Internet

Seit dem es das Web gibt, seit dem gibt es Suchmaschinen und ebenso Klagen über quantitativ und qualitativ unzureichende Suchergebnisse, schlechte Rankingverfahren und so weiter. Auch die Entwicklung von Meta-Suchmaschinen hat daran nicht viel geändert. Und alle Ansätze, die darauf setzten, dass die Seitenersteller oder die Nutzer selbst sich im Umgang mit dem Web qualifizieren, sind bislang gescheitert. Dennoch gibt es neue und Erfolg versprechende Verbesserungen, die sich die Suchenden zunutze machen können.

»Find what I mean not what I say«
Neuere Ansätze zur Qualifizierung von Suchmaschinen-Ergebnissen

Dirk Lewandowski

Aufbau und Probleme konventioneller Suchmaschinen

Um die vielfältigen Probleme, die sich bei der Arbeit mit konventionellen Suchmaschinen ergeben, beurteilen zu können, soll zunächst noch einmal kurz dargestellt werden, wie diese arbeiten.

So genannte *Gatherer* durchkämmen das Web, indem sie die in einem bekannten Dokument vorgefundenen Hyperlinks besuchen, die neu gefundenen Seiten weiter an den *Indexer* geben und wiederum den gefundenen Links folgen. So ergibt sich bei der fortschreitenden Suche ein immer größer werdendes Netz von Seiten, welches im Idealfall ein Abbild des kompletten World Wide Web bildet.

Der Indexer genannte Teil des Suchsystems erstellt eine Datenbank der gefundenen Seiten und bereitet diese für spätere Suchanfragen auf. Der dritte Teil des Suchsystems ist die Software, welche die eingehenden Suchanfragen bearbeitet und die gefundenen Treffer durch ein so genanntes *Ranking-Verfahren* in eine bestimmte Reihenfolge bringt.

Unterschiedliche Ranking-Verfahren

Die beschriebene Vorgehensweise ist prinzipiell bei allen Suchmaschinen gleich, erhebliche Unterschiede ergeben sich allerdings bei den Ranking-Verfahren. Allgemein gilt jedoch, dass Suchbegriffe, die an exponierter Stelle des Dokuments (Titelzeile, Überschrift, Unterüberschrift, verlinkter Text, Anfang des Dokuments…) auftauchen, höher bewertet werden als Begriffe, die einfach im laufenden Text vorkommen. Die genauen Verfahren werden von den Suchmaschinenbetreibern geheim gehalten, da sie das Kernstück des jeweiligen Systems bilden und das individuell Besondere desselben darstellen.

Rankingverfahren wurden notwendig, weil es sich beim WWW um einen immensen Datenbestand handelt. Während konventionelle Online-Datenbanken meist einen »überschaubaren« Bestand anbieten und deshalb die Treffermenge bei normalen Suchanfragen in einem Bereich liegt, in dem alle Dokumente durch den Benutzer geprüft werden können, geht die Anzahl der Treffer bei WWW-Suchen oft in die Tausende, wenn nicht gar Hunderttausende.

> Die genauen Ranking-Verfahren werden von den Suchmaschinenbetreibern geheim gehalten, da sie das Kernstück des jeweiligen Systems bilden und das individuell Besondere desselben darstellen.

Es wird davon ausgegangen, dass der durchschnittliche Suchende etwa zehn bis zwanzig ausgegebene Treffer überprüft. Alle Treffer, die beim Ranking nicht unter die ersten zwanzig kommen, müssen damit als verloren angesehen werden.

Unzureichende Abdeckung des Web

Trotz der immensen Menge der indexierten Dokumente erreicht keine Suchmaschine einen Abdeckungsgrad, der sich den idealen hundert Prozent auch nur annähert[1]. Dies liegt zum einen an Faktoren, die eine Indexierung schlicht unmöglich machen: von einer Datenbank kann eine Suchmaschine nur die in HTML geschriebene Formularseite indexieren, nicht jedoch die Inhalte der Datenbank. Dazu müssten *alle* theoretisch möglichen Suchanfragen an die Datenbank gestellt werden, um dann jeweils die Ergebnisseiten zu indexieren. Nicht nur im Falle großer Datenbestände wie beispielsweise Bibliothekskatalogen muss dies scheitern.

Problematisch an dieser Tatsache ist, dass in den letzten Jahren immer mehr Seiten dynamisch erzeugt werden, das heißt, es findet erst in dem Moment, in dem ein Nutzer die Seite anfragt, die Erstellung derselben einerseits aus dem Inhalt und andererseits aus den Designelementen statt.

Diese Methode wird vor allem im Zusammenhang mit Content-Management-Systemen verwendet, die ein einheitliches Erscheinungsbild und eine leichtere Pflege der Webseiten gewährleisten. Allerdings sind einige große Anbieter gerade wegen der geringen Indexierungstauglichkeit von diesem Ansatz abgekommen und verwenden Content-Management-Systeme, die statische Seiten erstellen.

