Internet

»Alle benutzen Google«. So lässt sich kurz und prägnant die Entwicklung im Bereich der WWW-Suche in der letzten Zeit auf den Punkt bringen. Durch gute Suchergebnisse und eine schlichte, ausgesprochen gut bedienbare Benutzerschnittstelle hat sich Google als die Suchmaschine für alle Zwecke etabliert. Im Zuge dieser Entwicklung fanden größere Veränderungen auf dem Markt statt: Einige Betreiber mussten ihre Suchwerkzeuge aufgeben, neue Firmen haben dafür die Herausforderung angenommen, dem Benutzer noch bessere Ergebnisse oder wenigstens innovative Features zu bieten.

Alles nur noch Google?
Entwicklungen im Bereich der WWW-Suchmaschinen

Dirk Lewandowski

Marktbereinigung

Insgesamt fand im Laufe des letzten Jahres eine Marktbereinigung statt: Einige Betreiber von Suchmaschinen haben aufgegeben (darunter *Direct Hit* und *Excite*). Besonders bedauerlich ist der Wegfall der innovativen Suchmaschine *Northern Light*, deren Idee der Ergebnisclusterung aber inzwischen von neuen Anbietern aufgenommen wurde. Desweiteren scheint auch *Gerhard* (bibliothekarisch interessant durch seine automatische Einordnung der Ergebnisse in die UDK) nicht mehr fortgeführt zu werden.

Andere Suchmaschinen verwenden keine eigene Datenbank mehr, sondern kaufen ihren Datenbestand bei anderen Suchmaschinen ein. So z.B. *Lycos:* diese Suchmaschine »der ersten Stunde« zeigt inzwischen Ergebnisse aus der Datenbank von *FAST* an, die auch direkt über die *FAST*-eigene Suchmaschine *Alltheweb (www.alltheweb.com)* recherchierbar sind.

Im Bereich der großen internationalen Suchmaschinen sind nur wenige Anbieter übrig geblieben. Eigene Datenbestände werden noch von *Google, Alltheweb* und *Alta Vista* verwaltet. Einen Sonderfall bildet der *Inktomi*-Index: Die Suchergebnisse sind über verschiedene Oberflächen abrufbar (zum Beispiel *Hotbot, AOL Search*), die Firma *Inktomi* bietet aber selbst keine Suchoberfläche an, sondern verkauft ihre Suchtechnologie und ihre Datenbestände nur an andere Anbieter.

Neue Suchmaschinen

Erfreulich ist, dass nach einer längeren Phase, in der kaum neue Suchmaschinen und Technologien vorgestellt wurden, seit dem letzten Jahr wieder einige neue Suchmaschinen entstanden sind. Diese legen meist ihren Schwerpunkt auf eine innovative Technologie der Ergebnisgewichtung oder auf Zusatzfeatures. Ein Wettbewerb bei der Zahl der indexierten Dokumente findet nur in wenigen Fällen statt.

Dieser Punkt wird wohl in der Zukunft wieder an Bedeutung gewinnen: Bei Erfolg einer Technologie werden die Indices wohl kontinuierlich erweitert werden. Nahezu alle neuen Suchmaschinen machen sich die Linkstruktur des WWW zunutze, um die Relevanz einzelner Seiten für eine Suchanfrage zu bewerten[1].

Erfreulich ist, dass nach einer längeren Phase, in der kaum neue Suchmaschinen und Technologien vorgestellt wurden, seit dem letzten Jahr wieder einige neue Suchmaschinen entstanden sind.

Zu den viel versprechenden, innerhalb des letzten Jahres neu gestarteten Werkzeugen zählen unter anderem *Teoma, Wisenut, Vivisimo* und *Openfind.* Zwei von ihnen sollen im Folgenden kurz beschrieben werden.

Teoma

Teoma (www.teoma.com) verwendet ein Rankingverfahren, das auf der Idee der Linkpopularität aufbaut. Während allerdings *Google,* das sich auch wesentlich auf dieser Methode gründet, prinzipiell jeden Link auf eine Seite als Stimme für diese bewertet (wenn auch mit unterschiedlichen Faktoren), fließen bei *Teoma* nur Links von solchen Seiten in die Bewertung mit ein, die auch selbst die Suchbegriffe enthalten. Dadurch soll gewährleistet werden, dass nur die »Experten« für ein Thema die Qualität anderer Seiten zum gleichen Thema bewerten. Der Ansatz ist viel versprechend und bringt auch gute Suchergebnisse zu Tage. Die Probleme von *Teoma* liegen eher an der relativ kleinen Ergebnisdatenbank und deren mangelnder Aktualität.

