Stand der Technik von WWW-Retrievalsystemen

Internet-Suchwerkzeuge im Vergleich (III)

Informationslinguistik und -statistik: AltaVista, FAST und Northern Light

Suchmaschinen im World Wide Web arbeiten automatisch: Sie spüren Dokumente auf, indexieren sie, halten die Datenbank (mehr oder minder) aktuell und bieten den Kunden Retrievaloberflächen an. In unserem Known-Item-Retrievaltest (Password 11/2000) schnitten - in dieser Reihenfolge - Google, AltaVista, Northern Light und FAST (All the Web) am besten ab. Die letzten drei Systeme arbeiten mit einer Kombination aus informationslinguistischen und informationsstatistischen Algorithmen, weshalb wir sie hier gemeinsam besprechen wollen. Im Zentrum unserer informationswissenschaftlichen Analysen stehen die "Highlights" der jeweiligen Suchwerkzeuge.

AltaVista: Der "Klassiker"

AltaVista nimmt seine Arbeit im Dezember 1995 in den in Palo Alto beheimateten Laboratorien der Digital Equipment Corp. auf. In kurzer Zeit entwickelt sich das Projekt zu einer der populärsten Suchmaschinen im World Wide Web. Nach der Übernahme von Digital Equipment kommt AltaVista Anfang 1999 in den Besitz von Compac Computer Corp. und wird als selbständiges Unternehmen ausgegliedert. Mitte 1999 verkauft Compac die Mehrheit seiner Anteile an AltaVista an den Internet-Mischkonzern CMGI Inc. in Andover. Zunächst verfolgt der neue Mehrheitseigentümer die Strategie, AltaVista als umfassenden Portalservice auszubauen. Mittels des Schwesterunternehmens First Up wird sogar ein kostenloser Internetzugang angeboten. Gegen Mitte 2000 ändert AltaVista wiederum die Strategie (vgl. Klostermeier 2000). Der Internetzugang wird ab Jahreswechsel 2000/2001 nur noch gegen Bezahlung offeriert; das Personal im Portalbereich teilweise entlassen; ein geplanter Börsengang ist bis jetzt nicht zustandegekommen. Mit der Einführung eines neuen Suchsystems ("Raging Search") sowie drei weiterer Retrieval-Varianten besinnt man sich auf seine alte Stärke: die Suchtechnik. Wir beschränken unsere Analyse von AltaVista auf genau diese Stärken und fragen erstens nach dem Algorithmus zum Einsammeln der Web-Dokumente sowie zweitens nach der Technik des automatischen Indexierens mit deren Kernstück, dem Ranking-Algorithmus. Kurz gehen wir drittens auf die vier Varianten der Suchoberflächen ein. Auch wenn AltaVista auf seiner Homepage verkündet, diese Algorithmen seien so geheim wie das Rezept von Coca-Cola, erhoffen wir uns doch von der Durchsicht der immerhin 38 Patente (siehe Tabelle 1), die unser Unternehmen hält, zumindest soviele Anhaltspunkte zu finden, dass die ver-

5,724,033	Method for encoding delta values
5,745,889	Method for parsing information of databases records using word-location pairs and metaword-location pairs
5,745,890	Sequential searching of a database index using constraints on word-location pairs
5,745,894	Method for generating and searching a range-based index of word-locations
5,745,898	Method for generating a compressed index of information of records of a database
5,745,899	Method for indexing information of a database
5,745,900	Method for indexing duplicate database records using a full-record fingerprint
5,765,149	Modified collection frequency ranking method
5,765,150	Method for statistically projecting the ranking of information
5,765,158	Method for sampling a compressed index to create a summarized index
5,765,168	Method for maintaining an index
5,787,435	Method for mapping an index of a database into an array of files
5,797,008	Memory storing an integrated index of database records
5,809,502	Object-oriented interface for an index
5,832,500	Method for searching an index
5,852,820	Method for optimizing entries for searching an index
5,864,863	Method for parsing, indexing and searching World-Wide-Web pages
5,909,677	Method for determining the resemblance of documents
5,914,679	Method for encoding delta values
5,915,251	Method and apparatus for generating and searching range-based index of word locations
5,963,954	Method for mapping an index of a database into an array of files
5,966,703	Technique for indexing information stored as a plurality of records
5,966,710	Method for searching an index
5,970,497	Method for indexing duplicate records of information in a database
5,974,455	System for adding new entry to Web page table upon receiving Web Page including link to another Web page not having corresponding entry in Web page table
5,974,481	Method for estimating the probability of collisions of fingerprints
6,005,503	Method for encoding and decoding a list of variable size integers to reduce branch mispredicts
6,016,493	Method for generating a compressed index of information of records of a database
6,021,409	Method for parsing, indexing and searching World-Wide-Web pages
6,032,196	System for adding new entry to Web page table upon receiving Web Page including link to another Web page not having corresponding entry in Web page table
6,047,286	Method for optimizing entries for searching an index
6,067,543	Object-oriented interface for an index
6,073,135	Connectivity server for locating linkage information between Web pages
6,078,923	Memory storing an integrated index of database records
6,105,019	Constrained searching of an index
6,112,203	Method for ranking documents in a hyperlinked environment using connectivity and selective content analysis
6,119,124	Method for clustering closely resembling objects
6,138,113	Method for identifying near duplicate pages in a hyperlinked database

Tabelle 1: US-Patente von AltaVista
Quellen: AltaVista (persönliche Mitteilung v. 12.12.2000); US Patent & Trademark Office

wendeten Algorithmen durchschaubar werden. Mit der doch beachtlichen Menge an Patenten kann AltaVista durchaus das Prädikat des Technologieführers bei den Suchmaschinen verliehen werden.

Die Quellen von AltaVista und ihr "Einsammeln": Scooter

Zentrales Element für den Erfolg von AltaVista ist seine Sammlung von WWW-Dokumenten. Sie fundiert auf dem Programm Scooter, das von Louis M. Monier entwickelt worden ist (vgl. Monier 2000). Grundidee von Scooter ist das Verfolgen von Links in bereits bekannten Dokumenten (siehe Abbildung 1). Für jede einzelne Seite werden deren Links abgearbeitet. Zunächst wird die URL des Links mit den bereits vorhandenen URLs verglichen. Ist sie noch nicht vorhanden, wird nach parallelen Sites (etwa Spiegelungen ein und derselben Site auf unterschiedlichen Servern) gesucht. Ist auch dieses negativ, wird ein Eintrag "nicht erfasst" notiert. Nach dem Bearbeiten aller Links kopiert Scooter das Dokument zur Erschließung auf den Server von AltaVista. In einem nächsten Schritt werden die Webseiten, die die Markierung "nicht erfasst" tragen - wie gerade beschrieben - analysiert.