Ein weiteres Hindernis für SM sind durch Passwortabfragen geschützte Bereiche, die natürlich auch für die Gatherer nicht zugänglich sind. Ebenso verhält es sich mit Seiten und Verzeichnissen, die durch den Autor mit einem Vermerk gekennzeichnet wurden, dass sie durch SM nicht indexiert werden sollen (*non-follow*). ▷

1 Vgl. Steve Lawrence/Giles, C. Lee: Accessibility of information on the web. Nature 400(1999)8, S. 108, fig. 2. Wichtig ist hier nicht der genaue Abdeckungsgrad, welcher sich stetig ändert, sondern allein die Tatsache, dass dieser sehr gering ist.

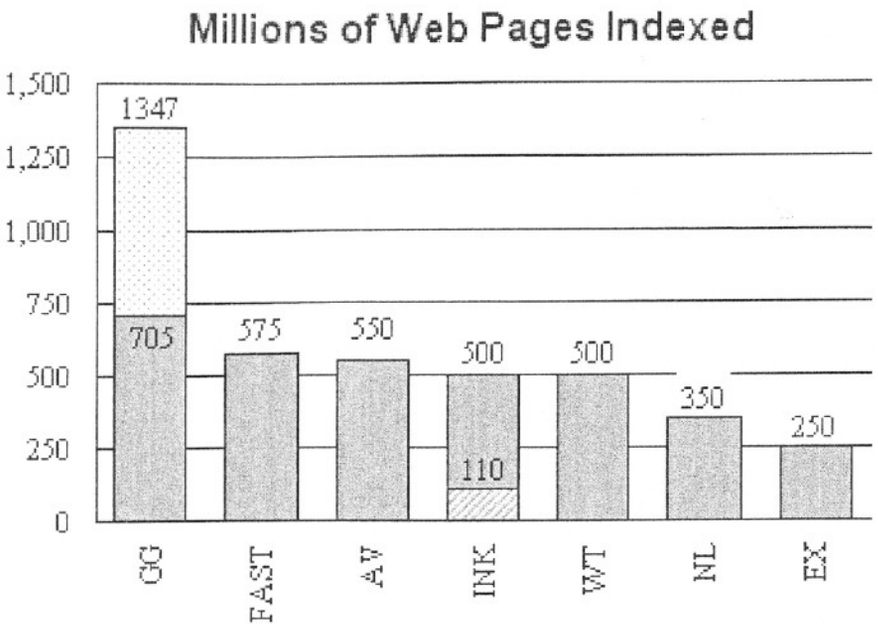


Abbildung 1: Indexgrößen der wesentlichen internationalen Suchmaschinen, Stand: 6. April 2001 (Quelle: Search Engine Watch, *www.searchenginewatch.com/reports/sizes.html);* Legende: GG=Google, FAST=FAST, AV=AltaVista, INK=Inktomi, WT=WebTop.com, NL=Northern Light, EX=Excite

Auch wenn die nicht indexierbaren Seiten außen vor gelassen werden, erreichen die Suchmaschinen keine vollständige Abdeckung des Web. Hier spielen erstens technische Gegebenheiten eine Rolle: Je größer der Index einer Suchmaschine, desto leistungsfähiger muss das dahinter stehende Rechnersystem sein. Dies ist natürlich auch eine Geldfrage. In Abbildung 1 (oben) sind die Indexgrößen der wesentlichen internationalen Suchmaschinen dargestellt.

Probleme der Indexierung

Zweitens geht es um die Tiefe der Indexierung. Während bei privaten Domains beispielsweise davon ausgegangen werden kann, dass es nur wenige Verzeichnisstufen in der Hierarchie geben wird, so kann es bei großen Anbietern sehr weit in die Tiefe gehen. Hier setzt sich jede Suchmaschine eine Grenze der Indexierungstiefe. Wäre dies nicht der Fall, so würden relativ wenige Domains sehr ausführlich erfasst, während ein Großteil der Angebote nicht erfasst werden würde. Dazu kommt, dass eine Suchmaschine die indexierten Seiten in regelmäßigen Abständen wieder überprüfen muss, da sich jederzeit Änderungen ergeben können oder die Seite inzwischen gelöscht worden sein kann.

Drittens beschränkt sich die Indexierung durch Suchmaschinen weitgehend auf textuelle Informationen. Das bedeutet, dass beispielsweise Bilder nur mittels der sie umgebenden Beschreibungstexte erfasst werden, nicht jedoch als eigenständige Informationseinheit. So würde eine Suchmaschine auf die Anfrage nach »Gerhard Schröder« und »Bild« eben nicht primär Seiten finden, die ein Bild des Bundeskanzlers enthalten, sondern zu einem großen Teil solche, auf denen er entweder im Zusammenhand mit der Bild-Zeitung steht oder aber die Texte enthalten wie »Gerhard Schröder machte sich ein Bild von…«

Die Problematik der nicht-textuellen Informationen soll im Folgenden jedoch ausgespart bleiben, da sie von den Suchmaschinen andere Ansätze erfordert und den hier vorgegebenen Rahmen sprengen würde.

