Stand der Technik von WWW-Retrievalsystemen

Internet-Suchwerkzeuge im Vergleich (IV) Relevance Ranking nach "Popularität" von Webseiten: Google

In unserem Retrievaltest von Suchwerkzeugen im World Wide Web (Password 11/2000) schnitt die Suchmaschine Google am besten ab. Im Vergleich zu anderen Search Engines setzt Google kaum auf Informationslinguistik, sondern auf Algorithmen, die sich aus den Besonderheiten der Web-Dokumente ableiten lassen. Kernstück der informationsstatistischen Technik ist das "PageRank"-Verfahren (benannt nach dem Entwickler Larry Page), das aus der Hypertextstruktur des Web die "Popularität" von Seiten anhand ihrer ein- und ausgehenden Links berechnet. Google besticht durch das Angebot intuitiv verstehbarer Suchbildschirme sowie durch einige sehr nützliche "Kleinigkeiten" wie die Angabe des Rangs einer Seite, Highlighting, Suchen in der Seite, Suchen innerhalb eines Suchergebnisses usw., alles verstaut in einer eigenen Befehlsleiste innerhalb des Browsers. Ähnlich wie RealNames bietet Google mit dem Produkt "AdWords" den Aufkauf von Suchtermen an.

Nach einer Reihe von nunmehr vier Password-Artikeln über Internet-Suchwerkzeugen im Vergleich wollen wir abschließend zu einer Bewertung kommen. Wie ist der Stand der Technik bei Directories und Search Engines aus informationswissenschaftlicher Sicht einzuschätzen? Werden die "typischen" Internetnutzer, die ja in der Regel keine Information Professionals sind, adäquat bedient? Und können auch Informationsfachleute von den Suchwerkzeugen profitieren?

Eine 1 mit 100 Nullen

Ähnlich wie das norwegische System FAST (AllTheWeb) entstammt Google einem akademischen Umfeld. Es ist aus dem Stanford WebBase-Projekt der Data Mining Group und der Digital Libraries Group der Stanford University hervorgegangen. Gründer sind der aus Michigan stammende Lawrence ("Larry") Page (heute der CEO) und der gebürtige Moskauer Sergey Brin (heute Präsident), beide Doktoranden der Computerwissenschaft in Stanford. Page entstammt dem Institut für Computerwissenschaft von Terry Winograd, wo er das Projekt leitet, aus dem sich Google entwickelt.

Google wird 1998 gegründet und hat seinen Sitz in Mountain View, Kalifornien. Google, Inc., ein privates Unternehmen, wird finanziell unterstützt durch Venture Capital-Unternehmen wie Kleiner Perkins Caufield & Byers und Sequoia Capital, durch die Stanford Universität sowie auch durch individuelle Investoren.

Der Name unseres Suchwerkzeugs leitet sich von "Googol" ab, eine von Edward Kasner und James Newman in ihrem Buch "Mathematics and the Imagination" 1940 eingeführte Bezeichnung für die Zahl zehn hoch einhundert, also eine 1, gefolgt von 100 Nullen. Ausgedacht hat sich die Bezeichnung der seinerzeit neunjährige Neffe Kasners. Als "Googolplex" wird in der Literatur die Zahl zehn hoch zehn hoch hundert (oder zehn hoch googol) genannt, mithin eine 1, gefolgt von googol Nullen. Die Assoziation der Google-Entwickler ist offenbar, dass diese Suchmaschine auf Masse aus ist, auf die größte Menge an Datensätzen bei allen Suchwerkzeugen. Anfang 2001 umfasst Google rund 1,3 Milliarden Datensätze und ist damit in der Tat das Web-Retrievalsystem mit den meisten nachgewiesenen Dokumenten. (Voll indexiert ist davon jedoch nur knapp die Hälfte; die anderen Nachweise sind ausschließlich durch Anchor-Texte in anderen Dokumenten inhaltlich ausgewertet.)

PageRank

Das Herzstück unseres Suchwerkzeugs beruht auf einem neuartigen Rankingverfahren, dem sog. "PageRank", wobei "Page" sich nicht etwa auf eine Web"page", sondern auf den Namen des Entwicklers Larry Page bezieht. Page veranschaulicht an einem fingierten Beispiel, wie die Gewichtung einer Webseite einigermaßen objektiv gemessen werden kann. Gesucht werde die Site der Stanford Universität. Angenommen, ein heimtückischer Webmaster, der andere Leue glauben lassen möchte, seine Site sei Stanford, kopiere die Stanfordhomepage auf seinen eigenen Rechner unter seiner URL. Am Text erkennt eine Suchmaschine nicht, dass es sich bei der Kopie nicht um die eigentliche Seite der Stanford Universität handelt. Erst die Linkstruktur des Web gibt einer Suchmaschine Anhaltspunkte, um die "richtige" Seite zu finden. "We look at what the Web thinks Stanford is, and it turns out there are plenty of clues sprinkled around. One is that lots of people link to the Stanford home page, and the people who link to the Stanford home page also tend to have a lot of people linking to them, so they're not just random pages. There's plenty of quality pages that link to the Stanford home page, and that's significant, too. So when we run a search, we look at all of the data, and since there are ten thousand links out there that all point to Stanford , we calculate that it's really Stanford University" (Page / Pemberton 2000, 42). Google misst die Seite als "richtige" Seite, weil - so Page - die ganze Welt denkt, dass es die richtige Seite sei.

Durch das PageRank-Verfahren wird sozusagen die "Popularität" jeder Webseite gemessen. Für mathematisch Interessierte seien hier die Grundgedanken des Verfahrens berichtet. Gegeben sei eine Webseite **A**, auf die via Links von anderen Webseiten **T1**, ...**Ti**, ... **Tn** verwiesen wird. In den entsprechenden Ankertexten auf **T1**, ..., **Tn** wird thematisch auf A Bezug genommen. Es wird ein Faktor d eingeführt, der zwischen 0 und 1 eingestellt werden kann. Google verwendet derzeit d = 0,85. Der Wert **C(Ti)** zählt die Menge der Links, die von der

Webseite **T1** zu anderen Seiten weiterführen. Der PageRank **PR** von **A** errechnet sich als Summe der Quotienten **PR / C** für alle **Ti**, die auf **A** verweisen. Für jede Seite wird also deren PageRank durch die Anzahl der weiterführenden Links dividiert. Unter Einschluss des Faktors **d** lautet die Formel zur Berechnung des PageRank **PR** von Seite **A** (vgl. Brin / Page 1998, 110):

PR(A) = (1 - d) + d [PR(T1) / C(T1) + ... + PR(Tn) / C(Tn)].