Grundsätzlich nicht erreicht werden bei diesem Verfahren alle die Internetseiten, auf die an keiner anderen Stelle verwiesen wird. Frames werden in die einzelnen Teile ("panes") zerlegt, so dass das Gesamtbild einer Seite verloren geht. (Web-Master sind demnach gut beraten, zusätzlich zur Frames-Version eine Non-Frames-Version ihrer Seiten zu erstellen und diese bei AltaVista anzumelden.) In

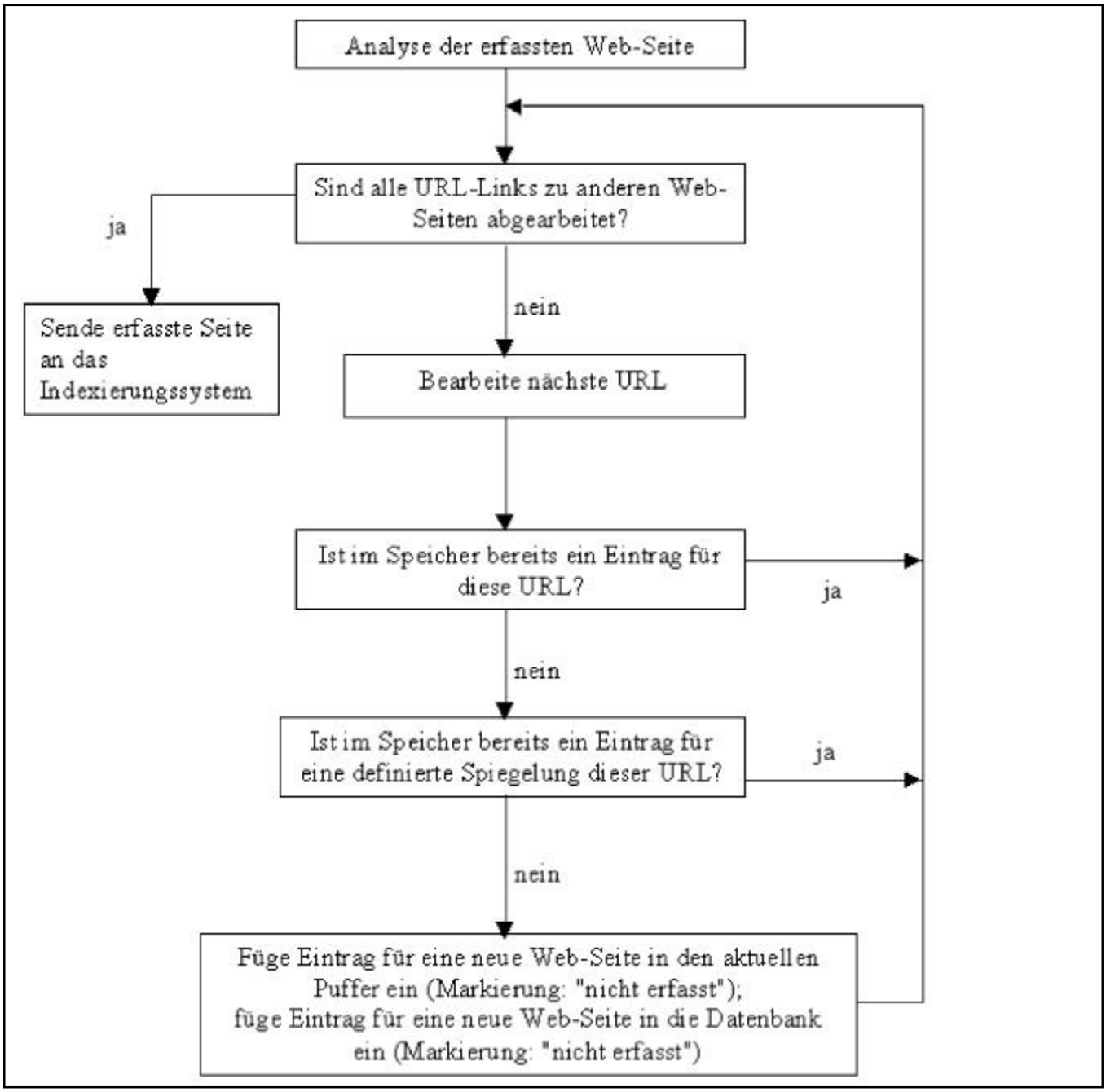


Abbildung 1: Algorithmus des Lokalisierens von Web-Seiten von "Scooter"
Quelle: Monier 2000, Fig. 4B

Graphiken eingebetteter Text kann nicht gelesen werden; natürlich auch keine "Texte" innerhalb von Audio- bzw. Videodateien (wohl aber in deren Beschreibung). Probleme bereiten Java Applets, XML-kodierte Seiten sowie pdf-Dateien, deren Inhalt von AltaVista nicht aufbereitet wird. URL-Links zu für AltaVista nicht indexierbare Dateien werden von Scooter ignoriert. Übergangen werden beim Einsammeln die Informationen aus dem Comments-Tag, also alles zwischen <!-- und -->. Innerhalb der Hierarchie einer Site verfolgt der Crawler nicht alle Ebenen nach unten; i.d.R. wird nicht tiefer als fünf Ebenen vorgedrungen. Die Stärke des Monier-Patentes liegt vor allem darin, dass die Schritte zum Einsammeln der Web-Seiten sehr schnell ablaufen, was übrigens den Namen "Scooter" begründet.

Es ist zusätzlich zum automatisch verlaufenden Scooter-Prozess für Web-Master jederzeit möglich, ihre Seiten manuell bei AltaVista anzumelden oder sie zu sperren. Beim Ausschluss von Sites hält sich AltaVista an den "Robot Exclusion Standard", der drei Varianten kennt (NOINDEX: alles wird ausgeschlossen; NOFOLLOW: alle in der Hierarchie der Site nach unten führende Links sind gesperrt; NOIMAGEINDEX: der Text wird erfasst, Bilder jedoch nicht).

In einer Untersuchung von Vivien Petras und Matthias Bank zeigt sich für AltaVista im Vergleich zu Hotbot (nach Lawrence / Giles die seinerzeit umfassendste Suchmaschine des Internet; siehe Password 11 / 2000, S. 24) eine größere Abdeckung des Web. Allerdings ergibt sich für AltaVista eine recht große Verzögerung des Inputs, die durchaus bei mehreren Monaten liegen kann (vgl. Petras / Bank 1998). Im aktuellen Vergleich der Datenbankgröße liegt AltaVista mit rund 350 Millionen Dokumenten hinter den beiden großen Systemen Google (über eine Milliarde URLs) und FAST / AllTheWeb (ca. 575 Millionen Seiten) zurück. (Der direkte Vergleich der Zahlen ist jedoch wenig aussagekräftig, da Scooter bei mehreren parallelen Seiten - sinnvollerweise - nur eine auswertet.)

Automatische Indexierung: Topics statt Suchargumente

Beim Abgleich einer Sucheingabe und der Datenbank kommen auf die automatisch vorgehenden Werkzeuge im Internet zwei Aufgaben zu. Erstens: Welche Datensätze werden ausgegeben? Und zweitens: In welcher Reihenfolge geschieht dies? Ein Problem für Suchmaschinen ist stets eine geringe Anzahl von Suchargumenten; Nutzeruntersuchungen berichten über durchschnittlich 1,5 Suchworte pro Suchfrage. Dies ist nicht genug, um Auswahl und Rangordnung von Dokumenten adäquat zu steuern. AltaVista nutzt die Linkstruktur des Web, um eine Suchfrage anzureichern. Das Patent von Krishna A. Bharat und Monika R. Henzinger (siehe Abbildung 2) geht bei der Zusammenstellung der Startmenge für das Retrieval vom Nachbarschaftsgraphen (N-Graphen) derjenigen Seiten aus, die beim exact match gefunden worden sind. Der Graph zeigt HTML-Dokumente als Knoten und Links als Kanten. In die Startmenge gehen alle Seiten ein, die auf die direkten Treffer via Link verweisen oder auf die von den direkten Treffern verwiesen wird. Die Terme in den Dokumenten der Startmenge müssen nicht mehr mit den (vom Nutzer eingegebenen) Suchargumenten übereinstimmen. Über einen Abgleich von Suchargument und URL-Termen sowie eine Analyse der Linkstruktur bildet AltaVista eine Teilmenge von Seiten, in die die jeweiligen Top-Treffer eingeordnet werden. Die Terme der ersten 30 Dokumente bilden die Basis für die Formulierung des Anfrage-Topic. Hierbei werden sowohl Stoppworte entfernt als auch Grundformen (durch Suffix-Stemming) gebildet. Der Anfrage-Topic Q gilt als Vergleichswert für das anschließende Relevance Ranking, indem jede Seite mit Q verglichen wird.