Viertens schließlich besteht seitens der SM-Betreiber nicht unbedingt der Wille, einen Index größtmöglicher Vollständigkeit zu erstellen. Das hat zum einen mit den Kosten für die Geräte zu tun, zum anderen mit der fehlenden Notwendigkeit, da sich damit nicht unbedingt Nutzer gewinnen lassen. Werbung, Zusatznutzen und Gimmicks scheinen hier bessere Instrumente zu sein. Außerdem ist die beste Suchmaschine nicht unbedingt diejenige mit dem größten Datenbestand. Da aber die Rankingverfahren aufgrund der Geheimhaltung nur sehr eingeschränkt vergleichbar sind und die Indexgröße das einzig objektiv vergleichbare Merkmal darstellt, kommt es oft zur Gleichsetzung von Indexgröße und Qualität des Suchwerkzeugs.

Vielfältige Rankingverfahren

Während die aufgeführten Punkte die Problematik der Gesamtdokumentenmenge betreffen, so ist als zweiter großer Problemkreis die Methode des Rankings anzusehen. Grundsätzlich können Rankingmethoden bei den hier behandelten Dokumentenmengen im Falle von Ein- bzw. Zwei-Wort-Anfragen nur sehr eingeschränkt greifen. Nach Angaben der Suchmaschinenbetreiber werden aber hauptsächlich solche Anfragen gestellt.

Eine klassische Methode des Rankings, nämlich die Dokumentstruktur anhand der verwendeten HTML-Tags wie <h1> für die Hauptüberschrift und so weiter zu erfassen, scheitert seit einigen Jahren zunehmend am veränderten Einsatz der Sprache HTML: Wurde sie in der Anfangszeit des WWW noch als echte Textauszeichnungssprache verwendet (wobei das Layout weitgehend den Benutzervorgaben überlassen wurde), so ist diese Funktion gegenüber den Designelementen in den Hintergrund gerückt oder ganz verschwunden.

Es kann inzwischen beispielsweise nicht mehr davon ausgegangen werden, dass eine Überschrift, die relevante Begriffe enthält, auch tatsächlich durch einen <h$_n$> Tag ausgezeichnet ist. Vielmehr ist der entgegengesetzte Fall zu erwarten, nämlich dass die Überschrift aus gestalterischen Gründen gleich als Grafik erstellt wurde und höchstens noch im Ersatztext der Grafik als Text vorhanden ist.

Drei Ansätze, um Suchergebnisse zu verbessern

Um Suchergebnisse zu verbessern, können prinzipiell drei Ansätze verfolgt werden: entweder setzt man beim Suchenden, bei den Webautoren oder aber bei den Suchmaschinen selbst an.

Der erste Ansatz bedeutet schlicht, dass die Benutzer sich erweiterte Suchtechniken aneignen müssen, um gezielter recherchieren zu können. Dabei müssen sie sich mit komplexen Auswahlmenüs und der Struktur von HTML-Seiten auseinander setzen. Verständlicherweise sind viele Benutzer nicht willens oder fähig, sich diese Techniken anzueignen.

Lohnend ist die Benutzung von Metasuchmaschinen vor allem bei Anfragen, die nur sehr wenige Treffer versprechen, oder bei solchen, bei denen es auf eine höchstmögliche Vollständigkeit ankommt.

Bei den Autoren anzusetzen, heißt, diese teilweise mit der Erschließung ihrer eigenen Seiten zu betrauen. Sie sollen ihre Seiten durch Strukturvorgaben und vor allem durch Metadaten besser indexierbar machen. Dies mag wünschenswert sein, in der Praxis hat sich jedoch gezeigt, dass Metadaten nur selten vergeben werden oder bei der Vergabe oft unehrliche Beschreibungen abgegeben werden, um die Rankingverfahren der Suchmaschinen zu täuschen.

Der dritte Ansatz setzt bei den Suchmaschinen selbst an: sie sollen durch zusätzliche Features die Suchergebnisse aufwerten; sei dies durch für den Nutzer nicht sichtbare Verfahren oder durch solche, die das in einem Prozess der Interaktion mit dem Benutzer qualifizieren.

Im Weiteren sollen nun einige Verfahren von Seiten einzelner Suchmaschinen betrachtet werden, die diesen dritten Ansatz verfolgen.

Meta-Suche

Die bekannteste und auch weitgehend etablierte Art, das Problem der für eine Suchanfrage zu geringen Gesamtdokumentenmenge oder das Problem des unzureichenden Rankings anzugehen, ist die Benutzung von Metasuchmaschinen. Diese sammeln selbst keine Daten, sondern machen sich die Vorarbeit anderer Suchmaschinen zunutze: Wird eine Suchanfrage gestellt, so wird diese an verschiedene SM weitergeleitet. Die eintreffenden Ergebnisse werden wiederum geordnet und dem Nutzer als eine Ergebnisseite präsentiert.

Der große Vorteil dieser Methode ist der geringe Aufwand für den Nutzer: Müsste er sonst die Suchmaschinen einzeln abfragen, so kann er mit einer Metasuchmaschine die gleiche Anfrage an viele Datenbanken stellen und erhöht so die Wahrscheinlichkeit eines Treffers gegenüber der Anfrage an nur eine Suchmaschine. Lohnend ist die Benutzung von Metasuchmaschinen vor allem bei Anfragen, die nur sehr wenige Treffer versprechen, oder bei solchen, bei denen es auf eine höchstmögliche Vollständigkeit ankommt. Im ersten Fall kann die Treffermenge signifikant erhöht werden[2], im zweiten wird der Abdeckungsgrad des Web erhöht.