Die Berechnung der PageRanks erfolgt in einem iterativen Verfahren und bedarf der Linkanalyse aller Datensätze der Datenbank. Da die PageRanks aller Seiten eine Wahrscheinlichkeitsverteilung aufspannen, ist die Summe der PageRanks aller Seiten stets gleich 1. Die intuitive wahrscheinlichkeitstheoretische Deutung der Formel führt zum Modell des Benutzerverhaltens eines "Zufallssurfers", der rein zufällig Seiten auswählt bzw. zufällig Links verfolgt (aber nie "zurück" navigiert). Die Wahrscheinlichkeit, mit der unser Zufallssurfer eine Webseite besucht, ist deren PageRank. Der Faktor 1-d steht für die Wahrscheinlichkeit, nach der ein Zufallssurfer nicht den angegebenen Links folgt, sondern - wieder nach dem Zufallsprinzip - eine andere Seite besucht. Google hat d zur Zeit für alle Seiten auf 0,85 eingestellt; es wäre jedoch möglich, d für jede einzelne Seite oder für Gruppen von Seiten differenziert zu vergeben.

Neue noch unbekannte Webseiten, die man vielleicht intuitiv als qualitativ wertvoll einschätzen würde, erhalten bei Google sehr wahrscheinlich einen niedrigen PageRank-Wert, weil sie innerhalb der Linkstruktur im World Wide Web noch eine untergeordnete Rolle spielen: Nur wenige Seiten verweisen auf sie und darunter sind auch nur sehr wenige bereits populäre Seiten.

PageRank hat Ähnlichkeiten mit der Berechnung der Zitationsraten bei der Citation Analysis. Sie unterscheidet sich davon allerdings durch ihren rekursiven Charakter. Ein Link von Seite A zu Seite B wird von Google - wie bei der Zitationsanalyse - als eine Stimme von Seite A für Seite B gedeutet. Die Wichtigkeit einer Seite wird einerseits nach den abgegebenen Stimmen und aber auch zusätzlich - und dies unterscheidet Google von Zitationsanalysen - von der Wichtigkeit der Seiten, von der die Stimmen herkommen, bewertet. Nicht alle Links sind demnach vor Google gleich. Es macht einen Unterschied, ob eine Webseite einen Link von einer eher unbekannten Site oder beispielsweise von Yahoo! erhält. Page argumentiert, "Now the reason why Yahoo! is important is because lots of people link to Yahoo!, and lots of important people link to Yahoo!" (Page / Pemberton 2000, 43). Je größer die Datenbasis von Google ist, desto zutreffendere Ergebnisse bringt PageRank.

PageRank findet in unterschiedlichen Kontexten Einsatz:

- es ist ein Gewichtungswert für eine einzelne Webseite, unabhängig von Suchfragen
- es ist ein Sortierkriterium für Relevance Ranking von Suchergebnissen und
- es ist ein Ordnungshinweis für den Crawler, welche Seiten zur Indexierung aufgesucht werden (vgl. Cho / Garcia-Molina / Page 1998).

"Popularität" versus "Qualität"

Bei der Interpretation der PageRank-Messwerte sollte man nicht den Fehler begehen, die gemessene Popularität einer Webseite mit deren Qualität gleichzusetzen. Der grüne Balken verleitet einen Nutzer nämlich leicht zu solch einer Fehlannahme. Hier liegt dasselbe Problem wie bei der Zitationsanalyse vor. Eine hohe Zitationszahl oder Zitationsrate gilt auch bei der Citation Analysis nicht als Qualitätsfaktor, sondern lediglich als Hinweis auf eine besondere Stellung in der wissenschaftlichen Kommunikation. Eugene Garfield, der "Vater" der wissenschaftlichen Zitierindices, schreibt dazu: "Citation analysis is not a substitute or shortcut for critical thinking; it is, instead, a point of departure for those willing to explore the avenues to thorough evaluation" (Garfield 1986, 408).

Bei der Balkendarstellung wird der PageRank auf das Intervall null bis zehn (maximale Gewichtung) geeicht. Der Leser möge sein Qualitätsurteil über folgende Dokumente mit dem PageRank vergleichen:

- Yahoo! (www.yahoo.com): PageRank: **10**
- Star Wars Episode 1 (www.starwars.com / episode-i /): PageRank: **8**,
- Britney Spears (www.britneyspears.com / welcome.html): PageRank: **7**
- Kants Kritik der reinen Vernunft (www.gutenberg.aol.de / kant / krvb / krvb.htm): PageRank: **6**
- das Schematismuskapitel in der Kritik der reinen Vernunft: PageRank: **5**
- Höhlengleichnis in Platons "Staat" (www.netvip.com / users / piaries / dhoehle.htm): PageRank: **2**.

Dieses kleine Experiment dürfte verdeutlicht haben, dass "Qualität" nicht mittels PageRank operationalisierbar ist.

Dokumentationseinheit

Für Google haben Links eine besondere Bedeutung. Deshalb ist es verständlich, dass auch den Ankertexten großes Gewicht beigemessen wird. Umfang einer Dokumentationseinheit bei Google ist der gesamte Text einer Webseite sowie der Text aller Anker, die auf unseren Text verweisen. Die Ankertexte auf der Seite werden demgemäß doppelt zugeordnet, zu der Webseite, auf der sie aufscheinen, und zusätzlich zu der Webseite, auf die sie verweisen. Auf Abbildung 1 sehen wir den Umfang der Dokumentationseinheit der Internetseite A. Die Texte der Anker zu X und Y gelten als Teil von A, aber auch von X und Y, die - unserer Abbildung gemäß - noch gar nicht in der Datenbasis vorliegen. Teile der Dokumentationseinheit A sind auch die Ankertexte der Webseiten B, C und D, unter denen Links auf A liegen. Ankertexte können eine genauere Beschreibung der Webseite besitzen als die fragliche Webseite selbst. Ankertexte existieren auch für Webdokumente, die von textorientierten Suchmaschinen z.T. überhaupt nicht indexiert werden können wie z.B. Programmdateien, Bilder oder ganze Datenbasen. "We use anchor propagation mostly because anchor text can help provide better quality results" (Brin / Page 1998, 110).

In der Datenbank von Google liegen sowohl voll erschlossene Dokumentationseinheiten (über 600 Millionen Dokumente Anfang 2001) als auch Dokumentationseinheiten, von denen derzeit nur die Ankertexte (die in den 600.000 Dokumenten gefunden worden sind) die Indexierungsgrundlage bilden (über 700 Millionen Dokumente).