"Beschneiden" des Nachbarschaftsgraphen

Durch die Aufnahme der Seiten der ankommenden und der weiterführenden Links ist die Menge der Seiten im N-Graphen u.U. sehr groß geworden. Nun gilt es, die Menge zusammenzustreichen. Hierzu kommen das Vektorraummodell (von Gerald Salton) sowie das probabilistische Modell (von W.B.Croft und D.J.Harper) zum Einsatz. Berechnet wird die "Ähnlichkeit" jedes Dokuments mit dem Anfrage-Topic Q. Aspekte der Berechnungsformel sind (a) die Häufigkeit des Auftretens eines Terms (bzw. genauer: dessen Grundform) in Anfrage-Topic und Dokument und (b) die inverse Dokumenthäufigkeit (IDF) der Terme in einer konstanten Datenbank. Zur Berechnung der IDF-Werte analysierten die AltaVista-Forscher 400.000 Dokumente aus der Yahoo-Sammlung. Diese konstante Datenbank widerspricht durchaus den Intentionen des IDF, ändern sich doch in variablen Datenbanken die IDF-Werte mit jedem neuen Dokument. Hier müssen die theoretischen Erfordernisse offenbar hinter der praktischen Durchführbarkeit bei solch großen Datenbanken zurückstecken. Bei der Bewertung der Ähnlichkeit zwischen Q und jedem Dokument ist ein Schwellenwert definiert; alle Seiten, die diesen Wert nicht erreichen, werden gestrichen. Bharat und Henzinger verwenden an dieser Stelle den Begriff "beschneiden" - "to prune" -; wie ein Baum wird eine Menge von Web-Seiten im Nachbarschaftsgraphen auf die wesentlichen Äste zurückgeschnitten, wobei das "tote Holz" zurückbleibt.

Relevance Ranking nach Stellung in der Linkstruktur

Im letzten Schritt wird die Rangordnung innerhalb des zusammengestrichenen Graphen bestimmt. Aus der Linkstruktur der Dokumente errechnet der Algorithmus für jede Seite ein Maß für deren Eigenschaften als "Mittelpunkt" ("hub") und "Autorität" ("authority") im WWW. Ein guter "hub" verweist auf viele andere Seiten; eine gute "Autorität" kommt durch viele Links zustande, die auf diese Seite verweisen. Die Ergebnismenge wird nach der Höhe der jeweiligen Mittelpunkt- und Autoritätswerte sortiert ausgegeben.

Homepage - Power Search - Advanced Search - Raging Search

AltaVista bietet - neben Bildschirmen für Bilder, Videos usw. - vier Suchoberflächen zur Lokalisierung von Web-Dokumenten an. Die Suchfunktion der Homepage ist in diverse weitere Dienste des Portalservices eingebettet. Ein Button führt zur Advanced Search. Die Raging Search ist unter einer eigenen URL aufrufbar. Letztere ist Suchwerkzeug "pur", keine störenden

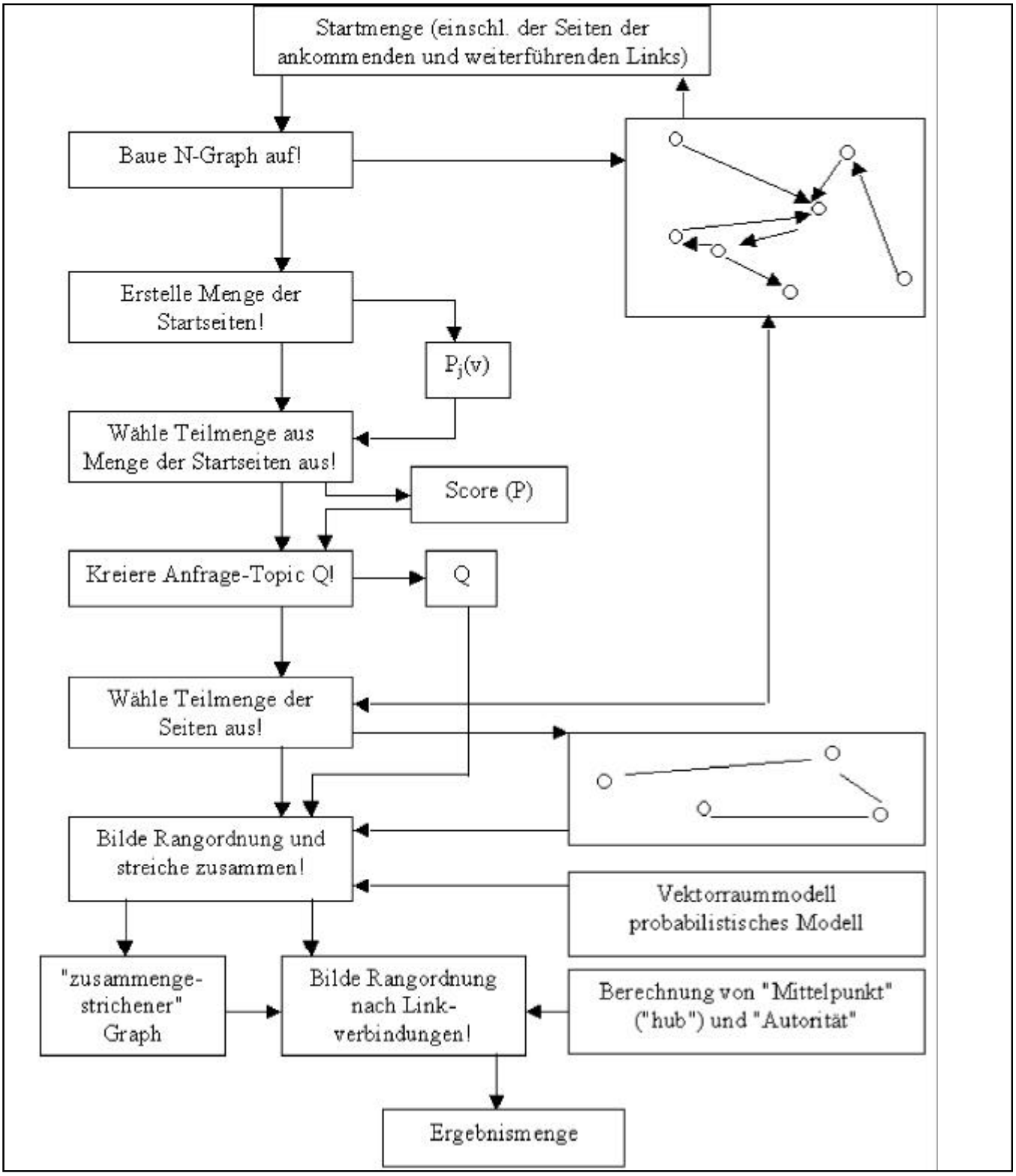


Abbildung 2: Ranking-Algorithmus von AltaVista
Quelle: Bharat / Henzinger 2000, Fig. 2 (vereinfacht)

Portaldienste wurden aufgenommen. Die Power Search wurde etwas versteckt; in der Homepage liegt der Link unter "More Search Options". Je nach Suchargument (nur ein Term, Phrase, mehrere Terme) unterscheiden sich die Suchalgorithmen der einzelnen Oberflächen. (Der oben beschriebene Algorithmus findet bei Raging Search und Homepage Search Einsatz.)

