Reisen ohne Karte: wie funktionieren Suchmaschinen?

Baumgärtel, Tilman

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
SSG Sozialwissenschaften, USB Köln

Schriftenreihe der Abteilung
"Organisation und Technikgenese" des
Forschungsschwerpunkts Technik-Arbeit-Umwelt am WZB

FS II 98-104

Reisen ohne Karte

Wie funktionieren Suchmaschinen?

Von Tilman Baumgärtel
tilman@icf.de

PROJEKTGRUPPE
KULTURRAUM INTERNET

Institut für Sozialwissenschaften
Fachbereich Umwelt und Gesellschaft, TU Berlin

und

Wissenschaftszentrum Berlin für Sozialforschung GmbH (WZB)
Reichpietschufer 50, 10785 Berlin
Telefon (030) 254 91 - 0, Fax (030) 254 91 - 684

Zusammenfassung

Suchmaschinen sind die Eingangstore zum Internet. Doch obwohl Netz-User die Suchmaschinen bei fast jeder „Surftour" benutzen, wissen nur die wenigsten ihrer Nutzer, wie sie funktionieren. Dieses Papier geht der Frage nach, ob die zentrale Position von Suchmaschinen im Netz auch zu einer Zentralisierung des Internet fuhren kann. Durch eine detaillierte Darstellung der Funktionsweise der Search Engines wird gezeigt, daß sie nicht pauschal als Instanzen betrachtet werden können, die dem „unstrukturierten", „hierarchiefreien" Internet ein Zentrum und eine Hierarchie aufzwingen. Technische Einzelaspekte erscheinen jedoch trotzdem problematisch: Besonders Suchmaschinen, die - wie **Lycos** - die angebliche Beliebtheit einer Site zum Maßstab ihrer Beurteilung machen, tragen dazu bei, im Netz eine hierarchische Gliederung zu voranzutreiben.

Abstract

Search Engines are the entrance doors to the internet. But despite the fact that net surfer use them almost every time they go online, most of them don't know how they actually work. This discussion paper addresses the question, whether the central position of search engines can lead to a centralization of the internet. A detailed technical analysis of the search engines leads to the conclusion that they cannot simply be considered as the main authorities that impose a center and a hierarchy on the supposedly „unstructured", „non-hierarchical" Internet. Yet some technical aspects appear to be problematic: especially search engines which use the assumed popularity of a site as a standard, are to some extent responsible for the introduction of a hierarchical structure on the internet.

Inhalt

1. Einleitung

Als der Abt Hugues de Saint-Cher 1240 das erste Stichwortverzeichnis der Bibel aufstellen ließ, waren mit dieser Aufgabe 500 Mönche beschäftigt. Dabei hat die Bibel in der heute üblichen Druckfassung nur etwa 800 Seiten. Das sind - in Computereinheiten - knapp fünf Megabyte.

Das WorldWideWeb, so wird geschätzt, enthält zur Zeit etwa 50 Millionen Seiten oder 400 Gigabyte, und es wächst monatlich um ungefähr 20 Prozent.[1] Mit der Aufgabe, diese Informationsflut zu erfassen und durchsuchbar zu machen, wären heute wohl sämtliche Mönche der Welt überfordert.

Für diese Aufgabe gibt es darum Suchmaschinen, die seit 1993 das immer unübersichtlicher werdende Internet durchkämmen. Wegen der dezentralen Struktur des Internets kann es keinen zentralen Index oder ein Gesamtverzeichnis des Netzes geben. Das Projekt der Suchmaschinen ist es darum, das gesamte WorldWideWeb zu erfassen und über Stichwörter durchsuchbar zu machen. In diesem Aufsatz soll erklärt werden, wie Suchmaschinen funktionieren, wie ihre technische Infrastruktur aussieht, und welche Bedeutung ihnen bei der inhaltlichen Strukturierung des Internet zu kommt.[2]

1.1. Die Bedeutung der Suchmaschinen

Ohne die Suchmaschinen wären die „Surftouren" durch das Internet wie Reisen ohne Karte. Suchmaschinen sind sozusagen die „Eingangstore zum Internet". Wer nicht genau weiß, wo er in den unübersichtlichen Informationsweiten des „Netz der Netze" eine bestimmte Information oder einen gewünschten Text findet, geht in der Regel über eine Suchmaschine seiner Wahl - und das sind die meisten User des Internets: Wie verschiedene Studien gezeigt haben, werden fast 90 Prozent aller Websites über die Search Engines gefunden und nicht über Werbung in Printmedien oder Fernsehen, die „Bannerwerbung" auf anderen Websites, Mundpropaganda oder andere Informationswege.