Suchfunktionalität

Wie sucht Google? Hier muss man zunächst notieren, was Google an erwarteter Suchfunktionalität nicht bringt. Nicht unterstützt werden: Jokerzeichen für Fragmentierungen, Groß- bzw. Kleinschreibung, Klammersetzung in der Suchfrage und automatische Wortstammbildung. Da nur exakt zeichengetreu gesucht wird, würden Tippfehler stets zu falschen Ergebnissen führen. Dieses Problem wird durch eine Tippfehleranalyse (bei Eingabe eines nicht bekannten Terms: "Did you mean: XXX?") etwas gemildert. Voreingestellter Boolescher Operator ist UND. Disjunktive Suchen mit OR sind durchführbar, dabei bindet OR stärker als der Standardoperator UND. Die Eingabe

Urlaub Baltrum OR Borkum

findet demnach die Schnittmenge aus Texten,

die "Urlaub" enthalten, und der Vereinigungsmenge der Dokumente, die "Baltrum" bzw. "Borkum" beinhalten. Stoppwörter werden standardmäßig ignoriert, sie sind jedoch mit Hilfe des +-Operators in Suchargumente einzubinden. So findet die Eingabe

Episode +I

Sites zu Star Wars I, obgleich "I" ein Stoppwort darstellt. Phrasen werden wie üblich durch Anführungszeichen markiert, als UND NICHT-Operator fungiert das Minuszeichen. Suchen mit Feldnamen (wie "site:" oder "link:") werden unterstützt. Bei der "Advanced Web Search" (siehe Abbildung 2) sind die Suchmöglichkeiten durch unterschiedliche Eingabefelder benutzerfreundlich angeordnet.

Da bei der automatischen Indexierung außer dem PageRank einer Seite für jeden Term seine Position im Text sowie sein Druckfont gespeichert werden, arbeitet das Retrievalsystem bei der Erstellung der Rangordnung mit mehreren Kriterien. Bei Suchen mit mehreren Argumenten ist dabei der Abstand (das ist die minimale Differenz der Positionen im Text) ein leitendes Sortierkriterium. Nach der Anzeige von Suchergebnissen ist mit "Search within results" eine Verfeinerung innerhalb der Ergebnismenge möglich.

Darstellung der Suchergebnisse

Besonders interessant ist bei Google die Anzeige der Suchergebnisse. Ist eine Seite voll indexiert, so wird deren Titel und diejenige Stelle im Dokument angezeigt, die das Suchargument betrifft. Die Suchbegriffe sind hierbei hervorgehoben. Gibt es mehrere Seiten innerhalb einer Domain, wo das Suchargument vorkommt, so werden die Textstellen (ab der zweiten Anzeige dabei eingerückt) präsentiert (siehe Abbildung 3). Ist eine Seite noch nicht voll indexiert, sondern nur durch einen Ankertext, so erscheint ausschließlich deren URL als Trefferanzeige. Nutzt man die Suchoption "I'm Feeling Lucky" ("Auf gut Glück"), so gibt es keine Trefferliste, sondern die direkte Anzeige des erstplatzierten Treffers.

Führt ein Link ins Leere (404; File not found), so hat der Nutzer stets die Möglichkeit, auf die Google-Datenbank zurückzugreifen, in der die Webseiten in derjenigen Version archiviert sind, wie sie Google indexiert hat. Diese Funktion hat besondere Relevanz bei Websites, deren Inhalte sich häufig ändern (wie z.B. Nachrichten), hat man doch im Cache die ursprüngliche, "alte" Information.

Via "Similar pages" wird das Web automatisch nach Seiten durchsucht, die der Ausgangsseite ähneln. Das hierbei eingesetzte Programm GoogleScout analysiert die URLs nach Hierarchien. Angenommen, die Ausgangsseite ist die Homepage einer Universität, so wird GoogleScout Homepages anderer Universitäten suchen. Starten wir von der Homepage eines Fachbereichs - sagen wir "Library and Information Science" - so sucht das Programm nach anderen Library and Information Science-Departments.

Durch Kooperationen mit anderen informationswirtschaftlichen Unternehmen gibt es diverse weiterführende Informationsmöglichkeiten. Abbildung 4 zeigt Verbindungen zu in Google eingebundenen Services. Hierzu gehören u.a.:
● RealNames, ● das Open Directory-Projekt (ODP), ● Wörterbucheinträge via Dictionary.com, ● News Anbieter, ● Börsennotierungen.

Bei der Ausgabe von Treffern innerhalb der ODP-Klassen werden die Seiten nach Relevance Ranking sortiert ausgegeben, eine Option, die ODP auf seiner Homepage nicht anbietet.

Die Google Toolbar

Google bietet wichtige Befehle in der Form einer eigenen Menüleiste an (siehe die obere Leiste in den Abbildungen 2