Tabelle 2 zeigt Gemeinsamkeiten und Unterschiede der vier Retrievaloberflächen. Die Homepagesuche ist in diverse weitere Funktionen eingebettet, darunter das klassifikatorische System von LookSmart sowie das Angebot von News usw. Power Search erlaubt eine feldspezifische Suche, Advanced Search erfordert eine mengentheoretisch formulierte Suchfrage sowie die Eingabe von Sortiertermen. (Diese Zusatzoptionen verhindern den Einsatz des Bharat-Henzinger-Patentes.) Übersetzungen und More like this-Suchen bieten alle vier Systeme. Einschränkungen auf eine

bestimmte Web-Site gestatten Home und Raging Search. Der RealNames-Link ist bei Raging Search nicht implementiert. Über "Extent your search" erreicht man bei den ersten drei Suchoptionen weitere Dienste wie Gelbe Seiten, eBay oder LookSmart. Bei Raging Search werden mit "Related Searches" Suchfragen angezeigt, die dem System gestellt worden sind und die Anregungen zur Reformulierung des Sucharguments geben. Interessant ist das Angebot einer benutzerdefinierten Ausgabe bei Raging Search; man wählt die Anzeige (oder Unterdrückung) folgender Felder bzw. Optionen:

- Hervorheben des Titels
- Beschreibung
- Hervorhebung der Beschreibung
- Datum der letzten Änderung
- URL
- Größe der Datei
- Sprache
- Übersetzungsoption
- Nachweise nur von der vorliegenden Web-Site
- assoziative Suche (more like this).

Nur eine knappe Bemerkung zum Übersetzungsprogramm: Mit dem Projekt "Babelfish" (Kenner von Douglas Adams "Per Anhalter durch die Galaxis" wissen, worum es geht) wird unter Einsatz der Software Systran eine rudimentäre Möglichkeit zur Wort-für-Wort-Übersetzung vorgehalten. Man bekommt - mit Glück - einen ersten Eindruck zum übersetzten Text, auf keinem Fall auch nur eine Art Rohübersetzung. Ruth Balkin stellt fest: "AltaVista is good for searching, but not translation" (Balkin 1999, 57)

FAST: All the Web - All the Time

Fast Search & Transfer (FAST) wird nach jahrelangen Vorarbeiten an der wissenschaftlich-technischen Universität Trondheim, Norges Teknisk-Naturvitenskapelige Universitet (NTNU), 1997 gegründet und schlägt neben der Niederlassung in Trondheim sein Hauptquartier in Oslo auf. Die norwegischen Teams widmen sich der Forschung und Entwicklung von Internetsuche und Datenübertragung. Ihre Forschungsschwerpunkte richten sich auf die Suchmaschinentechnik, Textretrieval im Web - einschließlich der Suchmaschine "AllTheWeb"-, korporative Suche (Intranetsystem), Bild- und Video-Kompression sowie die Suche nach WAP-gängigen Informationen. FAST knüpft schnell und ausgiebig internationale Kontakte. 1998 entsteht in den USA an der Ostküste in Massachusetts eine Tochtergesellschaft, dessen Tätigkeiten auf weltweite Geschäftsentwicklung, Marketing, Verkauf, Technik und Betrieb basieren. Für Verkauf und Betrieb an der Westküste ist ab 1999 eine Geschäftsstelle in Kalifornien zuständig. In Großbritannien erwirbt FAST 1998 Hercules und nennt den Inhalteanbieter in "Fast Web Media Ltd." um. Ein Ergebnis der Partnerschaft mit Dell offeriert die Suchmaschine AllTheWeb, die von Dell selbst als "The Next Generation Search Engine" gelobt wird: "The ability that FAST has to continue its image and video compression technology with its search technology makes it the only company in the world to be able to provide its customers with a complete multimedia solution".

Eigenlob und Expansionsdrang von FAST und Partner sind groß. Während im Juni 2000 Lycos bekanntgibt, gemeinsam mit FAST die größte Datenbank im Web geschaffen zu haben und für das kommende Jahr einen Ein -Milliarden URL-Katalog aufbauen möchten, preist sich FAST vier Monate später als den größten und schnellsten Suchmaschinenservice der Welt an: Nach der Dubletteneliminination bleiben von den derzeit durchsuchten 1,5 Milliarden URL ca. 575 Millionen Webdokumente übrig. 12 Millionen Suchfragen werden nach eigenen Angaben täglich bearbeitet, wobei die durchschnittliche Antwortzeit unter einer halben Sekunde liegen soll. Lycos und FAST dehnen ihre Partnerschaft auf den asiatischen Raum aus.

FAST führt die Leistungsfähigkeit des Systems bzl. Datenbankgröße, Suche und Transfer auf die Art der Architektur der Hardwarekomponenten zurück: den parallelen Einsatz von vielen kleinen Servern. Außer dieser Hardwarekomposition sehen wir zwei Stärken von FAST, zum einen die Suche nach nicht-textbasierten WWW-Informationen, zum andern die Techniken der Datenkompression beim Handling von Videos und Bildern.

An Dokumenttypen kann via FAST auf einen jeweils sehr großen Datenbestand zurückgegriffen werden:

- Webdokumente,
- Multimediadokumente (Audio, Video, Bilder),
- "Mobile Search" (WAP),
- FTP-Search (Dateien),
- MP3 (Musik).

Für die einzelnen Dokumenttypen liegen jeweils eigene Suchoberflächen vor. Abbildung 3 zeigt einen Ausschnitt der Suchmaske der "FAST Multimedia Search".

Zeichenbäume von Worten und Wortfolgen

Bei der Recherche nach Nicht-Text-Materialien sind zwangsläufig nur wenige Informationen vorhanden, die den Stoff der automatischen Indexierung bilden. Das Material muss demnach optimal ausgeschöpft werden. Grundidee ist die Zerle-

Funktion	**Home**	**Power Search**	**Advanced S.**	**Raging S.**
Klassifikation	LookSmart	---	---	---
Portalfunktionen	News etc.	---	---	---
Suchart	einfache Suche	einfache Suche, Feldsuche	Boolesche Suche, Sortierung	einfache Suche
Übersetzung	Babelfish	Babelfish	Babelfish	Babelfish
More like this	related pages	related pages	related pages	more like this
Site	more pages from this site	---	---	results from this site only
Homepagesuche	RealNames	RealNames	RealNames	---
Suchvorschlag	Extend your search	Extend your search	Extend your search	Related Searches

Tabelle 2: Suchfunktionalität der Retrievaloberflächen von AltaVista
Quelle: www.altavista.com; ragingsearch.altavista.com