Die Suchmaschinen sind aber nicht einfach nur wichtige Anlaufpunkte im Internet, sie sind vielmehr die zentralen Knotenpunkte des angeblich dezentralen, hierarchiefreien Internets, an denen die Aufmerksamkeit der Netznutzer „umgeschlagen" und „weiterverteilt" wird. Den Suchmaschinen kommt damit eine entscheidende Rolle in der „Ökonomie der Aufmerksamkeit" (Georg Franck) des Internets zu (http ://www.t0.or.at/franck/gfeconom.htm).

[1] Angaben nach: Steinberg, Steve: Seek and Ye Shall Find (Maybe), Wired 5 1996, S. 108 -114,172 - 182. Bei der unübersichtlichen Natur des Internets sind solche Zahlen freilich mit allergrößter Vorsicht zu genießen.

[2] Der vorliegende Text entstand im Rahmen des von der Volkswagen-Stiftung geförderten Projektes „Kulturraum Internet, Netzkultur und Netzwerkorganisation in offenen Datennetzen" (Helmers u.a. 1996).

Auch mit Suchmaschinen kann die Recherche nach einer bestimmten Information oft mühsam oder unfruchtbar sein: **Vista** oder **Hotbot** geben bei gängigen Suchworten manchmal Millionen von „Treffern" an, die Ergebnisse scheinen jedoch oft beliebig oder unverständlich.

Das Vorhaben, das Internet zu katalogisieren und durchsuchbar zu machen, ist freilich immer nur in Annäherungen zu realisieren. Das Netz wächst ununterbrochen und fast unkontrollierbar. Es enthält heute Daten (und vor allem: Dateiformate!), an die niemand gedacht hatte, als es in den siebziger Jahren als militärisch-akademisches Netzwerk entstand. Auch die Suchmaschinen sind in gewissem Sinne noch auf dieses „klassische", textlastige Internet ausgerichtet, d.h. sie suchen im immer multimedialer werdenden Netz fast ausschließlich nach Geschriebenem. 1994 und 1995, als die „klassischen" Suchmaschinen wie **Alta Vista, Lycos** oder **Excite** entstanden, war noch nicht abzusehen, daß zum im Internet gesammelten „Weltwissen" schon bald auch Live-Videoaufnahmen von Kaffeemaschinen oder Galerien von Navigationsbuttons gehören würden.

Um es gleich vorwegzunehmen: keine der existierenden Suchmaschinen kann wirklich von sich sagen, daß sie ein vollständiger Index des Internet wäre, oder gar Volltext-Suchen des Netzes bieten könne, auch wenn einige Anbieter - zum Beispiel - **Vista** sich schon zu dieser Behauptung verstiegen haben. Im Gegenteil: wie Experimente mit den Suchmaschinen und selbstgeschriebenen Seiten gezeigt haben, schließt z.B. **Alta Vista** Seiten von bestimmten Servern (zum Beispiel von Geocities, CompuServe oder Tripod, bei denen man sich kostenlos eine eigene Homepage anlegen kann) gezielt aus.

1.2. Wozu dieses Papier?

Obwohl Netz-User die Suchmaschinen bei fast jeder „Surftour" benutzen, wissen nur die wenigsten ihrer Nutzer, wie sie funktionieren. Suchmaschinen sind technologische „Black Boxes", die zwar eine sehr wichtige Aufgabe ausführen, deren Funktionsweise aber keineswegs transparent ist. Einige der Firmen, die Suchmaschinen anbieten, sind sogar regelrecht darum bemüht, die Funktionsweise ihrer Angebote als „Geschäftsgeheimnisse" zu schützen. Wer es "von außen", d.h. als Nutzer genauer wissen will, wie die Suchmaschinen funktionieren, ist daher oft auf Experimente mit der "Trial & Error"-Methode angewiesen.

Die undurchsichtige Situation sollte eigentlich mißtrauisch machen: Nach welchen Kriterien werden die Ergebnisse, die eine Stichwortsuche mit einer Suchmaschine bringt, eigentlich geordnet? Können die Schöpfer einer Homepage Einfluß darauf nehmen, an welcher Stelle eine Site bei einer Suchmaschine angezeigt wird? Können Firmen sich in die „Hit"-Hierarchien einkaufen, so daß zum Beispiel die Homepage eines Autoherstellers immer an erster Stelle erscheint, wenn jemand als Suchwort „Car" oder „Automobile" eingibt?