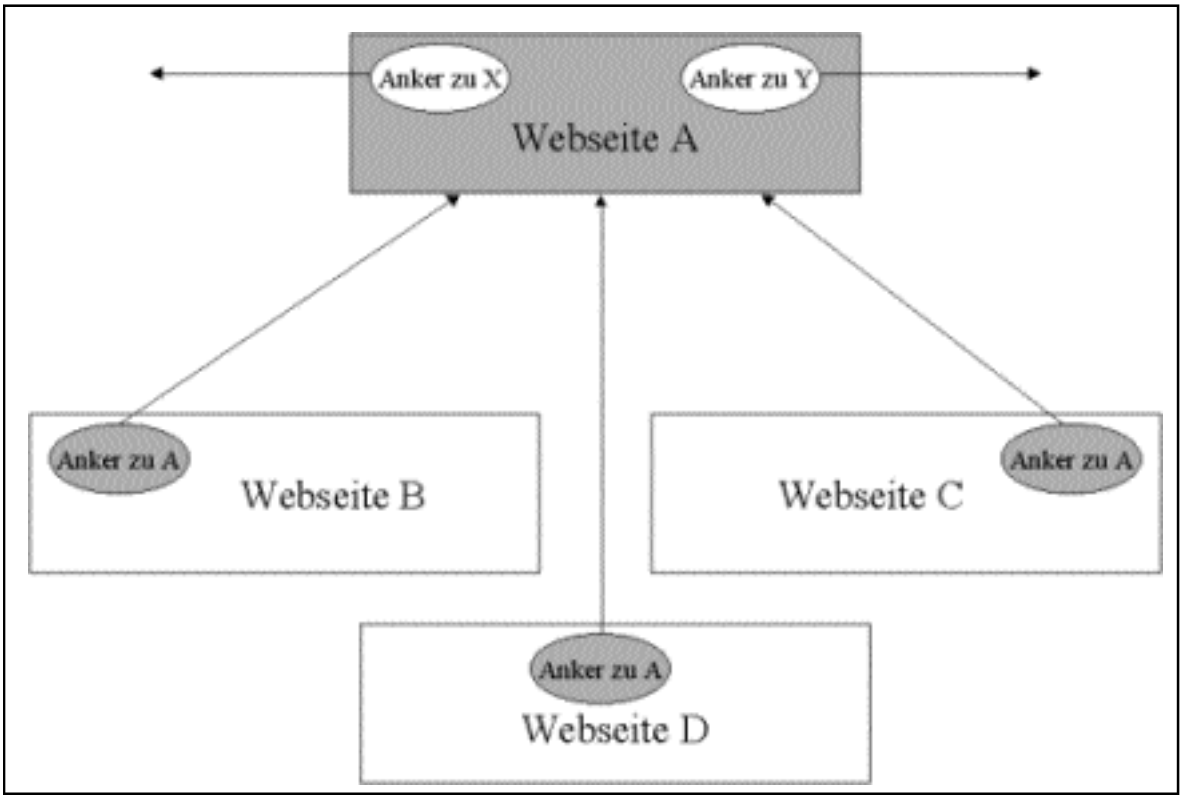

Abbildung 1: Dokumentationseinheit bei Google

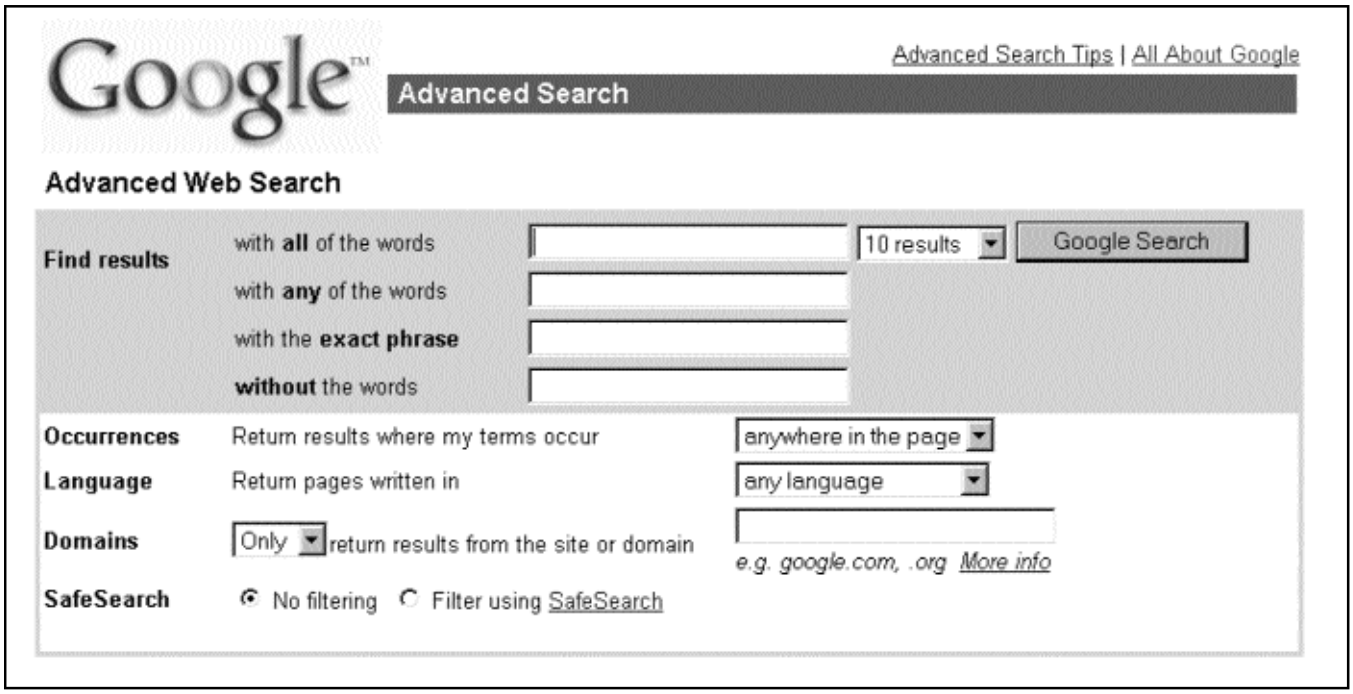

Abbildung 2: Suchbildschirm der fortgeschrittenen Suche bei Google (Ausschnitt)

und 3). Voraussetzung für die - uneingeschränkt empfehlenswerte - Installation ist der Einsatz des Internet Explorer (ab Version 5.0). Die Button haben folgende Funktionen: ● Searchbox, ● Search site (spezifiziert eine bereits formulierte Suchfrage zusätzlich nach einer Domain), ● PageRank (die PageRank-Einstufung einer Webseite wird von Google auf eine Skala, die von null bis zehn reicht, für den Nutzer anschaulich in Form eines grünen Balkens abgebildet), ● Page Info (enthält drei Funktionen: [1.] Link zum Cache, [2.] "Similar Pages": Suche mittels GoogleScout zu ähnlichen Seiten, [3.] "Backward Links": Suche nach Seiten, die zur gefundenen Seite führen), ● Up (Sprung zur nächsthöheren Hierarchieebene einer URL), ● Highlight (markiert die Suchterme im Treffer), ● Word Find (nennt die einzelne Suchterme, die nach Anklicken eine "Find in page"-Suche initialisieren).

Verkauf von Suchargumenten durch "AdWords"

Google verzichtet auf Bannerwerbung. So bleibt der Bildschirm übersichtlich und einfach strukturiert. Aber auch Google muss Geld verdienen. Ähnlich wie RealNames (vgl. Password 11/1999) verkauft Google Suchargumente an Unternehmen. Im Gegensatz zu RealNames, die vornehmlich Produkt- und Unternehmensnamen unterstützen, richtet sich Google auf komplexe Keywords, die (auch) aus Allgemeinbegriffen bestehen. RealNames verkauft ein Suchwort genau einmal und verweist auf die entsprechende Homepage, Google läßt innerhalb des Service "AdWords" alle Keywords