Wichtig ist allerdings, darauf zu achten, dass die gewählte Meta-Suchmaschine auch die wesentlichen Suchmaschinen, das heißt die mit den größten Datenbeständen und den etablierten Rankingverfahren, abdeckt. Wird beispielsweise die größte SM nicht mit abgefragt, so kann es durchaus der Fall sein, dass die Menge der gesamten durchsuchten Dokumente bei der Metasuche geringer ist als die der größten konventionellen Suchmaschinen.

Für ungeübte Benutzer ist die Popularitätsanalyse eine ideale Hilfe: Man muss sich keinerlei Kenntnisse über Suchvorgänge, Operatoren oder einschränkende Befehle aneignen, um zu brauchbaren Ergebnissen zu kommen.

Qualitätskriterien für die Metasucher

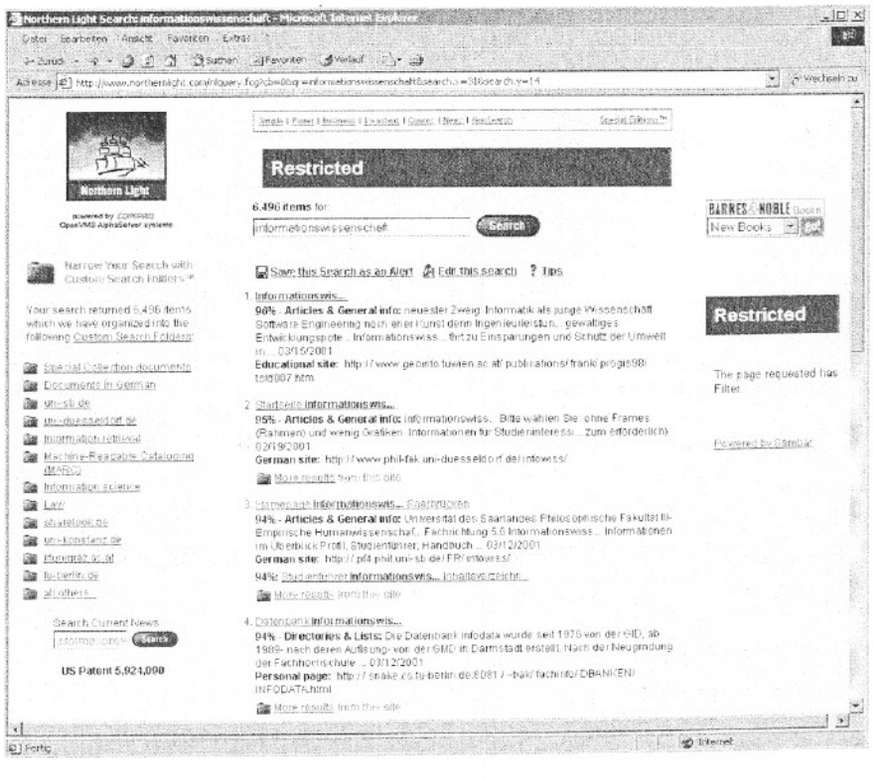
Inzwischen gibt es allein für deutschsprachige Seiten etwa 35 Metasucher. Diese hohe Zahl rührt zu einem Teil daher, dass diese Art im Vergleich zu konventionellen SM relativ leicht und kostengünstig zu erstellen und zu betreiben ist. Viele Metasucher zielen auf das schnelle Geld durch Bannerwerbung und vernachlässigen, dass eine einfache Zusammenführung fremder Suchergebnisse auf einer Seite allein noch keinen wesentlichen Mehrwert bringt.

Die folgenden Qualitätskriterien (nach *Sander-Beuermann*[3]) sollten als verbindlich angesehen werden: parallele Suche (keine All-in-one-Formulare), die Ergebnisse müssen zusammengeführt und in einem einheitlichen Format dargestellt werden, Dubletten-Eliminierung, AND- und OR-Operatoren, Übernahme der von einzelnen SM gelieferten Kurzbeschreibungen, »Search Engine Hiding« (das heißt, die spezifischen Eigenschaften der unter der Meta-Maschine liegenden Suchdienste dürfen für die Bedienung keine Rolle spielen), Vollständigkeit der Suche.

Die meisten (deutschsprachigen) Angebote sind allerdings weit davon entfernt, diese Kriterien zu erfüllen. Eine geeignete Meta-Suchmaschine ist also mit Sorgfalt auszuwählen.

Ein Nachteil, der bei Meta-SM allerdings immer vorhanden sein wird, ist die Unmöglichkeit, Spezifika einzelner SM zu nutzen. Neben den oben genannten Minimalanforderungen an Operatoren, bieten einige SM umfangreiche Möglichkeiten, eine Suche zu qualifizieren: So lassen sich Anfragen beispielsweise auf Informationen in der Titelzeile einer Seite oder auf bestimmte Domains einschränken. ▷

Abbildung 2: »Custom search folders« bei Northern Light



2 Vgl. Lawrence/Giles (a.a.O.), S. 108: »The overlap between the engines remains relatively low; combining the results of multiple engines greatly improves coverage of the web during searches.«