Eine Besonderheit von *Teoma* ist die Aufteilung der Ergebnisseite in drei Teile (vergleiche Abbildung 1, Seite 559): Zum einen wird im Hauptteil wie bei anderen Suchmaschinen auch die Ergebnisliste angezeigt. Dazu kommen »Suggestions to narrow your search«, also Vorschläge, mit welchen Begriffen sich die Suche weiter einschränken lässt. Diese Vorschläge sind oft brauchbar, vor allem bei nicht-englischsprachigen Suchbegriffen aber meist irrelevant bis unsinnig.

Der dritte Teil der Ergebnisseite verspricht einen höheren Nutzen: hier werden »link collections from experts and enthusiasts« aufgelistet. Dabei handelt es sich um Linksammlungen, die von den durch das Ranking ermittelten »Experten« erstellt wurden. Oft bieten sie einen exzellenten Einstieg in ein Thema, da sie wesentliche Quellen erschließen, unter Umständen auch so genannte *Invisible-Web*-Ressourcen (siehe unten).

1 Vgl. dazu: Dirk Lewandowski: »Find what I mean not what I say«: Neuere Ansätze zur Qualifizierung von Suchmaschinen-Ergebnissen. In: BuB 53(2001)6/7, S. 381–386

2 Steve Lawrence; Giles, C. Lee: Accessibility of information on the web. Nature 400(1998)8, S. 108

Vivisimo

Die Firma *Vivisimo* (*www.vivisimo.com*) bezeichnet sich selbst als »the document clustering company«. Wie bei der inzwischen eingestellten Suchmaschine *Northernlight* werden die Suchergebnisse thematisch in Ordner einsortiert. Mit einem Klick auf den Ordnernamen werden die Ergebnisse nach dem Ordnernamen gefiltert und man erhält nach einem oder mehreren Klicks eine überschaubare Trefferliste (siehe Abbildung 2 auf dieser Seite). Die für unsere Suche nach »Bibliothek« vorgeschlagenen Cluster sind sinnvoll, sichtbar wird jedoch die geringe Indexgröße: gerade einmal 191 Treffer wurden zu diesem Suchbegriff gefunden.

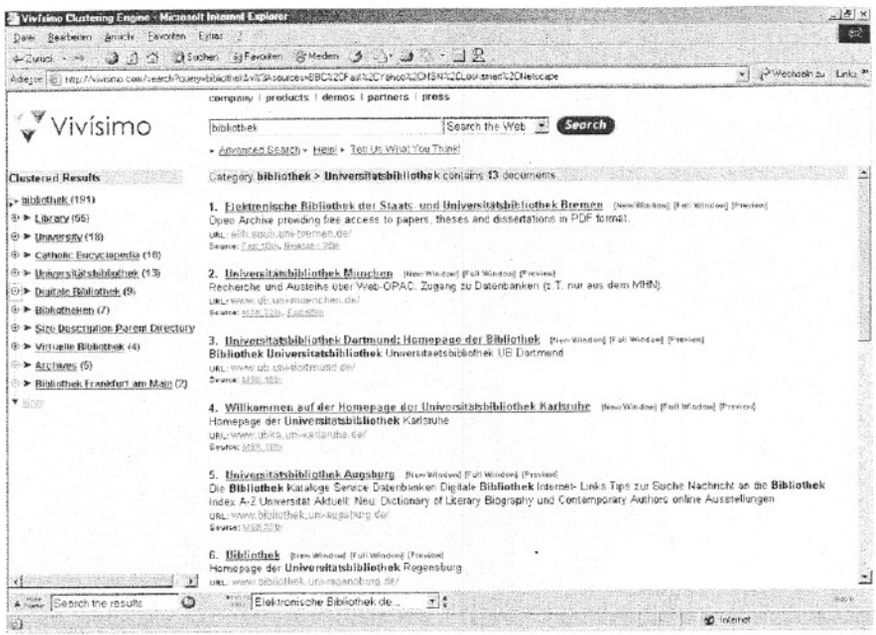

Abbildung 2. Ergebniscluster bei *Vivisimo*

Indexgrößen

Im Lauf des letzten Jahres haben einige der größten Suchmaschinen ihre Datenbanken weiter ausgebaut. Einen Überraschungserfolg erzielte die Suchmaschine *Alltheweb* im Juni 2002, als sie ankündigen konnte, nun mit über zwei Milliarden indexierten Dokumenten den noch vor *Google* größten Datenbestand anbieten zu können.

