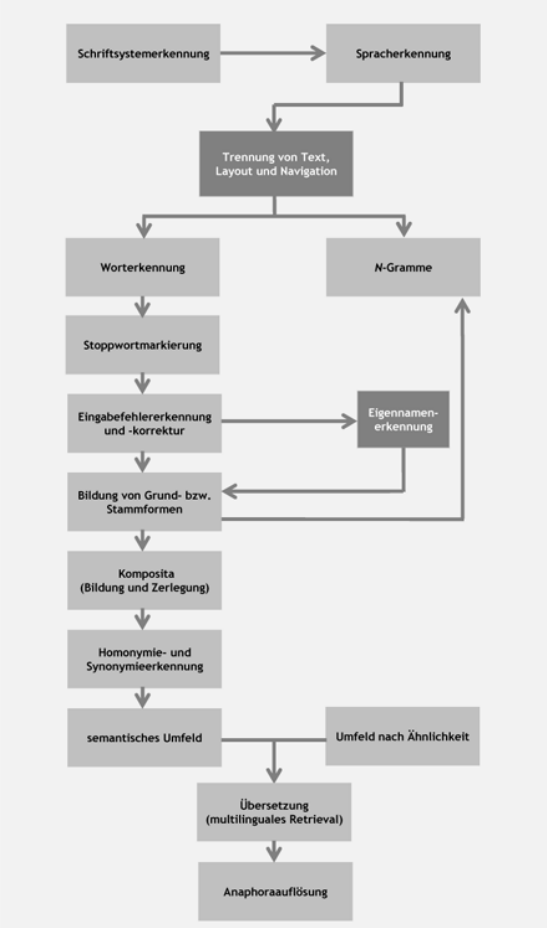

Abb. 1: Aufgaben der Informationslinguistik

Wichtiges sollte oben stehen

Informationswissenschaft und Informatik kennen mehrere Vorgehensweisen, Dokumente und deren Retrievalprozesse zu modellieren (Abbildung 2). Wir beschränken uns hier auf eine Aufzählung der gängigsten Ansätze.

Vor der Zeit des World Wide Web verfügt die Informationswissenschaft über folgende grundlegende Modelle: die Boolesche Logik, das Vektorraummodell sowie das probabilistische Modell. Außer dem Booleschen Modell erfüllen alle Ansätze die Aufgabe, die (mittels informationslinguistischer Arbeitsschritte) aufgefundenen Dokumente in eine Rangordnung nach Relevanz zu bringen. Mit dem Aufkommen des WWW wird die Menge der Retrievalmodelle größer (Lewandowski, 2005). Da die Hypertextdokumente untereinander verlinkt sind, ist deren Stellung innerhalb des Hyperspace im Rahmen der Linktopologie bestimmbar.

Die Boolesche Logik geht auf den englischen Mathematiker und Logiker George Boole zurück, der in seinem Werk "The Laws of Thought" (1854) eine binäre Sicht auf Wahrheitswerte (0 und 1) verbunden mit drei Funktionen UND, ODER und UND NICHT fundiert. Das starre Korsett Boolescher Modelle erfordert stets den Funktoreneinsatz; ist in einem Dokument das Suchargument vorhanden, wird es als Treffer ausgegeben, ist es nicht vorhanden, kommt es nicht zur Anzeige.

Null-Eins-Entscheidungen durch Gewichte ergänzen

Da wir grundsätzlich mit diesen Null-Eins-Entscheidungen arbeiten, ist eine Sortierung nach Relevanz prinzipiell unmöglich. In erweiterten Booleschen Modellen versucht man, diesem Missstand durch Gewichtungswerte abzuhelfen. Die Booleschen Operatoren müssen dazu entsprechend uminterpretiert werden.

Eine automatische, maschinelle Indexierung erfordert eine quantitative Analyse der in den Dokumenten auftretenden Terme. Gemäß Hans-Peter Luhn (bereits 1957 formuliert) geschieht dies durch Statistik. Als zwei grundlegende Gewichtungsfaktoren erweisen sich die Auftretenshäufigkeit eines Wortes in einem Dokument sowie die Anzahl der Dokumente in einer Datenbank, in denen das Wort erscheint.