Abbildung 3: Google Toolbar und Ergebnisanzeige

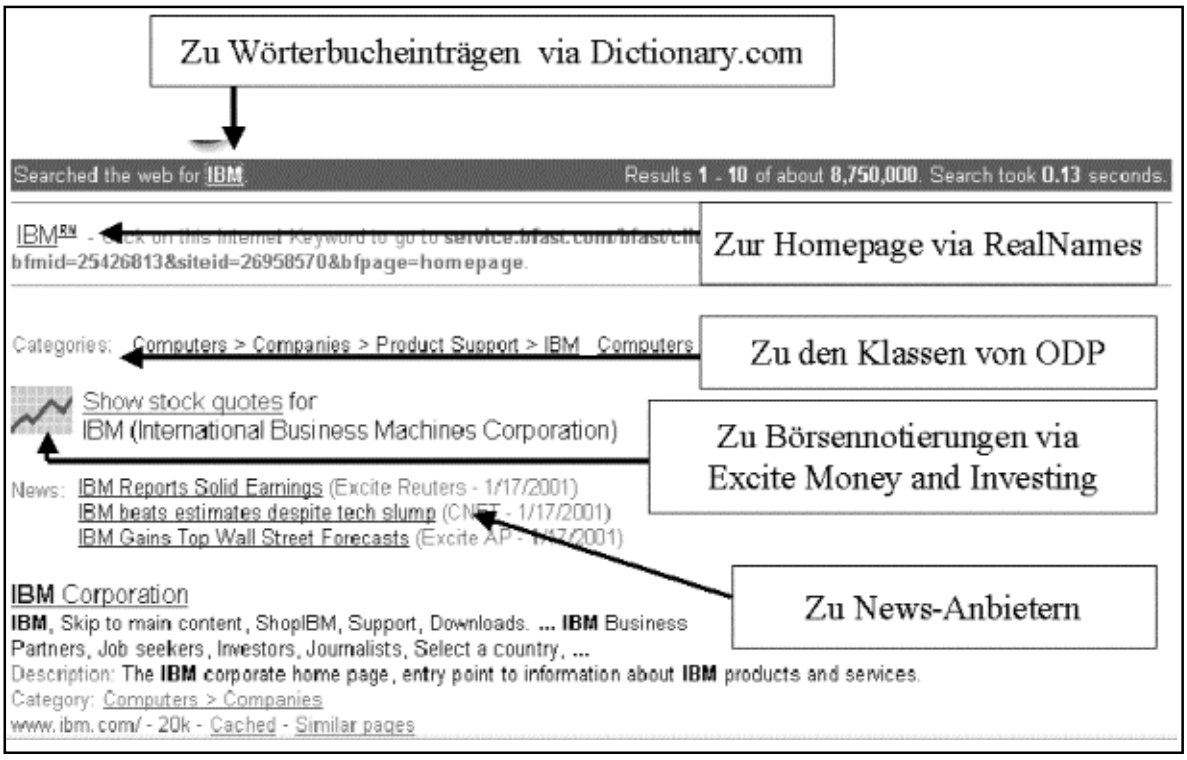

Abbildung 4: Weiterführende Links in der Ergebnisanzeige bei Google

- auch mehrfach - zu, stellt jedoch nur maximal drei Treffer in einer eigenen Spalte (im Bildschirm rechts) dar (siehe Abbildung 5). Zusätzlich gibt es bei Google ein "Premium Sponsorship", deren Treffer an der Spitze der Ergebnisanzeige aufscheinen. In Abbildung 5 ist die erste Zeile (zu www.artinstitutes.edu) ein "Premium Sponsorship", die beiden rechts positionierten (zu www.sendtraffic.com und zu theinternetbiz.com) sind AdWords. Der große Unterschied zwischen beiden Arten von "Sponsored Links" liegt im Preis; AdWords sind billiger, dafür liegen die Links etwas ungünstiger.

Wenn mehr als ein AdWords-Interessent ein Suchwort abonniert hat, rotiert Google zunächst die Rangfolgen eins bis drei. Über die Klickrate kristalliert sich eine Reihenfolge heraus, an der letztendlich die Position des Werbelinks bestimmt wird. Der Preis folgt der Position: $ 15 für je 1.000 angezeigte AdWords auf Position 1, $ 12 für 1.000 Anzeigen für die Mitte und $ 10 für Rang 3. Im Gegensatz zur recht unspezifischen Bannerwerbung ist die AdWords-Werbung stets kontextabhängig. "Unlike other Web sites that offer adversiting targeted by general content or user demographics, Google AdWords Program lets you reach only those people who have expressed an active desire for information related to your product or service", heißt es auf der Google-Homepage.

Definierbar als AdWords sind exakt zu bestimmende Keywords. Da Google keine Truncation und auch keine Wortstammanalyse anbietet, muss jede mögliche sprachliche Variante berücksichtigt werden. Bei mehreren Keywords wird standardmäßig eine UND-Verknüpfung generiert. Mit dem UND NICHT-Opera-

Abbildung 5: Verkaufte Links durch "Sponsored Links" und "AdWords" bei Google

tor lassen sich Terme negativ auszeichnen. Falls jemand z.B. "Lincoln" abonniert, aber nicht bei Suchfragen zu Lincoln, Nebraska, aufscheinen möchte, wird "Nebraska" als negatives Keyword eingesetzt. Zusätzlich sind Phrasen sowie "Exact matches" (genaue Übereinstimmung zwischen den AdWords und einer Suchfrage) formulierbar.

Bewertung

Google ist nicht nur als Testsieger unserer Known Item-Search hervorgegangen, dieses Retrievalsystem erscheint uns in der Tat - nach der Analyse von Suchtechnik und Nutzerschnittstellen - als eines der besten Suchwerkzeuge im Web. Probleme bereitet die schwache informationslinguistische Basis: kein Stemming und keine einzige Trunkierungsmöglichkeit. Genial ist der PageRank-Algorithmus, der - unabhängig von Nutzeranfragen - ein Maß für die "Popularität" einer Webseite errechnet. Bei Sortierungen nach Relevanz ist der PageRank neben dem Wortabstand (bei mehreren durch UND verknüpften Termen) und der Druckdarstellung der Terme ein entscheidendes Kriterium. Innovativ ist Google auch bei der Definition der Dokumentationseinheit, da die Ankertexte fremder Dokumente den Zieldokumenten zugerechnet werden. Das Screen Design erscheint gut gelungen, die Seiten sind einprägsam und selbsterklärend, sowohl was die Anordnung der Felder als auch was die Farbgestaltung betrifft. Nützlich ist der Cache; nach einer "File not found"-Fehlermeldung greift der Nutzer auf die Archivvariante des Dokuments zurück. Die Ergebnisdarstellung wartet mit einem weiteren Highlight auf: In den gefundenen Dokumenten werden nicht nur - wie sonst üblich - URL und Titel genannt, sondern zusätzlich der Kontext, in dem im Dokument die Suchterme aufscheinen. Als nützlich erweist sich die Google Toolbar, nicht nur, weil sie einige Vereinfachungen und Zusatzoptionen für die Suche bereithält, sondern auch beim Surfen im Web. Bei jeder aufgerufenen Seite (egal, ob via Google oder sonstwie gefunden) ist der PageRank angegeben, hier als kleiner grüner Balken, der dem Nutzer eine optische Unterstützung zur Einschätzung der Relevanz einer Seite bietet. Zu AdWords: Wenn schon Werbung, dann bitte kontextspezifisch, genau auf die Suche zugeschnitten. So erscheinen die Kleinanzeigen bei Google nicht ganz so aufdringlich, sondern vielleicht sogar hilfreich.