Angesichts dieser hohen Dokumentenzahl und dem weiterhin größer werdenden Abstand zu den Konkurrenten werden diese beiden Suchmaschinen zu einem Muss für den seriösen Benutzer. Metasuchmaschinen erreichen oft auch durch Kumulation der Datenbestände mehrerer Suchmaschinen nicht die Indexgröße von *Alltheweb* oder *Google*.

Weiterhin gültig bleiben die Untersuchungen von *Lawrence* und *Giles*[2], wonach unterschiedliche Suchmaschinen im Prozess der Suche nach Dokumenten keine deckungsgleichen Mengen aufstöbern, sondern im Gegenteil die Überschneidungen zwischen den Datenbeständen erstaunlich gering sind.

Paid Inclusions, Paid Listings

Das klassische Geschäftmodell der Suchmaschinen war es, sich über Werbeeinnahmen zu finanzieren. Die Werbung wurde auf den Ergebnisseiten in Form von Bannern angezeigt. Nachdem der Umsatz mit Bannerwerbung – nicht nur bei den Suchmaschinen – signifikant zurückging, waren neue Einnahmequellen vonnöten. Inzwischen geben alle Suchmaschinen mit eigener Technologie diese für Suchfunktionen in Intranets oder zur Suche innerhalb einer Firmen-Website in Lizenz. Eine weitere – für den Benutzer wesentlich folgenreichere – Tendenz sind die so genannte *Paid Inclusions* oder *Paid Listings*.

Dabei handelt es sich einerseits um die Aufnahme von Seiten in einen Suchmaschinen-Index gegen Bezahlung (*Paid Inclusion*). Den Anfang machte hier Yahoo, die schon vor einiger Zeit anfingen, von Gewerbetreibenden für eine garantierte Überprüfung ihres gewünschten Eintrags eine Gebühr zu verlangen. Andere Suchmaschinen garantieren gegen Bezahlung die vollständige Aufnahme der gesamten Webpräsenz und / oder die regelmäßige Überprüfung der Seiten in kurzen Intervallen.

Bei den *Paid Listings* handelt es sich um bezahlten Werbeplatz innerhalb oder noch vor den eigentlichen Trefferlisten. Mittlerweile zeigen die meisten Suchmaschinen bezahlte Treffer vor der eigentlichen Trefferliste an. Daran ist an sich

Abbildung 1. Ergebnispräsentation bei *Teoma*

Dirk Lewandowski, geboren 1973; Abschluss an der HBI Stuttgart 1997. Magisterabschluss an der Uni Düsseldorf in Philosophie und Informationswissenschaft, 2001. Momentane Tätigkeit: Projektleiter Datenbanken und Research bei der NRW Medien GmbH, Düsseldorf; Lehrbeauftragter an der Uni Düsseldorf. – Privatanschrift: Mühltaler Straße 24, 40221 Düsseldorf, E-Mail dirk.lewandowski@uni-duesseldorf.de

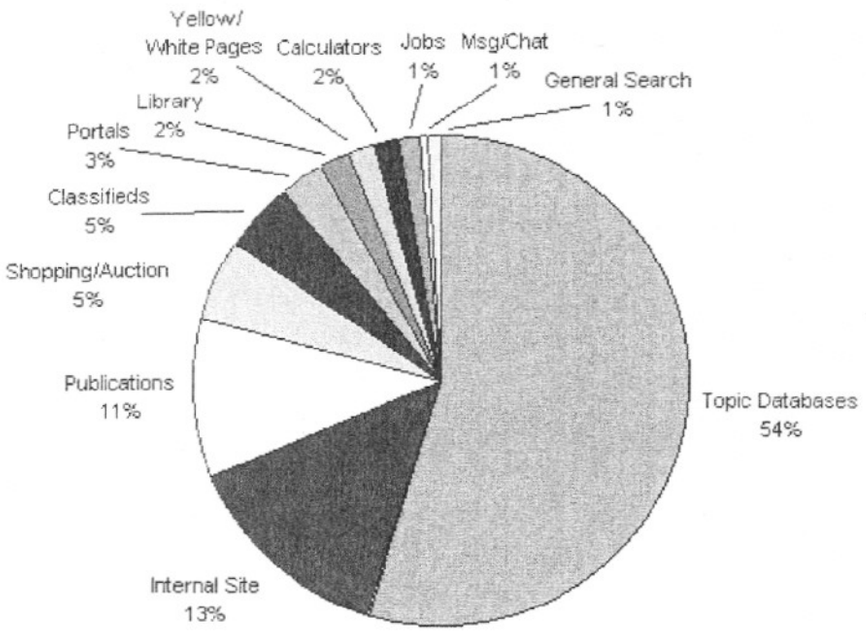

Abbildung 3. Inhalte des »Invisible Web« (*http://www.brightplanet.com/deepcontent/tutorials/DeepWeb/index.asp*)