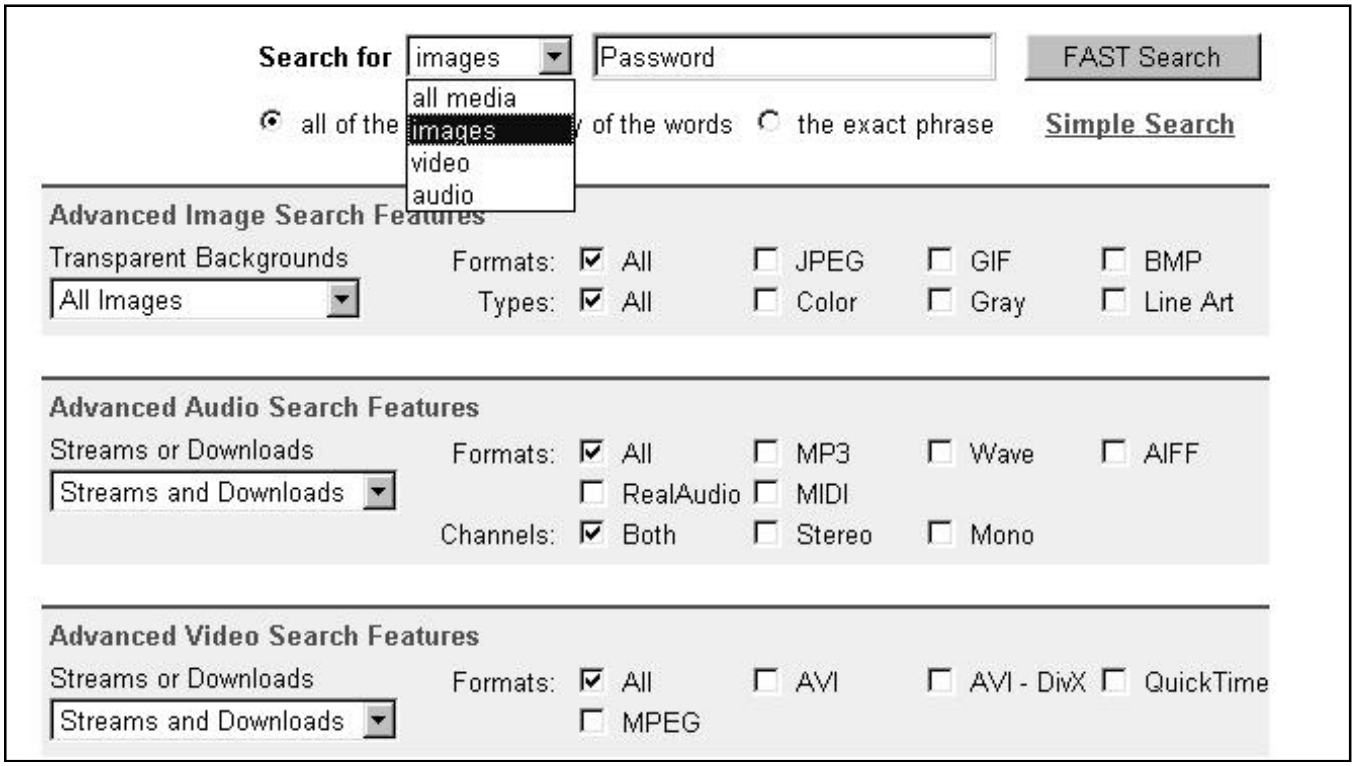


Abbildung 3: FAST Multimedia Search - Suchbildschirm

gung von Worten und Sätzen in Zeichenbäume, wobei Leerzeichen ausdrücklich als Zeichen und nicht (nur) als Begrenzer von Worten angesehen werden (vgl. Risvik 1998). Gesucht werden Worte (erkannt durch Begrenzungszeichen wie Leerzeichen oder Satzzeichen) sowie Wortfolgen (erkannt durch "sekundäre" Begrenzungszeichen wie etwa die Markierung des Dokumentendes). Innerhalb von Worten und Wortfolgen liegen Strings und - als Zerlegung von Strings - Substrings vor. Ein Beispiel einer Substringzerlegung bringt am Beispiel des Wortes "structure" Abbildung 4. Im entstehenden Zeichenbaum können alle in den Suchindex eingehenden Zeichenketten abgelesen werden. Unter "S" (dieses Zeichen kommt im String nur einmal vor) liegt das gesamte Wort, unter "T" (kommt zweimal vor) die jeweils dem "T" folgenden Substrings "tructure$" und "ture$", unter "R" "ructure$" und "re$" usw. Es ist klar, dass FAST bei diesem Verfahren grundsätzlich auf alle Zeichen zurückgreift und demnach keine Stopworte definieren kann.

So günstig dieses Verfahren ist, auch bei extrem kleinen Mengen an Textinformationen etwas zu finden, was über ein exact match hinausgeht, so ist es doch fehleranfällig. Wir fanden bei einer Bildsuche zu "Kieferorthopäde" (drei Bilder, davon die ersten zwei korrekt) ein Bild mit der Zeichenfolge "Stiefel" in der URL - nicht ganz das, was wir erwarten durften. Die Erklärung liegt im Substring "-iefe-", der den best match-"Treffer" bedingt hat.

Da bei Nicht-Text-Materialien große Datenmengen pro Video, Graphik usw. anfallen, das Retrieval und die Ergebnisdarstellung aber trotzdem schnell vonstatten gehen sollen, ist es vernünftig, mit Methoden der Datenkompression zu arbeiten. Die Verdichtung von Bild- und Videoinformationen ist einer der Forschungsschwerpunkte des Trondheimer Universitätsinstituts, dem einige der FAST-Entwickler entstammen ("Gruppe for Signalbehandling"); FAST nutzt die zum Patent angemeldete Technik von Arild Fuldseth (vgl. Fuldseth 1997).

Hybrid aus Web-Suchmaschine und Online-Archiv: Northern Light

1995 hat Kurt Müller bei Dataware Technologies die Idee, eine WWW-Suchmaschine zu entwickeln; 1997 beginnt das Ergebnis dieser Idee unter dem Namen "Northern Light" (dem Namen eines 1851 in Boston gebauten Schiffes) als eigenes Unternehmen in Cambridge, Massachusetts seine Arbeiten. Grundi-