Das Fazit

Wir wollen unsere Artikelreihe mit einem kurzen Rückblick beenden. Ausgang war die Frage nach der Qualität von Internet-Suchwerkzeugen. Entsprechen diese Retrievalsysteme dem technischen Stand des Information Retrieval? An Vorabverurteilungen mangelt es ja nicht: Dass "die Bratkartoffeln neu erfunden würden", ist an vielen Stellen zu hören.

Qualität von Suchwerkzeugen ist näherungsweise abschätzbar durch **Retrievaltests.** Die Informationswissenschaft hält mit Recall-, Precision- und Known-Item-Untersuchungen drei Methoden von Retrievaltests bereit. Der **Recall** ist das Verhältnis der Anzahl der bei einem Werkzeug gefundenen relevanten Dokumente und aller relevanter Dokumente. Da letzteres nicht erfassbar ist, analysiert man diverse Systeme und betrachtet - als "relative" Recallbasis - die hier insgesamt gefundenen relevanten Dokumente als "alle relevanten". In einer viel beachteten Studie legen Steve Lawrence und C. Lee Giles 1998 Zahlen für den relativen Recall ausgewählter Suchmaschinen vor. Für wissenschaftliche Anfragen ergibt sich für HotBot ein relativer Recall von 57,5%, für AltaVista 46,5% und Northern Light von 32,9%. Eine Hochrechnung vom relativen Recall auf den absoluten führt zu Schätzungen von 34% für HotBot, 28% für AltaVista und 20% für Northern Light. Eine Folgeuntersuchung 1999 zeigt eine Recallverminderung der Retrievalsysteme; nunmehr erreicht kein System mehr als 16%. Ein solcher Recall wäre kaum hinnehmbar schlecht. Aber die Lawrence / Giles-Studie hat gravierende methodische Probleme: Sie geht ausschließlich von wissenschaftlichen Informationsbedarfen aus, konzentriert sich auf die englische Sprache und vollzieht eine Hoch"rechnung" vom (korrekt erhobenen) relativen Recall auf einen absoluten anhand von Schätzwerten.

Die **Precision** von Retrievalsystemen errechnet sich aus dem Quotienten der Anzahl der gefundenen relevanten Dokumente und aller (bis zu einem Cut-off-Punkt) gefundenen Dokumente. Im Test von H. Vernon Leighton und Jaideep Srivastava stehen fünf Suchmaschinen, deren Precision innerhalb der jeweils ersten 20 Treffer ermittelt wird. Die Autoren unterscheiden nach "formaler Precision" (Ergebnis nachvollziehbar, aber nicht unbedingt inhaltlich zutreffend), "schwacher Precision" (Ergebnis ist entweder potentiell relevant oder relevant) und "starker Precision" (Ergebnis ist zweifelsfrei relevant). Der Median aller fünf Suchmaschinen ergibt einen Wert von 0,81 für die formale Precision, von 0,40 für die schwache Precision und von gerademal 0,06 für die starke Precision. Legen wir einen eher schwachen Relevanzbegriff zugrunde, so ist ein Wert von 0,4 für die schwache Precision gar nicht so schlecht; verschärfen wir jedoch den Relevanzbegriff, so liegen wir mit 0,06 bei der starken Precision in Bereichen, die kaum ein Nutzer wird tolerieren können: durchschnittlich gut ein relevanter Treffer auf 20 Anzeigen, und das bei den Top 20! Auch bei Precision-Untersuchungen sind wir mit methodischen Problemen konfrontiert. Hauptursache für Fehler dürfte in der Einschätzung der Relevanz liegen, einer subjektiven Angelegenheit, deren Ergebnis je nach Versuchsperson unterschiedlich ausfallen kann.

Known-Item-Analysen umgehen das Problem der Relevanzbeurteilung, indem sie von bekannten Vorlagen ausgehen und recherchieren, ob die bekannte Vorlage in einem Suchwerkzeug gefunden wird oder nicht. Auch hier gibt es methodische Fallstricke, vor allem bei der korrekten Formulierung der Suchfragen. Unsere Known-Item-Search berücksichtigt 20 URLs mit 20 Suchfragen in der "einfachen" Suche der Retrievalsysteme und weiteren 20 in der Profi-Suche. Die "Availability" ist der Quotient aus den gefundenen Dokumenten und allen gesuchten. Die Testergebnisse ergeben folgende Rangordnung (Top 12):

1. Google: Availability:	65%
2. AltaVista:	60%
3. Northern Light :	55%
4. FAST / All the Web:	50%
4. Fireball:	50%
6. Oingo:	40%
7. Infoseek / Go:	35%
8. HotBot:	30%
8. Lycos:	30%
10. Acoon:	20%
10. Excite:	20%
12. Web-Crawler:	15%
12. Yahoo!:	15%;
mit Google-Treffern insgesamt	45%.

Hier trennt sich deutlich die Spreu vom Weizen. Die Suchwerkzeuge von Platz sieben abwärts bringen nur ein Drittel aller Known Items und weniger. Der Testsieger kommt immerhin auf zwei Drittel aller bekannter URLs. Interessant ist ein weiteres Ergebnis der Studie: Beim geschickten Einsatz mehrerer Suchsysteme unter Ausnutzung jeweils aller Retrievaloptionen kann eine Gesamtavailability von 90% erreicht werden, so bei den Kombinationen

Google - AltaVista - FAST sowie
Google - Northern Light - FAST.

Die Top-Suchwerkzeuge im Überblick

Die von uns untersuchten **klassifikatorischen Systeme Yahoo!** und das **Open Directory-Projekt (ODP)** haben viel zu viele Klassen entwickelt, die zudem extrem präkombiniert sind. Die Größe der Datenbasen ist äußerst niedrig, dafür sind allerdings die aufgenommenen Seiten intellektuell geprüft. Schmerzlich vermisst wird der Rückgriff auf bereits bestehende Klassifikationssysteme wie u.a. die Dewey Decimal Classification (DDC). Für einen Einsatz im Web wäre jedoch die DDC zu reformieren, indem über Schlüssel zusammengesetzte Notationen zugunsten postkoordinierender unterschiedlicher Notationen ersetzt würden. Klassifikationstheoretische Fehler bei Yahoo! und ODP wie etwa die Kreation homonymer Klassennamen oder die Klassifikation nach dem Alphabet schaden der Nutzbarkeit. Mit dem Suchwerkzeug **Oingo** wird durch den Einsatz terminologischer Kontrolle das Homonymproblem gemildert. Oingo macht damit einen entscheidenden Schritt in Richtung Thesaurus.