Der erste Faktor wird als WDF (within document frequency weight) bezeichnet und relativiert die Anzahl eines Wortes in einem gegebenen Text auf die Gesamtanzahl der Worte. Im Grundsatz gilt: Je häufiger ein Wort in einem Text vorkommt, je größer ist sein WDF. Der zweite Faktor relativiert das Gewicht eines Wortes an seinem Auftreten in der Gesamtdatenbank. Da er in der Konstruktion gegenläufig zum WDF arbeitet, bezeichnet man ihn als IDF (inverse document frequency weight).

Der IDF errechnet sich als Quotient aus der Gesamtzahl der Dokumente einer Datenbank und der Anzahl derjenigen Dokumente, in denen das Wort vorkommt. Hier gilt: Je häufiger Dokumente mit dem Wort in einer Datenbank vorkommen, desto kleiner ist der IDF des Wortes. Die Gewichtung eines Terms in einem Dokument berechnet man als Produkt aus WDF und IDF.

Im Vektorraummodell sind sowohl die Dokumente als auch die Suchanfragen als Vektoren repräsentiert, wobei der Raum durch die jeweils vorhandenen Worte in den Dokumenten (einschließlich der Suchanfrage) aufgespannt wird. Liegen n Worte vor, so arbeiten wir in einem n-dimensionalen Raum. Die Ähnlichkeit zwischen Anfrage und Dokumenten sowie zwischen Dokumenten untereinander wird durch den jeweiligen Winkel der Dokumentvektoren bestimmt. Je kleiner der Winkel (oder je größer der Cosinus) ist, desto höher taucht der gegebene Text im Ranking auf.

Das probabilistische Modell fragt nach der Wahrscheinlichkeit, mit der ein gegebenes Dokument auf eine Suchanfrage zutrifft. Das Modell geht vom Vorhandensein von Relevanzinformationen im Anschluss an eine Rückkopplungsschleife ("Relevance Feedback") aus: Ein Nutzer muss gewisse Dokumente als relevant und gewisse andere als nicht relevant markiert haben. Dies kann stellvertretend - als Pseudo-Relevance Feedback - auch die Maschine übernehmen. Mittels der Feedbackinformationen können über die Auftretenswahrscheinlichkeiten der Worte in den relevanten oder nicht relevanten Dokumenten neue Gewichtungswerte kreiert werden, die das Relevance Ranking steuern.

Miteinander verlinkt: Dokumente im WWW

Dokumente im World Wide Web sind untereinander verlinkt. Es ist somit möglich, das WWW als Raum zu betrachten, in dem die einzelnen Dokumente liegen. In diesem Raum existieren Dokumente, auf die viele andere verweisen, und es gibt Dokumente, die ihrerseits auf viele andere linken. Solche Zusammenhänge macht sich das linktopologische Modell zunutze. Im Rahmen dieses Modells arbeiten der Algorithmus von Jon M. Kleinberg sowie der PageRank nach Sergey Brin und Lawrence Page.

Im Kleinberg-Algorithmus werden Seiten nach ihrer Funktion als "Hub" (nach ihren ausgehenden Links) und als "Authority" (nach ihren eingehenden Links) ausgezeichnet (Kleinberg, 1999). Die Gewichtung ergibt sich daraus, inwieweit Hubs auf "gute" Authorities (das heißt solche mit vielen "guten" eingehenden Links) linken und umgekehrt, inwieweit Authorities von "guten" Hubs (also solche mit vielen "guten" ausgehenden Links) gelinkt werden.

Der bei Google eingesetzte PageRank berücksichtigt nur das Authority-Maß (Brin & Page, 1998). Ein Webdokument erhält seinen PageRank durch die Anzahl und das Gewicht der eingehenden Links, indem die PageRanks aller dieser Backlinks, jeweils dividiert durch die Anzahl der ausgehenden Links, addiert werden. In einem theoretischen Modell ist der PageRank einer Webseite diejenige Wahrscheinlichkeit, mit der ein ausschließlich nach Zufall Surfender diese Seite findet.

Auch im Netzwerkmodell geht es um die Verortung. Die Basis liegt in der Theorie sozialer Netzwerke und in der Theorie "kleiner Welten". Es lässt sich zeigen, dass soziale Systeme - und auch Subsysteme wie Wissenschaftlergemeinschaften oder das WWW - keineswegs gleich verteilt, sondern stark "geklumpt" sind, wobei "Abkürzungen" zwischen unterschiedlichen Klumpen vorkommen. Innerhalb der Cluster lassen sich zentrale Dokumente oder - etwa bei wissenschaftlichen Autoren - zentrale Namen ausmachen. Das Ausmaß der Zentralität ist berechenbar und eignet sich als Rankingkriterium.