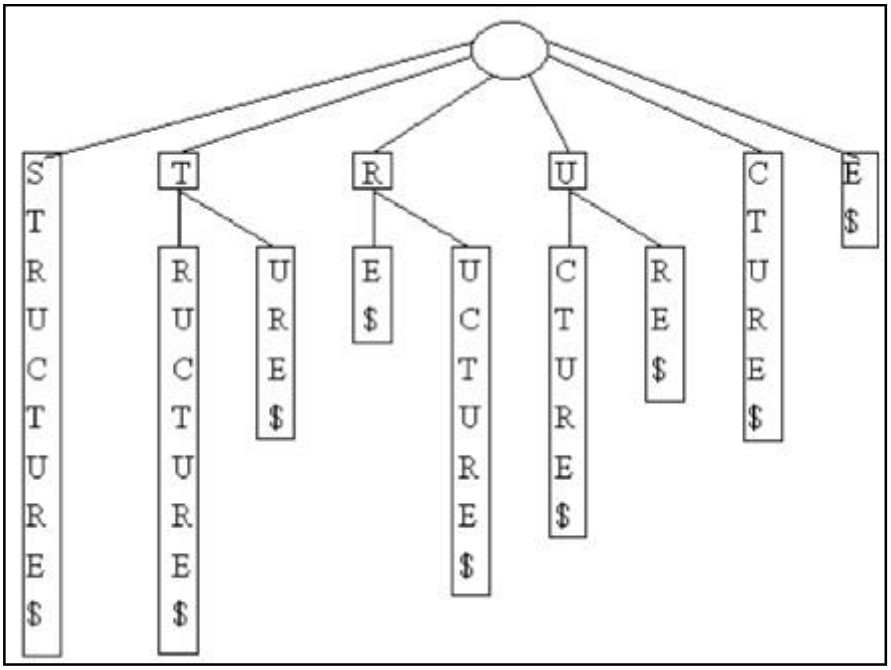
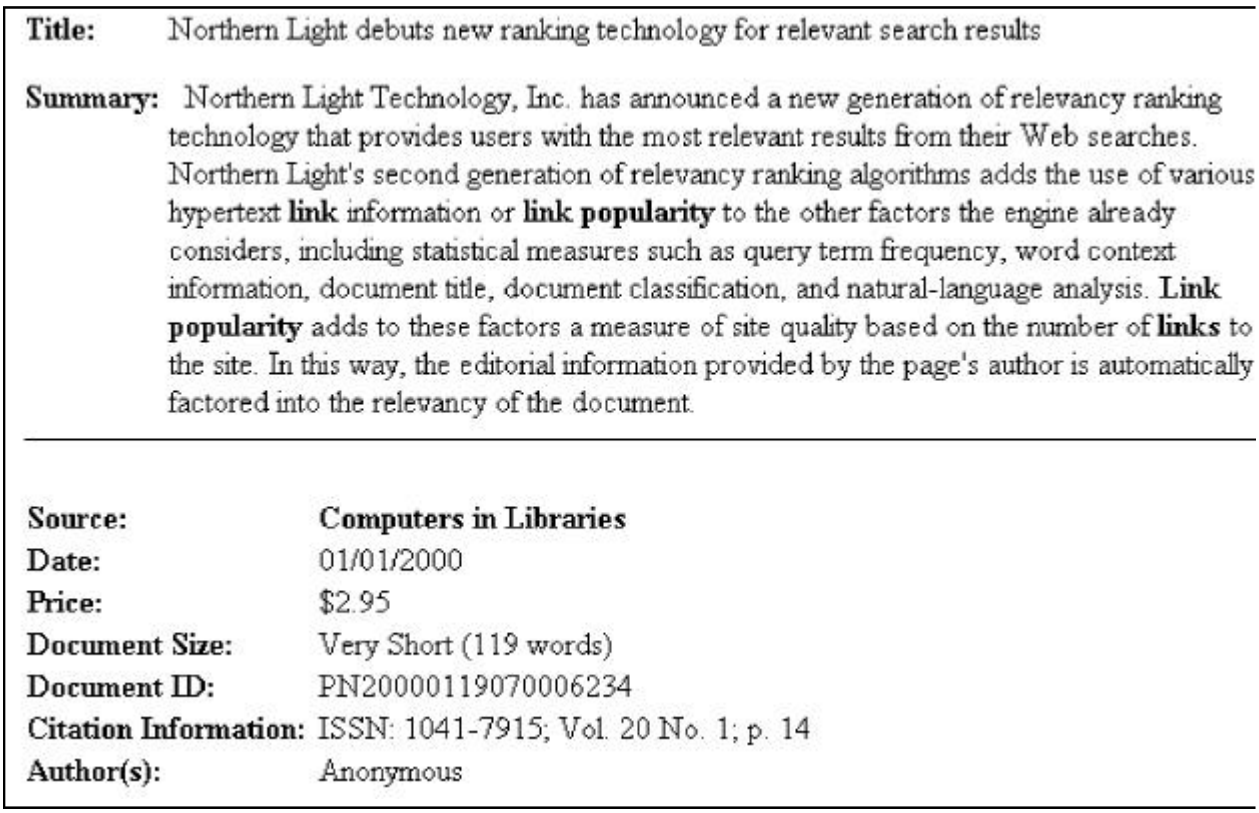


Abbildung 4: Zeichenbaum bei der Indexierung von FAST
Quelle: Risvik 1998, Fig. 1



Title: Northern Light debuts new ranking technology for relevant search results

Summary: Northern Light Technology, Inc. has announced a new generation of relevancy ranking technology that provides users with the most relevant results from their Web searches. Northern Light's second generation of relevancy ranking algorithms adds the use of various hypertext **link** information or **link popularity** to the other factors the engine already considers, including statistical measures such as query term frequency, word context information, document title, document classification, and natural-language analysis. **Link popularity** adds to these factors a measure of site quality based on the number of **links** to the site. In this way, the editorial information provided by the page's author is automatically factored into the relevancy of the document.

Source:	**Computers in Libraries**
Date:	01/01/2000
Price:	$2.95
Document Size:	Very Short (119 words)
Document ID:	PN20000119070006234
Citation Information:	ISSN: 1041-7915; Vol. 20 No. 1; p. 14
Author(s):	Anonymous

Abbildung 5: Kostenloser bibliographischer Nachweis bei Northern Light

dee von Northern Light ist die Vereinigung der beiden Welten der Informationen im World Wide Web und der Informationen in den kommerziellen Online-Archiven innerhalb einer Suchmaschine. Das aus dieser Vereinigung entstandene Hybridsystem hat derzeit eine Datenbasis von rund 315 Millionen Webseiten und über 50 Millionen Volltexten aus über 7.000 Quellen. Der jetzige CEO von Northern Light, David Seuss, kann 1999 u.a. Hewlett-Packard, Reuters und Times Mirror als Investoren gewinnen. Von Anfang an ist Northern Light ausschließlich ein Suchsystem, das nie an eine Ausweitung in Richtung Portal gedacht hat. "We're not a portal. The portals have the idea that the Internet is a lifestyle and entertainment experience ... We are a search engine. Our users want information" (Seuss in Barlas 1999).

Online-Archive der "klassischen" Art bieten einzelne Datenbanken mit deren spezifischen Werkzeugen (etwa einem datenbankspezifischen Thesaurus) jeweils einzeln an. Bei den technisch durchaus möglichen databankübergreifenden Suchen gehen die Vorteile der spezifischen Werkzeuge verloren. Endnutzerorientierte Systeme wie Profound von der Dialog Corp. oder Northern Light arbeiten mit **einem** Dokumentationswerkzeug für **alle** Bestände, unabhängig von deren Quelle. Profound setzt bekanntlich seinen InfoSort-Thesaurus ein (vgl. Password Nr. 10 / 1998, 22 ff.), Northern Light bevorzugt Clusteranalyse und ein Klassifikationssystem. Bei der automatischen Indexierung müssen sowohl Profound als auch Northern Light nicht nur informationslinguistische und -statistische Verfahren beherrschen, sondern zudem noch die Vergabe von Deskriptoren bzw. Klassenbezeichnungen steuern.

Da die Technik datenbankübergreifenden Retrievals mittels eines einzigen Werkzeugs auch im Intranet von Firmen Relevanz hat, bietet Northern Light sein System unter dem Namen "Single Point!" auch als Unternehmenslösung an. In Single Point! fließen die Informationen aus der Northern Light-Datenbank, Informationen aus weiteren Quellen (z.B. anderen externen Datenbanken) sowie aus unternehmensinternen Dokumenten zusammen.