Alle von uns analysierten **Suchmaschinen** verfügen über eine große Datenbasis. AltaVista, FAST (AllTheWeb) und Northern Light setzen bei der automatischen Indexierung auf informationslinguistische und informationsstatistische Algorithmen, Google vor allem auf die inhärente Struktur des Web. Genauer angesehen haben wir uns Scooter, **AltaVista**s Software zum Einsammeln von Webseiten. Scooter folgt den Links in bereits vorhandenen Dokumenten und führt dabei eine Dublettenkontrolle durch. Von gespiegelten Seiten speichert AltaVista (wie auch FAST) nur eine Variante. AltaVistas Oberflächen für Informationslaien (u.a. Raging Search) führen zu einem Retrievalsystem, das die Suchfrage durch weitere Suchterme zu einem "Suchtopic" anreichert. Das sich anschließende Verfahren sortiert mittels "klassischer" informationswissenschaftlicher Algorithmen (Vektorraummodell, probabilistisches Modell) die Ergebnismenge in eine Rangordnung und "beschneidet" die Gesamttreffermenge. Aus der beschnittenen Menge wird nochmals eine Rangordnung gebildet, diesmal mit Orientierung auf die Stellung der Dokumente im Web (Berechnung von "Mittelpunkt" und "Autorität"). AltaVista hält eine Suchoberfläche für Information Professionals bereit, die u.a. mengentheoretische und feldspezifische Suchen, Truncation und freie Einstellung der Sortierelemente beinhaltet. Bei der UND-Verknüpfung sortiert AltaVista (wie Google) nach dem minimalen Abstand der Terme in den gefundenen Dokumenten. Die große Stärke von **FAST (AllTheWeb)** liegt - neben der großen Datenbasis - im Retrieval von Nicht-Text-Dokumenten. Sowohl das Suchsystem, das auch mit geringen Textelementen in einer Webseite Treffer generiert, als auch Datenkompressionsverfahren ermöglichen einen bequemen und schnellen Zugriff auf Bilder, Videos usw., natürlich auch auf Texte. Technischer Höhepunkt von **Northern Light** ist die auf Clusteranalyse beruhende automatische Klassifikation, die bei großen Treffermengen eine Strukturierung und damit eine Verfeinerung der Suche bringt. Wie bei der Klärung der Homonyme bei Oingo geschieht die Verfeinerung des Sucharguments bei Northern Light durch mehrere Dialogschritte zwischen Nutzer und Suchwerkzeug. Modellcharakter hat Northern Light durch seine Hybridstruktur aus Web-Suchmaschine und kommerziellem Online-Archiv, seinem Preismodell und dem Angebot eines Profildienstes. Ein Highlight von **Google** ist PageRank, das objektive Gewichtungsverfahren von Web-Dokumenten, das auf der Hypertextstruktur des Webs aufbaut. Als Sortierkriterium für Relevance Ranking wird ein solcher Wert bei Google selbst, aber auch bei Northern Light und (mit dem "Mittelpunkt" und der "Autorität" in verwandter Form) bei AltaVista eingesetzt. Google überzeugt des weiteren durch sein Screen Design, die Archivdatenbank, seine Ergebnisdarstellung sowie die Menüleiste. Während die anderen Suchwerkzeuge mit - im Vergleich zu einem Sucharguments - unspezifischer Bannerwerbung Geld verdienen, setzt Google mit AdWords auf gezielt kontextsensitive Werbeanzeigen.

Suchwerkzeuge: State of the Art

Betrachten wir den Stand der Technik bei den Suchwerkzeugen, den Web-Directories und den Search Engines! Wo stehen die Suchwerkzeuge Anfang 2001? Randolph Hock, Kenner der beiden Welten der Suchwerkzeuge im Web und der Retrievalsysteme der kommerziellen Hosts, stellt fest: "Unfortunately - and the impact of this continues into the present - none of the search engines took advantage of the heavy-duty searching technology and approaches found in online services such as Dialog and Lexis-Nexis. Neither did the search engines nor their cousins, the Web directories, take advantage of the extensive subject classification theory and practice of the last hundred or so years" (Hock 1999, S. 2). Wir können feststellen: Inzwischen entdecken einige Internet-Unternehmen die "alte" Informationswirtschaft sowie die noch ältere Informationswissenschaft und Dokumentationspraxis. Teilweise werden auch völlig neue informationswissenschaftliche Aspekte aufgegriffen und gelöst. Wir wollen dies an einigen Beispielen belegen:

- Mit Klassifikationssystemen wie der DDC wird experimentiert (unser Beispiel ist die Wolverhampton Web Library).
- Mit Terminologiekontrolle löst man die Probleme von Homonymie und Synonymie (Oingo).
- Inhalte der "alten" Online-Archive werden gemeinsam mit Web-Inhalten innerhalb einer Datenbank vorgehalten (das Hybridsystem Northern Light).
- Clusteranalyse führt zu einer automatischen Klassifikation von Dokumenten (Northern Light).
- Die UND-Verknüpfung innerhalb eines Klassifikationssystems umfasst nicht nur direkte, sondern auch indirekte Treffer, wo eines der Suchargumente in Unterbegriffen aufscheint (Yahoo!).
- "Klassische" Algorithmen des Relevance Ranking (Saltons Vektorraummodell; probabilistische Modelle oder auch schlicht das Produkt aus IDF und WDF) finden Einsatz (AltaVista, Northern Light).
- Benutzer des Internet ohne Ahnung von Retrievaltechniken suchen anders als Information Professionals. Entsprechend halten einige Systeme unterschiedliche Suchoberflächen bereit. Bei AltaVista arbeiten Professionals mit der "Advanced Search" mit diversen Optionen (Boolesche Suche, Relevance Ranking, Truncation, Feldsuche usw.). Der Laie hat bei der AltaVista Homepage Search oder der Raging Search ein auf ihn zugeschnittenes System.
- Gerade Laien brauchen einfache, ergonomisch optimale Bildschirmoberflächen (wie sie z.B. Google anbietet).
- Laien geben nur wenige Suchargu-

mente ein (im Durchschnitt etwa 1,5). Damit die Suchmaschinen überhaupt Material für ihre Algorithmen bekommen, muss das Suchargument durch weitere Terme angereichert werden (AltaVistas Raging Search).

- Durch die Hypertextstruktur des World Wide Web ergeben sich neue Retrievaltechniken. In Weiterentwicklung von Zitierindices werden die Links (wie vormals die Fußnoten) verfolgt und als Gewichtungswert für WWW-Dokumente benutzt (Google).

Da die Internet-Suchwerkzeuge grundsätzlich den gelegentlichen Nutzer ansprechen und nicht immer eine Profi-Oberfläche bereitstellen, muss sich auch der Information Professional mit den natürlichsprachigen Suchen vertraut machen, und dies betrifft gleichermaßen den Gebrauch der Nicht-Booleschen Funktionen und das Verständnis, wie die Suchmaschinen überhaupt arbeiten. Danny Sullivan betont: "Understanding the innerworkings and mechanical tendencies of the various search engines ... will enable every searcher to hone in on the appropriate service and retrieve the best results" (Sullivan 1999, 38).