3 Wolfgang Sander-Beuermann: Schatzsucher: die Internet-Suchmaschinen der Zukunft. In: c't 13/98, S. 178

Dirk Lewandowski, geboren 1973; Abschluss an der HBI Stuttgart 1997. Seitdem Studium der Philosophie, Informationswissenschaft und Medienwissenschaft an der Heinrich-Heine-Universität Düsseldorf und Angestellter in der Bibliothek und Informationsvermittlung des Ministeriums für Wirtschaft und Mittelstand, Energie und Verkehr des Landes NRW. – Privatanschrift: Mühltaler Straße 24, 40221 Düsseldorf; E-Mail dirk.lewandowski@uni-duesseldorf.de

Für einfache Suchanfragen oder um sich einen ersten Überblick zu schaffen, sind Meta-SM durchaus zu empfehlen. Der fortgeschrittene Sucher kommt allerdings oft schon mit den erweiterten Suchformularen der konventionellen SM schneller zu besseren Ergebnissen.

Popularitätsanalyse / Zitationsbewertung

Die Berechnung von Zitationsraten einzelner Webseiten ist eine Spezialität der Suchmaschine *Google*[4]. Das System orientiert sich an den wissenschaftlichen Zitierindices und lässt sich grob gesagt auf die folgende Formel bringen: »Je mehr Seiten auf eine gegebene Seite X verweisen, desto wichtiger ist diese.«

Allerdings wird nicht nur positiv bewertet, *dass* ein Link auf diese Seite gesetzt wurde; wichtig ist auch, *wer* diesen Link gesetzt hat. So gilt ein Verweis durch eine populäre Seite wie beispielsweise Yahoo als wertvoller als der von einer weniger populären Seite.

Dabei wird nicht auf eine intellektuelle (und damit manipulierbare) Bewertung zurückgegriffen, sondern ein eigenes Verfahren entwickelt, bei dem für jede Seite aufgrund der Rückverfolgung ihrer Verlinkung ein Wert bestimmt wird, der bei einer Suchanfrage die Rankingposition der gefundenen Seiten wesentlich bestimmt. Der Algorithmus ist dokumentiert[5] und gewährleistet ein Ranking nach objektiven Kriterien.

Ein wesentlicher Vorteil ist, dass das System weitgehend unanfällig gegen *spamming* (die absichtlich falsche Aufbereitung von Webseiten zum Zweck der Werbung oder Irreführung) ist. *Page/Pemberton* verdeutlichen den Mechanismus an einem Beispiel:

Ein Webmaster möchte die Benutzer glauben machen, dass seine Seite diejenige der Stanford University wäre. Dazu kopiert er den kompletten Inhalt der echten Seite auf seinen Rechner. Für eine Suchmaschine sind die beiden Seiten nicht nach Original und Kopie zu unterscheiden. Erst die Verlinkung der Seiten gibt Aufschluss über die echte Seite: Diejenige Seite, die von den Benutzern für richtig gehalten und deshalb auch verlinkt wird, ist das Original[6]. Das System wertet also die intellektuelle Leistung der Verlinkung maschinell aus und macht sich damit die Meinung der Nutzer zu eigen.

Die größte Stärke der Popularitätsbewertung liegt in der Bearbeitung einfacher Suchanfragen: So gibt Google bei einer Anfrage nach »bub« an zweiter Stelle die Seite der gesuchten Zeitschrift aus, da diese von vielen und als wichtig angesehenen Seiten verlinkt wurde. Für die Konkurrenz, die mit konventionellem Ranking arbeitet, ist eine solche Suchanfrage faktisch nicht bearbeitbar: sie ist zu allgemein und zu mehrdeutig.

Für ungeübte Benutzer ist die Popularitätsanalyse eine ideale Hilfe: Man muss sich keinerlei Kenntnisse über Suchvorgänge, Operatoren oder einschränkende Befehle aneignen, um zu brauchbaren Ergebnissen zu kommen. Die Methode orientiert sich am tatsächlichen Nutzerverhalten, nämlich an den Ein- oder Zwei-Wort-Anfragen, und kann diese auch größtenteils zufriedenstellend beantworten.

Erweiterung der Dokumentationseinheit

Während Suchmaschinen normalerweise verlinkten Text dem Dokument zurechnen, welches verlinkt, geht dieser Ansatz davon aus, dass der verlinkte Text zusätzlich demjenigen Dokument zuzurechnen ist, auf welches verwiesen wird. Die drei wesentlichen Vorteile dieses Ansatzes sind:

- Verlinkter Text enthält oft eine treffendere Beschreibung des Inhalts einer Seite als diese Seite selbst.
- Verlinkter Text existiert auch für Dokumente, die durch Suchmaschinen überhaupt nicht indexiert werden können (zum Beispiel Bilder, Programme, Datenbanken).
- Suchmaschinen können durch verlinkten Text Seiten indexieren, ohne sie selbst zu besuchen[7].

Auch dieser Ansatz macht sich also die intellektuelle Leistung der Webautoren zunutze. Bis zu einem gewissen Grad kann er auch dazu dienen, der Synonymproblematik Herr zu werden. Es ist wahrscheinlich, dass unterschiedliche Autoren für die Beschreibung ein und derselben Seite unterschiedliches Vokabular verwenden werden. Dieser Umstand kann für Suchanfragen verwertet werden und wird insbesondere für spezifische Anfragen mit einer geringen Treffermenge wesentliche Vorteile bringen.