nichts auszusetzen, solange der bezahlte Werbeplatz deutlich von den eigentlichen Suchergebnissen unterschieden ist, beispielsweise durch farbliche Unterlegung (wie bei *Google*). Leider häufen sich aber die negativen Beispiele wie etwa bei *Alta Vista*, wo sich die eigentliche Trefferliste (»AltaVista-Ergebnisse«) zwischen mehreren Bereichen von bezahlten Ergebnissen befindet, wobei sich für den Benutzer nicht in allen Fällen klar ergibt, was nun Werbung und was tatsächliche Treffer oder redaktionelle Empfehlungen sind. So handelt es sich sowohl bei »sponsored Listings«, »relevanten Ergebnissen«, »weiteren Listings«, »empfohlenen Links« und den »weiteren Angebote« um bezahlten Werbeplatz.

Noch weiter geht die *T-Online*-Suche, bei der vor der Trefferliste erst einmal sämtliche bezahlten Links angezeigt werden. Die Kennzeichnung der Werbung ist als mangelhaft zu bezeichnen. In den USA hat die *Federal Trade Commission* die Suchmaschinen-Betreiber aufgefordert, Werbung deutlicher als bisher zu kennzeichnen.

Die Konsequenz aus den aufgezeigten Entwicklungen sollte sein, sich primär den Suchmaschinen zuzuwenden, deren Ergebnisse nicht mit Werbung vermischt sind und deren Anzeigen deutlich gekennzeichnet sind. Besonders Portale wie *T-Online, Netscape, MSN* und so weiter sind zu vermeiden, da sie den Werbenachteil auch nicht durch bessere oder wenigstens andere Suchergebnisse ausgleichen. Denn die angezeigten Ergebnisse kommen sowieso von Anbietern wie *FAST* oder *Google*, deren Datenbanken auch direkt zugänglich sind.

»The Invisible Web«

Nicht nur die stark ansteigende Zahl von Webseiten macht den Suchmaschinen zu schaffen, sondern erst recht ein weiterer Bereich, der in den Ansätzen der konventionellen Suchmaschinen vernachlässigt wird: das »invisible Web«, also das »unsichtbare Netz«.

Damit ist jener Teil des Internet gemeint, der durch Suchmaschinen nicht erschlossen werden kann, also zum Beispiel die Titelaufnahmen in einem Opac. Zwar kann der Suchroboter die Einstiegsseite des Katalogs finden, eine Recherche in der dahinter liegenden Datenbank ist ihm jedoch nicht möglich.

Sherman und *Price* definieren das invisible Web wie folgt: »Text pages, files, or other often high-quality authoritative information available via the World Wide Web that general-purpose search engines cannot, due to technical limitations, or will not, due to deliberate choice, add to their indices of Web pages[3].«

Der erste Teil dieser Definition betont die oft hohe Qualität solcher Ressourcen. Es kann davon ausgegangen werden, dass es sich erst ab einer gewissen Datenmenge lohnt, diese in einer Datenbank statt durch konventionelle HTML-Seiten zu erfassen. Um aber eine größere Datenmenge zu verwalten, bedarf es Zeit und Personal. Wer diese investiert, wird sich auch um die Qualität seiner Daten bemühen. Entweder er verkauft nun die Daten (die damit für Suchmaschinen von vornherein verloren sind, da sie geschützt werden) oder aber er bietet sie mittels einer Datenbankschnittstelle kostenlos im Web an. Oft handelt es sich bei solchen Anbietern um staatliche Stellen, die qua ihres Auftrags ihre Daten kostenlos zugänglich machen. Daher ist zumindest in vielen Fällen von einer hohen Qualität der Invisible-Web-Ressourcen auszugehen.

So etwa bei Bibliothekskatalogen: Die Datensätze werden durch Experten erstellt und in die Datenbank eingegeben. Die Finanzierung erfolgt aus öffentlichen Mitteln, die sowohl Kontinuität als auch Qualität gewährleisten.

Eine wesentliche Unterscheidung zwischen sichtbarem und unsichtbarem Web steckt in der Festlegung, dass die Inhalte des invisible web »via the WWW« (über das WWW) zugänglich sind. Normale HTML-Seiten dagegen sind im Web zugänglich (zu den Unterschieden vergleiche die Tabelle unten).