Die beiden Oberflächen Northern-Light.com und NLSearch.com unterscheiden sich nur unwesentlich; letztere wird hauptsächlich Unternehmen empfohlen, die hier via Passwort oder IP-Adresse den Zugang ihrer Mitarbeiter kontrollieren können. Northern Light offeriert wie jede andere Suchmaschine retrospektive Recherchen sowie zusätzlich - und dies ist einmalig im Web - kostenlose individuelle Profildienste, geliefert via E-Mail. Die Trefferanzeige ist stets umsonst, bei Dokumenten aus kommerziellen Quellen führt ein Link zum bibliographischen Nachweis, der bei längeren Berichten ein Inhaltsverzeichnis bzw. - je nach Quelle - ein Abstract enthält (siehe Abbildung 5). Kosten entstehen erst beim Aufruf eines Volltextes bzw. einer konkreten Seite eines Reports. Um es zu betonen: Das, was bei "klassischen" Online-Archiven bereits Geld kostet - der bibliographische Nachweis, ist hier frei, erst die komplette Information - der Volltext - kostet etwas. Gefällt die Information nicht, wird dem unzufriedenen Kunden durch die "Money Back Guarantee" der bezahlte Betrag gutgeschrieben. (Northern Light hat offenbar gute Erfahrungen mit der Ehrlichkeit seiner Nutzer, kann man doch eine zur Kenntnis genommene Informationen nicht wirklich "zurückgeben".)

Automatische Klassifikation

Der technische Höhepunkt von Northern Light ist sein Vorgehen bei der automatischen Klassifizierung der Web- und Volltextdokumente. Im Dialog mit dem Nutzer erstellt das System für gefundene

Mengen von Dokumenten Teilmengen, die in Ordnern ("Custom Search Folders; siehe Abbildung 6) dargestellt werden. Mit jedem Schritt wird die Menge weiter eingeschränkt und das Suchergebnis entsprechend präzisiert. Für Greg Notess ist klar, "Northern Light's Custom Search Folders are a major advance in the realm of Web search engines" (Notess 1998, 34). Die Basiserfindung für diesen Fortschritt in der Suchtechnik stammt von Marc F. Krellenstein (vgl. Krellenstein 1999) und arbeitet mit der Clusteranalyse als Werkzeug zur Teilmengenbildung innerhalb komplexer Treffermengen.

Ausgang der Clusteranalyse ist ein Suchergebnis (egal welchem Umfangs) als Antwort auf eine Suchfrage. Wie bei anderen Suchmaschinen auch, ist die jeweilige Treffermenge nach Relevanz sortiert. Der Ranking-Algorithmus beinhaltet - laut Krellenstein-Patent - folgende Aspekte:

(1) Je mehr Worte zwischen Suchfrage und Dokument übereinstimmen, desto höher wird ein Dokument einsortiert,

(2) Falls in der Suchfrage eine Phrase enthalten ist, werden diejenigen Dokumente höher sortiert, die die genaue Phrase beinhalten,

(3) Falls in der Suchfrage Groß- und Kleinschreibung vorkommt, werden diejenigen Dokumente höher sortiert, die der gesuchten Schreibweise entsprechen,

(4) Falls ein Wort der Suchfrage im Titel des Dokuments vorkommt, so wird dieses höher sortiert,

(5) Falls ein Wort der Suchfrage im Abstract des Dokuments vorkommt, so wird dieses höher sortiert,

(6) Falls ein Wort der Suchfrage in den Keywords des Dokuments vorkommt, so wird dieses höher sortiert.

Angesichts des Erfolges, den Google beim Ranking nach Linkpopularität erhalten hat, modifiziert Northern Light 1999 seinen Kriterienkatalog (vgl. Feldman 1999). Eingefügt wird u.a.:

(7) Je mehr Links auf ein Dokument verweisen, desto höher wird es sortiert.

Da die Clusteranalyse bei großen Treffermengen zu viel Rechenzeit beanspruchen würde, arbeitet Northern Light nur mit Dokumentmengen bis maximal 200 Datensätzen. Hierbei wird geschickt das informetrische Verteilungsgesetz ausgenutzt. Informetrische Verteilungen sind nämlich extrem linksschief; die - im quantitativen Sinne - wichtigen Items liegen sehr wahrscheinlich am Anfang der Rangordnung. Northern Light sortiert die Treffer nach Relevanz und bearbeitet die ersten 200 der entstehenden Rangordnung. Ordnungskriterien der Clusteranalyse sind:

- Thema,
- Typ,
- Quelle,
- Sprache.

Formale Ordnungskriterien (wie die Sprache) gehen gleichberechtigt mit inhaltlichen Kriterien (wie das Thema) in unseren Kriterienkatalog ein. Die Einordnung in Klassen nach den beschriebenen Kriterien folgt zwei Regeln: Erstens sollen auf einer Hierarchieebene überschaubar viele Klassen (das Patent spricht von 7 +/- 2) entstehen, und zweitens muss jede Klasse mindestens 20% der Musterdokumente enthalten (bei 200 Mustern wären dies 40). Die entstehenden Klassen werden - analog zu Dokumenten - in eine Rangordnung gebracht und dem Nutzer als Ordner angezeigt. Wählt der Kunde einen Ordner zur weiteren Clusteranalyse aus, so wird das Verfahren wiederholt. Das Verfahren bricht ab, wenn weniger als rund 20 Datensätze in einer Treffermenge vorgefunden werden. Northern Light geht davon aus, dass