Die Verwertung von verlinktem Text wird von den wesentlichen Suchmaschinen unterstützt. Auch dieser Ansatz ist für den Benutzer unsichtbar; die Qualität der Ergebnisse wird verbessert, ohne dass der Nutzer sich selbst weitere Suchtechniken aneignen muss.

Der große Vorteil der Clusteranalyse liegt einerseits in der individuellen Gruppierung. Der andere große Nutzen liegt in der intuitiven Verständlichkeit des Systems.

Verbindung Web-Suchmaschine mit klassischen Datenbanken

Bisher wenden sich die Betreiber klassischer Online-Datenbanken im Wesentlichen an professionelle Infobroker und Informationsvermittler und vernachlässigen dabei den Endkunden, welcher das System vielleicht nur ein oder wenige Male benutzen möchte. Für diesen sind die Barrieren zur professionellen Information zu hoch: Neben der Auswahl des richtigen Anbieters ist eine Anmeldung

4 www.google.de

5 Sergey Brin / Page, Lawrence: The Anatomy of a Large-Scale Hypertextual Web Search Engine. siehe: www7.scu.edu.au/programme/fullpapers/1921/com1921.htm [zuletzt geprüft: 16.4.2001]

6 Lawrence Page / Pemberton, Jeff: Organizing the world's information. Online 24(2000)3, S. 41–48

7 Brin / Page (a.a.O.). Die Problematik des letzten Punktes soll hier ausgespart bleiben, wird jedoch in der Quelle diskutiert.

und oft auch die Erlernung einer Abfragesprache nötig.

Eine Lösung ist die Integration kostenpflichtiger Quellen in Web-Suchmaschinen. Als bisher einziger Anbieter ist *Northern Light*[8] diesen Weg gegangen und hat etwa 7 000 kostenpflichtige Quellen mit etwa 25 Millionen Volltexten aus Zeitschriften und Reports in seinen Datenbestand integriert. Eine Suchanfrage findet in der Standardeinstellung sowohl Webseiten als auch die so genannten »special collection documents«.

Für den Benutzer ergeben sich die folgenden Vorteile: Es ist nur noch ein Suchschritt auszuführen, die Einarbeitung in verschiedene Systeme entfällt. Dazu kommt eine *qualitative* Aufwertung der Treffermenge; es ist davon auszugehen, dass die Dokumente der »special collection« durchschnittlich ein höheres Niveau bieten als die gefundenen Webdokumente.

Bei einer geringen Treffermenge wird die Anfrage auch *quantitativ* aufgewertet. Der Benutzer ist hier sicher auch am ehesten bereit, für die gewünschten Informationen zu bezahlen. Auch kann generell die Kostenpflichtigkeit nicht als Nachteil angesehen werden, zumindest wenn sich – wie im Falle von Northern Light – die Suche leicht auch auf die kostenlosen Webdokumente beschränken lässt. Einen weiteren Vorteil bietet das vorgestellte System dadurch, dass bibliographische Angaben und eventuell vorhandene Abstracts stets kostenlos sind. Erst bei einem tatsächlichen Aufruf des Volltextes fallen Kosten an, die in einfacher Weise über eine Kreditkartenabbuchung beglichen werden können.

Als Nachteil des bestehenden Systems lässt sich die mit 25 Millionen Dokumenten relativ geringe Textbasis der »special collection« ansehen. Dazu kommt, dass der Schwerpunkt hier auf englischsprachigen Quellen liegt.

Es ist zu hoffen, dass der Ansatz einer hybriden Suchmaschine Nachahmer finden wird, speziell ein integrierter Zugang zur deutschsprachigen Presse würde den Nutzen einer SM hierzulande wesentlich aufwerten.

Clusteranalyse

Ein weiteres Feature, das nur von Northern Light angeboten wird, ist die so genannte Clusteranalyse. Hierbei werden