Der zweite Teil der Definition bezieht sich auf Erschließung durch Suchmaschinen: entweder diese indexieren entsprechende Dokumente nicht oder aber sie *können* sie nicht indexieren. Neben dem oben genannten Hauptgrund (der Inhalt ist Bestandteil einer Datenbank) gibt es weitere technische Gründe für das Ignorieren bestimmter Seiten:

Unterschiede zwischen »sichtbarem« und »unsichtbarem« Web

On the Web	Via the Web
Anyone with server access can place just about anything »on« the internet in the form of a Web page	Various databases, various providers, material not directly searchable via Web search tools
Very limited bibliography control, no language control	Typically highly structured and well indexed
Quality of info extremely varied	Uniformly high quality, often professional ressources
Cost is low or free	Invisible Web often low-cost or free: proprietary information services cost can vary, often expensive

- Kein Link zeigt auf die entsprechende Seite. Die Suchmaschine kann nicht wissen, dass es diese Seite überhaupt gibt.
- Die Seite besteht hauptsächlich aus Bildern, Musik oder Videos. Die Suchmaschine kann die Seite zwar finden und theoretisch auch in ihre Datenbank aufnehmen, sie kann jedoch nicht erkennen, um was es sich inhaltlich handelt; dazu fehlen erläuternde textuelle Informationen.
- Bei so genanntem *Real time content*, also Informationen, die sich stetig ändern (Börsenkurse, Staumeldungen), ist eine Indexierung sinnlos. Viele Suchmaschinen verzichten deshalb ganz auf Seiten solchen Inhalts.
- Dynamisch generierte Seiten bergen wegen ihrer Erzeugung durch serverseitige Skripte eine Gefahr für Suchmaschinen (»spider traps«, Endlosschleifen für Suchroboter). Die meisten Suchmaschinen verzichten deshalb vollständig auf die Erfassung dieser Seiten.

Um die Bedeutung des *Invisible Web* zu verdeutlichen, sei auf dessen Größe verwiesen: Die Firma *Bright Planet*, selbst ein Anbieter von Suchtechnologie für das *Invisible Web*, schätzt dessen Umfang auf etwa das 400–550fache des sichtbaren Web. Diese Zahlen mögen zu hoch liegen, verdeutlichen jedoch die bisher weitgehend vernachlässigte Informationsfülle.

Welche Arten von Informationen typischerweise im *Invisible Web* vorhanden sind, zeigt Abbildung 3 (Seite 560). Den wesentlichen Teil machen themenspezifische Datenbanken, seiteninterne Datenbanken (wie zum Beispiel die *Knowledge Base* auf der Microsoft-Site) und Artikeldatenbanken von Zeitungen und Zeitschriften aus. Gerade diese qualitativ hochwertigen Informationen machen die Bedeutung des *Invisible Web* aus. Informationsbeschaffer und -vermittler können hier Quellen benutzen und weiterempfehlen, die dem ungeübten Benutzer verschlossen bleiben.

Wie aber sind Informationen aus dem nicht sichtbaren Teil des Internet zu finden? Zwar bieten einige Anbieter schon Suchmaschinen an, die Teile des *Invisible Web* erfassen, von Vollständigkeit kann hier aber keine Rede sein. Bis entsprechende Werkzeuge zur Verfügung stehen, erscheint es angebracht, weiterhin nach den Quellen und nicht nach den Inhalten dieser Quellen zu suchen. Konkret bedeutet dies, nicht in einer Suchmaschine nach Literaturangaben zu Aufsätzen über ein bestimmtes Thema zu suchen, sondern lieber nach einer Datenbank, in der solche Aufsätze erschlossen werden.

Fazit

In das Suchmaschinen-Angebot ist wieder Bewegung geraten. Einige Anbieter mussten zwar aufgeben, ihre Ansätze wurden jedoch von neuen Websites aufgenommen und fortgeführt. Bedenklich ist die Vermischung von Suchergebnissen und Werbung, wie sie mittlerweile von den meisten Anbietern praktiziert wird. Aus diesem Grund empfiehlt sich die Konzentration auf wenige Suchmaschinen.

Im Bereich des *Invisible Web* bleibt abzuwarten, ob es gelingt, Ergebnisse aus diesem weitgehend unerschlossenen Bereich des WWW in Trefferlisten »normaler« Suchmaschinen zu integrieren. Bis dahin empfiehlt es sich, direkt nach Einstiegsseiten wertvoller Quellen zu suchen. Für Informationsvermittler bietet sich hier eine Möglichkeit, gegenüber dem Benutzer die eigene Kompetenz zu zeigen – denn »in *Google* suchen kann jeder«. ◁

3 Chris Sherman und Gary Price: The Invisible Web: uncovering information sources search engines can't see. Medford, 2001. S. 57