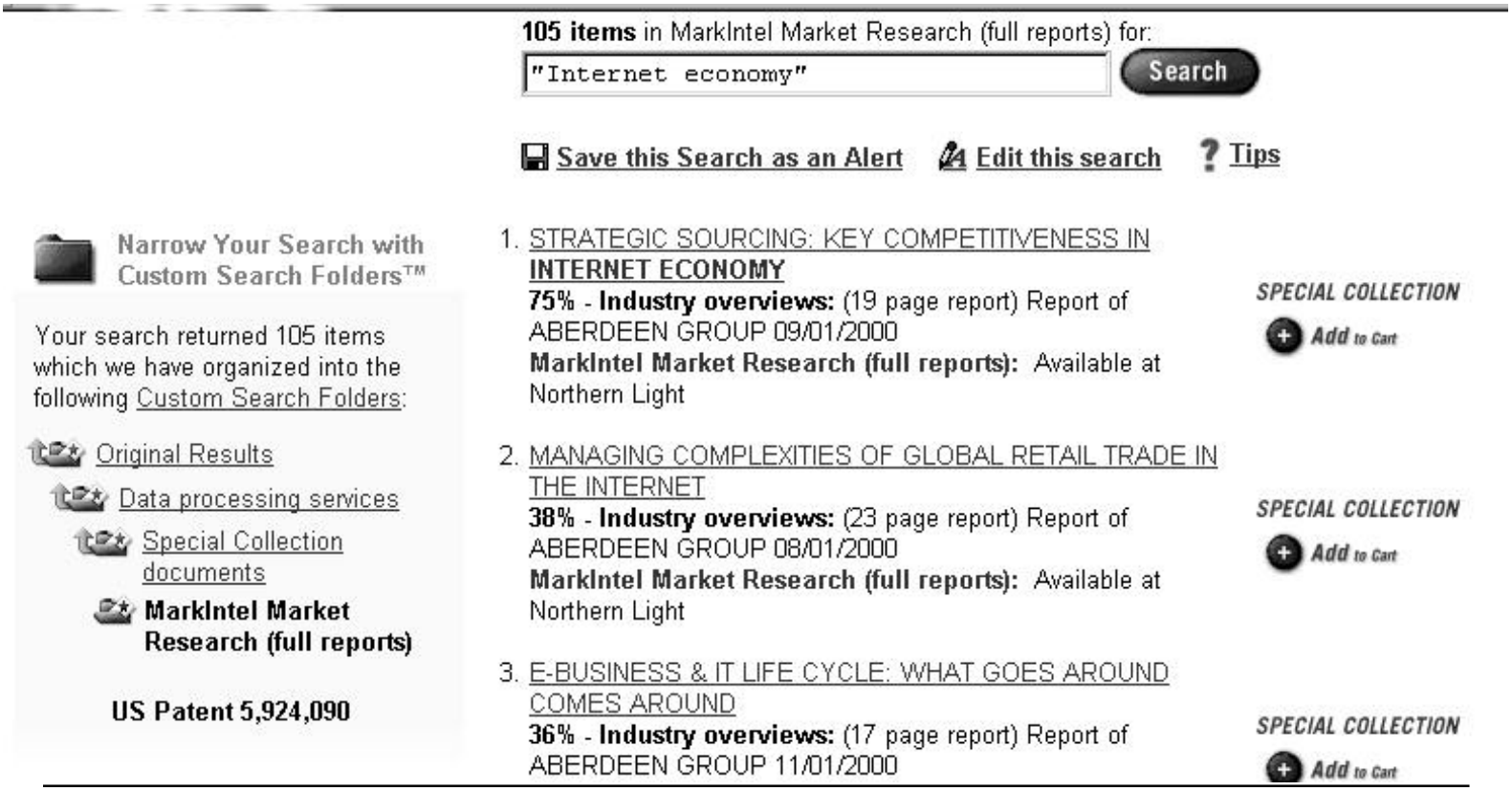


Abbildung 6: Clusteranalyse bei Northern Light: Anzeige der "Custom Search Folders"

Nutzer bis zu 20 Treffer intellektuell durcharbeiten.

Die Dokumente in den entstehenden Klassen werden analysiert und als automatisch generierte Suchfrage in das System zurückgespielt. Northern Light verwendet in diesem Retrievalschritt sowohl manuell vorgefertigte Suchargumente (z.B. 2000 thematische Suchformulierungen, 50 Typ- und sechs Sprachsuchfragen) als auch informationsstatistische Berechnungen der Gewichtung der Worte der gefundenen Dokumente. Eingesetzt wird das Produkt aus Termhäufigkeit im Dokument und der inversen Dokumenthäufigkeit in der Datenbank (IDF).

Die Anzeigen in den Ordnern (siehe Abbildung 6) sind stets empirisch gewonnene Taxonomien, die sich aus der ursprünglichen Suchfrage und der Analyse der jeweils gefundenen Dokumente ergeben; sie sind kein vorgegebenes Klassifikationssystem. Für jede Klasse (jeden Ordner, auf dem Bildschirm am linken Rand) wird die Gesamtmenge der jeweils gefundenen Dokumente nach Relevanz sortiert (auf dem Bildschirm rechts) angezeigt. Der Nutzer kann via "Edit this search" seine Suchfrage im Dialog modifizieren oder bei Gefallen über "Save this Search as an Alert" als Profil abspeichern.

Fazit

Die drei besprochenen Suchmaschinen AltaVista, FAST (AllTheWeb) und Northern Light verfügen dank elaborierter Einsammel- und Dublettenerkennungsverfahren über eine große Datenbasis. Bei der automatischen Indexierung setzen alle drei Suchwerkzeuge Kombinationen aus informationslinguistischen und informationsstatistischen Algorithmen ein, Northern Light darüber hinaus mit der Clusteranalyse ein Werkzeug der automatischen Klassifikation. Stärken von **AltaVista** sind die Behandlung der Suchfrage (durch Anreichern der Terme, Verfolgen der Linkstruktur und Beschneiden der Rohtreffermenge) sowie sein Relevance Ranking, das sowohl "klassische" informationswissenschaftliche Ergebnisse (Vektorraummodell, probabilistisches Modell) als auch die Spezifika von HTML-Dokumenten (Berechnung von "Mittelpunkt" und "Autorität" aus der Verteilung der Links) gebührend berücksichtigt. Betrachtet man die Patentaktivitäten der Suchmaschinen-Unternehmen, so dürfte AltaVista der Technologieführer in diesem Bereich sein. **FAST (AllTheWeb)** besticht neben seiner sehr großen Datenbank vor allem durch seine Zugriffsmöglichkeiten auf alle Arten von Nicht-Text-Dokumenten. Das Retrievalsystem ist so konzipiert, dass auch Multimediadokumente (selbst solche mit sehr wenig Textinformation) lokalisiert werden können. Datenverdichtungsverfahren ermöglichen einen schnellen Zugriff auf Bilder und Videos. **Northern Light** hat mit dem Einsatz der Clusteranalyse sein technisches Highlight. Hervorstechendes Alleinstellungsmerkmal ist jedoch der Hybridcharakter, der Webdokumente und Texte aus kommerziellen Online-Datenbanken unter eine Oberfläche zusammenführt. Das Preismodell scheint uns richtungweisend zu sein: Alle Webdokumente und alle bibliographischen Nachweise sind kostenlos, nur die "volle" Information, also der Zeitschriftenartikel oder der Report, wird bezahlt. Ausbaufähig dürfte der Umfang der kommerziellen Informationen sein; nur 7.000 Quellen sind - im Vergleich zu Lexis-Nexis oder Factiva - noch zu wenig. ■

Mechtild Stock & Wolfgang G. Stock

In Password 2/2001: Relevance Ranking nach Linkpopularität: Google

Literatur

AltaVista

Ruth Balkin: Babelfish. AltaVista's automatic translation program. - In: Database 22, Nr. 2 (1999), 56-57.

Krishna Asur Bharat; Monika R. Henzinger: Method for ranking documents in a hyperlinked environment using connectivity and selective content analysis / AltaVista Comp. - Patent Nr. US 6,112,203 vom 29. August 2000.

Johannes Klostermeier: Suchmaschine Altavista stoppt ihre kostenlosen Internetzugänge. - In: Net-Business Nr. 50 vom 11. Dezember 2000, 25.

Louis M. Monier: System for adding a new entry to a Web page table upon receiving a Web page including a link to another Web page not having a corresponding entry in the Web page table / Digital Equipment Corp. - Patent Nr. US 6,032,196 vom 29. Februar 2000.

Vivien Petras; Matthias Bank: Vergleich der Suchmaschinen AltaVista und HotBot bezüglich Treffermenge und Aktualität. - In: nfd. Information : Wissenschaft und Praxis 49 (1998), 453-458.

Richard Seltzer; Deborah S. Ray; Eric J. Ray: The AltaVista Search Revolution. - Berkeley: Osborne McGraw-Hill, 1997.

Chris Sherman: AltaVista gets a facelift, offers new search goodies. - In: Information Today 17, Nr. 8 (2000), 1+71.

FAST (All the Web)

Arild Fuldseth: A method in compression coding / FAST Search & Transfer ASA. - Patentanmeldung WO 99/05862. Priorität vom 28. Juli 1997.

Knut Magne Risvik: A search system and method for retrieval of data, and the use thereof in a search engine / FAST Search & Transfer ASA. - Patentanmeldung WO 00/03315 A2. Priorität vom 10. Juli 1998.

Northern Light

Pete Barlas: Northern Light guides Web's biggest engine. - In: Investor's Business Daily vom 12. Juli 1999.

Susan Feldman: Northern Light adds link popularity to its relevance ranking factors list. - In: Information Today 16, Nr. 11 (1999), 38.

Marc F. Krellenstein: Method and apparatus for searching a database of records / Northern Light Technology LLC. - Patent Nr. US 5,924,090 vom 13. Juli 1999.

Greg R. Notess: Northern Light: New search engine for the Web and full-text articles. - In: Database 21 (1998), Nr. 1, S. 32-37.

Sarah D. Scalet: See the Northern Light. Serious search capabilities for serious researchers. - In: Smart Computing in Plain English 6 (März 2000).