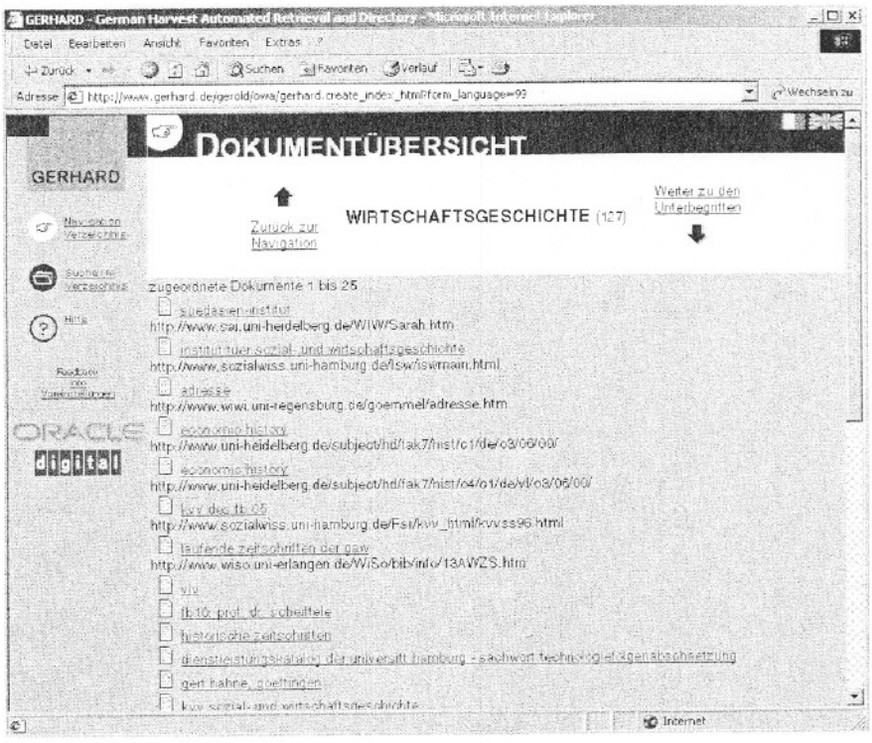


Abbildung 3: Ergebnisdarstellung bei Gerhard

für jede Suchanfrage individuell Gruppierungen vorgenommen, die bei der weiteren Einschränkung der Suche behilflich sein sollen.
Auf der Ergebnisseite erscheinen im rechten Teil, wie von anderen SM gewohnt, die durch Ranking in eine Reihenfolge gebrachten Ergebnisse.

Zusätzlich erscheint im linken Teil der Anzeige die Aufstellung der »custom search folders« (vergleiche Abbildung 2, Seite 382). Diese Cluster werden nach verschiedenen Ordnungskriterien erstellt: Thema, Art (zum Beispiel Presseerklärung), Quelle (zum Beispiel ein bestimmter Server) und Sprache. Zur Erstellung der Cluster greift das System sowohl auf manuell erstellte Suchformulierungen als auch auf informationsstatistische Verfahren zurück[9].

Bei der Beispielsuche nach »Informationswissenschaft« wurden sowohl thematische Cluster (Information Retrieval, Law, Information Science), ein sprachliches Cluster (Documents in German) als auch Cluster nach Servern (uni-duesseldorf.de, uni-sb.de) erstellt.

Wird eines der Cluster ausgewählt, wird die bisherige Suche auf dieses eingeschränkt. Es erfolgt also keine komplett neue Suchanfrage, sondern eine Auswahl aus den bisherigen Treffern.

Auch auf den so erzeugten weiteren Ergebnisseiten werden wieder Cluster gebildet; dies lässt sich fortführen, bis etwa zwanzig Treffer übrig geblieben sind. Die Betreiber von Northern Light gehen davon aus, dass diese noch verbleibende Restmenge vom Benutzer durchgesehen werden kann.

Der große Vorteil der Clusteranalyse liegt einerseits in der individuellen Gruppierung. Die Dokumente werden nicht fest in ein Klassifikationsschema geordnet, wie dies etwa bei den thematischen Katalogen der Fall ist, sondern sie können je nach Suchanfrage in den unterschiedlichsten Clustern auftauchen.

Der andere große Nutzen liegt in der intuitiven Verständlichkeit des Systems. Ist der Prozess der Clustererstellung relativ komplex und für den Nutzer kaum nachvollziehbar, so ist die Bedienung auf den ersten Blick einleuchtend. Auch sehr ungenaue Suchanfragen können durch wenige Klicks so eingeschränkt werden, dass sie zu brauchbaren Ergebnissen führen. Getreu dem Motto »find what I mean not what I say« wird der Nutzer in die Richtung der Suchanfrage gelenkt, die er zuvor vielleicht nicht auszudrücken wusste.

Seine vollen Stärken kann Northern Light leider nur bei englischsprachigen Anfragen entfalten, da die bereits vorgefertigten Suchanfragen in dieser Sprache gestellt wurden. Brauchbare Cluster werden aber in den meisten Fällen auch für Anfragen auf deutsch geliefert. Diese beziehen sich neben der obligatorischen Klasse »Documents in German« allerdings im Wesentlichen auf die Unterteilung nach den Servern, auf denen die Dokumente abgelegt sind. ▷

8 wwww.northernlight.com

9 Zur genauen Funktionsweise vgl. Wolfgang Stock / Stock, Mechthild: Internet-Suchwerkzeuge im Vergleich (III): Informationslinguistik und -statistik: AltaVista, FAST und Northern Light. In: Password 2001(1), S. 22 ff.

Automatische Klassifikation

Während die Clusteranalyse die Dokumente je nach gestellter Suchanfrage zuordnet, versucht die »echte« automatische Klassifikation, die Dokumente den einmal festgelegten Klassen zuzuordnen. Dabei kann ein bereits bestehendes Klassifikationssystem verwendet werden; im Fall der im Weiteren besprochenen Suchmaschine *Gerhard*[10] ist dies die Universale Dezimalklassifikation (UDK) in der Version der ETH Zürich.

In der Regel wird jedes Dokument einer Klasse zugeordnet und erhält mehrere Notationen. Auch hier soll das technische Verfahren nicht weiter erläutert werden; es ist an anderer Stelle dokumentiert[11].

Der wesentliche Vorteil der Verwendung einer Klassifikation gegenüber der Clusteranalyse ist die genauere Differenzierung. Dazu bietet sie dem bereits mit der Klassifikation vertrauten (Bibliotheks-)Benutzer einen einfachen Einstieg in das zu bearbeitende Wissensgebiet. Auch ist an eine mögliche Zusammenführung und einheitliche Klassifizierung von Print- und Onlineinformationen in einem einzigen System zu denken.