Soweit bei Suchmaschinen Informationen über die Nutzungshäufigkeit gegebener Webdokumente (zum Beispiel durch Auslesen von Protokollen bei Anwendern einer Toolbar der Suchmaschine) vorliegen, eignen diese sich unter Umständen als Rankingkriterium. Angaben zum Nutzer (beispielsweise über seinen aktuellen Standort) sind hilfreich bei Suchen, die ein geographisches Argument beinhalten ("Wo liegt die nächste Pizzeria?"). So ergibt sich ein Relevance Ranking durch die Berechnung des Abstandes zwischen dem Nutzerstandort und den Standortangaben in den Treffern.

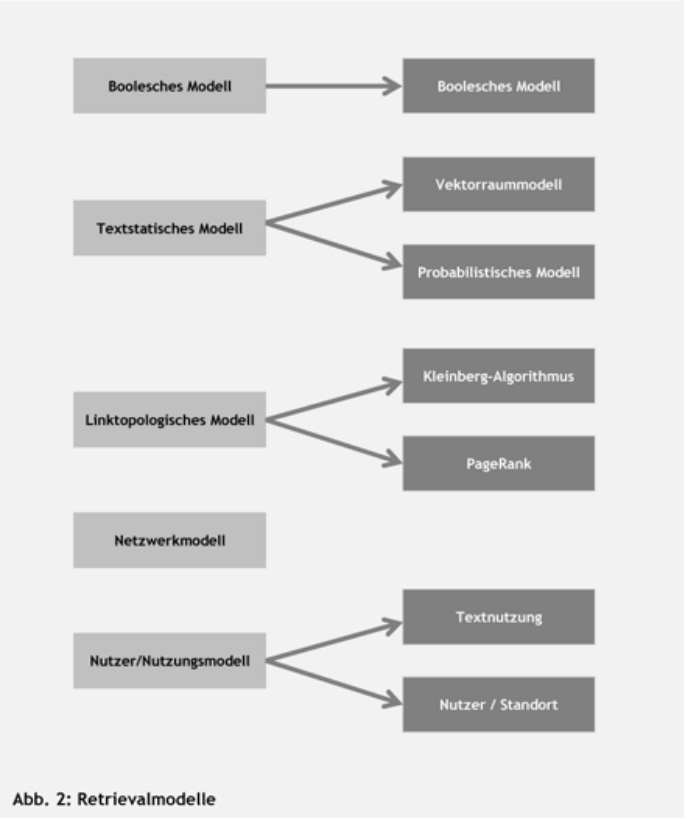

Abb. 2: Retrievalmodelle

Autor

Universitätsprofessor Dr. Wolfgang G. Stock ist Leiter der Abteilung für Informationswissenschaft der Heinrich-Heine-Universität Düsseldorf.

Literatur

Baeza-Yates, R. & Ribeiro-Neto, B. (1999): Modern Information Retrieval. New York: Addison-Wesley.
Brin, S. & Page, L. (1998): The anatomy of a large-scale hypertextual Web search engine. In: Computer Networks and ISDN Systems, 30, S. 107-117.
Feldman, S. (2000): Find what I mean, not what I say. Meaning based search tools. In: Online, 24(3), S. 49-56.
Ferber, R. (2003): Information Retrieval. Heidelberg: dpunkt.
Frakes, W.B. & Baeza-Yates, R., Hg. (1992): Information Retrieval. Data

Structures & Algorithms. Englewood Cliffs: Prentice Hall.
Kleinberg, J.M. (1999): Authoritative sources in a hyperlinked environment. In: Journal of the ACM, 46, S. 604-632.
Lewandowski, D. (2005): Web Information Retrieval. Technologien zur Informationssuche im Internet. Frankfurt/M.: DGI.
Peters, I. & Stock, W.G. (2007): Web 2.0 im Unternehmen. In: Wissensmanagement, 9(4), S. 22-25.
Stock, W.G. (2007): Information Retrieval. Informationen suchen und finden. München, Wien: Oldenbourg.
Stock, W.G. & Stock, M. (2008): Wissensrepräsentation. Informationen auswerten und bereitstellen. München, Wien: Oldenbourg.