Offene Probleme

Aus der Sicht von Informationswissenschaft und professioneller Informationspraxis bleiben einige Wünsche offen. Der von Northern Light eingeschlagene Weg der **Hybride** gehört u.E. weiterverfolgt und ausgebaut. Die unterschiedlichen Welten von

(1) Dokumenten im WWW,

(2) den Nachweisen und Texten in kommerziellen Online-Archiven,

(3) den weiteren nicht-kommerziellen Datenbanken im "Deep Web" und

(4) den Artikeln der elektronischen Versionen von Zeitschriften

könnten erfolgversprechend vereinigt werden. Lösbar ist dies durch Kooperationen einschlägig ausgewiesener Unternehmen, z.B. - um willkürlich Partner zu nennen - von (1) AltaVista, (2) Lexis-Nexis, (3) Wer liefert was? und (4) Springer Link.

Die **Inhaltserschließung** kann durchaus optimiert werden. Im Web eingesetzte Klassifikationssysteme wie bei Yahoo! oder dem Open Directory-Projekt sind fehlerhaft und unpraktisch. Nötig ist eine Facettierung in grundlegende Aspekte (wahrscheinlich reichen bereits zwei Facetten: Region und Sachthema aus). Eine Überarbeitung der Begriffsordnungen - auch im Vergleich mit etablierten Klassifikationssystemen (wie z.B. die Dewey Decimal Classification, die Internationale Patentklassifikation oder die Codesysteme von Predicasts) - ist dringend erforderlich. Die derzeit viel zu schmale Datenbasis kann durch automatische Klassifikation erweitert werden. Der von Oingo beschrittene Weg der terminologischen Kontrolle ist richtungweisend; außer der Synonymierelation kennen Thesauri allerdings auch Hierarchierelationen. Diese aufzubauen, wäre der nächste Schritt. Ob dies für "allgemeine" Suchwerkzeuge machbar ist, sei dahingestellt. Machbar ist es auf jedem Fall für Fachgebiete, und nichts spricht dagegen, Suchwerkzeuge nur für gegebene Spezialgebiete zu entwickeln. Bei der automatischen Indexierung sind die Möglichkeiten der Informationslinguistik und der Informationsstatistik bei weitem noch nicht ausgereizt; zum Beispiel: Wortstammanalysen arbeiten zwar für das Englische, für die meisten anderen Sprachen aber noch nicht; mit Pronomina können die Systeme überhaupt nichts anfangen.

Auch im **Retrievalsystem** sind Verbesserungen möglich und nötig. Ein Angebot von Profildiensten mit exakt formulierbaren Suchfragen (wie bei Northern Light) ist eigentlich selbstverständlich. Bei den Basisfunktionen professioneller Suche, die - hier allerdings auch jeweils nur teilweise - bereits in kommerziellen Online-Archiven realisiert sind, sieht es bei den Suchwerkzeugen recht düster aus. Aspekte wie u.a. Fragmentierungen (auch Linkstruncation), mehrere Abstandsoperatoren, Häufigkeitsoperatoren, hierarchische Suche usw., wie sie in unserer "Checkliste für Retrievalsysteme" (vgl. Stock 2000) enthalten sind, werden bei den Suchwerkzeugen noch schmerzlich vermisst.

Werden die Bratkartoffeln neu erfunden?

Wir wollen ein vorläufiges Urteil wagen: Die Suchwerkzeuge sind primär auf den "typischen" Internetnutzer, der ohne dokumentarische oder andere informationspraktische Vorkenntnisse an seine Suche herangeht, zugeschnitten. Hier arbeiten sie derzeit gar nicht so schlecht. Information Professionals müssen sich notgedrungen auf diese Art des Retrieval einlassen. Beim Durchschauen der dahinterstehenden Algorithmen und dem geschickten Umgang damit sind auch für den Experten gute Resultate zu erzielen. Die Expertenoberflächen sind jedoch (wo es sie überhaupt gibt), gemessen an einer ideal vollständigen Palette an Suchfunktionalität, ausgesprochen suboptimal. Hilfreich für Verbesserungen sind sicherlich Kooperationen zwischen den Unternehmen der Suchwerkzeuge mit anderen Firmen und Institutionen unserer Branche, mit Online-Archiven, Datenbankproduzenten, Bibliotheken und dem Buchhandel.

Die Relevanz von Suchergebnissen liegt einerseits im Retrievalsystem begründet. Die Top-Suchwerkzeuge erfüllen diese Funktion bereits durchaus zufriedenstellend. Andererseits hängt die Relevanz an der Qualität der Datenbasis, also hier der Seiten im WWW. Und diese ist im Gegensatz zu den Datensätzen kommerzieller Online-Archive anders: keine elaborierte Fachsprache, sondern restriktiver Alltagsdialekt, restriktiver Inhalt und restriktive Form.

Unsere provokative Eingangsfrage war: Erfinden die Suchwerkzeuge die Bratkartoffeln neu? Die Frage war falsch formuliert. Es geht überhaupt nicht um Bratkartoffeln, sondern um ganz andere Zutaten: Pommes frites (das wären immerhin noch Kartoffeln) oder Nudeln oder Reis oder ...? Und es geht nicht (nur) um "feine Küche", sondern für die breite Masse der WWW-Inhalte um intellektuelles Fastfood. ■

Mechtild Stock &
Wolfgang G. Stock

Literatur

Randolph Hock: The Extreme Searcher's Guide to Web Search Engines. - Medford, NJ: CyberAge Books, 1999.

Danny Sullivan: Crawling under the hood. An update on search engine technology. - In: Online 23 (1999), Nr. 1, 30-38.

Wolfgang G. Stock: Checkliste für Retrievalsysteme. Qualitätskriterien von Suchmaschinen. - In: Password Nr. 5 (2000), 22-31.

Google

Sergey Brin; Lawrence Page: The anatomy of a large-scale hypertextual Web search engine.- In: Computer Networks and ISDN Systems 30 (1998), 107-117.

Junghoo Cho; Hector Garcia-Molina; Lawrence Page: Efficient crawling through URL ordering. - In: Computer Networks and ISDN Systems 30 (1998), 161-172.

Eugene Garfield: Uses and misuses of citation frequency.- In: Eugene Garfield: Ghostwriting and Other Essays. Essays of an Information Scientist: 1985. - Philadelphia: ISI Press, 1986, 403-409.

John Harney: Value-based searching. - In: Knowledge Management Dec. 2000 / Jan. 2001, 10.

Stacy Nowicki: Google.- In: Journal of Government Information 27 (2000), 71-74.

Larry Page; Jeff Pemberton: Organizing the world's information. Google raises the bar on search technology. - Online 24 (2000), Nr. 3, S. 41-48.