Sinnvoll ist in jedem Fall die Einschränkung der zu durchsuchenden Server: So durchsucht Gerhard nur wissenschaftliche Server im deutschen Sprachraum. Die Liste der Server muss dabei manuell gepflegt werden.

Nach Angaben der Betreiber ordnet Gerhard etwa achtzig Prozent der Dokumente korrekt der jeweiligen Klasse zu. Probleme bereiten die Struktur der Dokumente, Spezifika der UDK sowie Mehrsprachigkeit und Homonyme. Dennoch kann davon ausgegangen werden, dass die Benutzer eine Fehlerquote von zwanzig Prozent tolerieren.

Gerhard wendet sich im Gegensatz zu den bisher erwähnten Suchmaschinen-Betreibern explizit an ein akademisches Publikum. Hier kann die Bereitschaft vorausgesetzt werden, sich vor der Suche mit dem Suchsystem auseinander zu setzen oder sich mit der Zeit an dieses zu gewöhnen.

Größtes Manko des bestehenden Systems ist die unzureichende Anzeige in der Trefferliste. Während nahezu jede SM eine Kurzzusammenfassung des Seiteninhalts anzeigt, beschränkt sich Gerhard auf die Titelinformationen aus dem <title>-Tag und den Link. Gerade Titelinformationen sind jedoch oft wenig aussagekräftig, da sie von Autoren oft als unwichtig angesehen oder schlicht vergessen werden (Abbildung 3, Seite 385).

10 www.gerhard.de

11 www.gerhard.de/info/dokumente/dokumentation/gerhard/bericht.pdf

Zur Zeit wird der Datenbestand von Gerhard weder aktualisiert noch ergänzt. Daher ist eine differenzierte Bewertung vor allem der Zuordnungsfähigkeit hier nicht möglich.

Fazit

Die alle Probleme der Suchenden lösende Suchmaschine gibt es nicht und wird es wohl auch in Zukunft nicht geben. Dies ist im Wesentlichen mit dem sehr großen und stets weiter steigenden Datenvolumen, mit der Heterogenität der vorhandenen Daten und schließlich mit mangelnden Kenntnissen auf Benutzerseite zu begründen. Trotzdem gibt es speziell von Seiten einzelner Suchmaschinen Ansätze, die – wenn sie schon Retrievalprobleme nicht lösen können – doch helfen, diese zu lindern. Zukunftsweisend scheinen solche Ansätze zu sein, die vom Benutzer entweder gar nicht bemerkt werden, weil sie verdeckt arbeiten, oder aber solche, die den Benutzer in einen Interaktionsprozess leiten.

Dagegen werden sich solche Funktionen, die erweiterte Kenntnisse der Benutzer erfordern, nur bei einem relativ kleinen Nutzerkreis durchsetzen; hier können sie jedoch durchaus gewinnbringend sein.

Bibliotheksbau

Auch in Sparzeiten ist innovative Bibliotheksarbeit nötig – und wird ermöglicht durch Weitsicht von Direktion, Stiftungsrat, Freundesgesellschaft und privaten Förderern der Gerd Bucerius Bibliothek im Museum für Kunst und Gewerbe in Hamburg. Seit ihrer Neueröffnung vor rund einem halben Jahr will die Bibliothek nicht nur eine »gewöhnliche« Arbeitsstätte sein, sondern fühlt sich in Dienstleistung und Präsentation der Gewerbetradition und dem Museumsgedanken verpflichtet.

Neu eröffnet: Die Gerd Bucerius Bibliothek im Museum für Kunst und Gewerbe Hamburg

Ein Forschungszentrum und bibliophiler Schauraum

Angela Graf

Justus Brinckmann, Museumsgründer

In Hamburg lebte einst ein Mann mit Visionen: *Justus Brinckmann* (1843–1915) ergriff die Initiative, nach der schließlich die »Patriotische Gesellschaft[1]« ein ›gewerbliches Museum‹ gründete.

Schon mit Errichtung des ersten Provisoriums 1874 begann die Geschichte der Museumsbibliothek[2]. 1877 übernahm die Stadt Hamburg das Museum, und es zog in einige Räume des neuen und heute weit über die Stadtgrenzen hinaus bekannten gelben Hauses am Steintorplatz ein. (Den Hamburger Hauptbahnhof gegenüber gab es damals noch nicht.)

1869 hatte das Gründungspromemoria einen ausdrücklichen Passus enthalten über die Einrichtung einer Fachbibliothek, die nicht nur »der Verwaltung des Museums für Sammlung, Ordnung und Nutzbarmachung der Museumsgegenstände« dienen sollte, »sondern es müsste auch dafür gesorgt werden, dass zur Benutzung der Bücher durch Gewerbetreibende und Lernende [...] möglichst bequeme Gelegenheit geboten werde[3].«

Umso wichtiger, als damals in Hamburg die einschlägige Literatur nur selten gesammelt und/oder für die erwarteten Benutzer schlecht zugänglich war[4]. Im Haus, das das Museum beherbergte, befanden sich auch verschiedene zur selben