Wie Hartmut Winkler in seinem Aufsatz „Suchmaschinen - Metamedien im Internet" (http://www.heise.de/tp/deutsch/inhalt/te/1135/Lhtml) schreibt, „bedeutet die

Häufung der Zugriffe einen signifikanten Umbau in der Gesamtarchitektur des Netzes.» Für das Netz bedeuten 12 Millionen Zugriffe am Tag einen Schub in Richtung Zentralisierung. Dies müßte all diejenigen hellhörig machen, die gerade den dezentralen und antihierarchischen Charakter des Netzes hervorgehoben haben und seine allgemeine Zugänglichkeit mit weitreichenden basisdemokratischen Hoffnungen verbinden." Dem wäre hinzuzufügen, daß auch die Platzierung von Dokumenten in den „Rankings" der Suchmaschinen die (Un-)Sichtbarkeit von Informationen im Internet beeinflußt.

Obwohl Suchmaschinen im Netz eine so zentrale Rolle einnehmen, ist die Literatur zu diesem Thema bisher auffallend dünn. Gedruckte Literatur gibt es jenseits von einigen oberflächlichen Artikeln in Computerzeitschriften so gut wie gar nicht; die Newsgroup „comp.infosystems.search", die sich mit Suchmaschinen beschäftigen soll, wurde erst im September 1997 eingerichtet und hat sehr wenige Postings; auf der „Robot-Mailingliste" (http://info.webcrawler.com/mailing-lists/robots/), bei der unter anderem über die Crawler von Suchmaschinen diskutiert wurde, findet fast keine Diskussion mehr statt.

Lediglich einige WWW-Sites bieten Informationen zum Thema „Suchmaschinen" an; fast alle sind freilich in Englisch und zum Teil offensichtlich von Geschäftemachern betrieben, die aus ihrem - oft bescheidenen - Wissen über Suchmaschinen Kapital schlagen wollen, und darum häufig mit unhaltbaren Behauptungen (z.B. „We can put you on top of any search engine for $ 250!") auftreten. Erst im Herbst 1997 sind in den USA zwei Bücher zum Thema erschienen, die allerdings eher Serviceleistung für Netz-User denn kritische Analyse der technischen Dispositive der search engines sind.

Um dieses Informationsdefizit zu beenden, soll das vorliegende Diskussionspapier erklären, wie Suchmaschinen funktionieren. Da die gedruckt vorliegende Literatur zu diesem Thema spärlich ist, mußte ich mich vor allem an „Online-Informationen" halten. Dazu gehören neben einigen Aufsätzen und speziellen, auf Suchmaschinen spezialisierten Websites unter anderem auch persönliche Auskünfte von Suchmaschinen-Betreibern. Besonders Louis Monier von **AltaVista** bin ich für schnelle, präzise Informationen per Email zu Dank verpflichtet. Außerdem habe ich eine eigene Homepage erstellt, die ich bei Suchmaschinen angemeldet habe, um zu prüfen, wie diese sie registrieren.

Ausgangspunkt der Arbeit ist die Frage, ob Suchmaschinen zu einer Hierarchisierung und Zentralisierung des Internets beitragen. Denn Suchmaschinen sind keine neutrale Technologie. In diesem Aufsatz soll gezeigt werden, wie die Search Engines aufgebaut sind, bevor am Beispiel von **AltaVista** die technischen Details einer besonders bekannten Suchmaschine erläutert werden sollen. Im Sinne einer thematischen Zuspitzungen werden einige Aspekte des Themas nicht berücksichtigt: Semantisch operierende Suchmaschinen werden in diesem Papier ebenso wenig behandelt wie „Meta-Suchmaschinen" oder neuere Angebote wie der Nachrichtensuchdienst „Northern Lights".

Die These dieser Arbeit lautet, daß Suchmaschinen zwar einerseits tatsächlich zu einer gewissen Hierarchisierung des Internets beitragen, andererseits aber sowohl die User der Suchmaschinen wie die Anbieter von content eine Reihe von - selten

ausgeschöpften - Möglichkeiten haben, Einfluß auf die Funktion der search engines zu nehmen.

Wie bei vielen Fragen, die die technische Infrastruktur des Internets aufwirft, steckt auch bei den Suchmaschinen der Teufel im Detail. Ein großer Teil dieser Arbeit besteht in der technischen Beschreibung und Analyse der Suchmaschinen. Mit dieser detaillierten Darstellung von Funktionsweisen der Search Engines will ich zeigen, daß es zu einfach wäre, sie pauschal als die Instanzen zu verurteilen, die dem angeblich „unstrukturierten", hierarchie-freien Internet ein Zentrum und eine Hierarchie aufzwängen.

Suchmaschinen tragen allerdings durch ihre Positionierung als „Eingangstore" zum Internet zu einer Zentralisierung des Netzes bei, indem sie ein „Streben zum Zentrum" verstärken, das dem Hypertext-System intrinsisch ist. Besonders Suchmaschinen, die - wie zum Beispiel Lycos - die angebliche Beliebtheit einer Site zum Maßstab ihrer Beurteilung machen, tragen dazu bei, im Netz eine hierarchische Gliederung voranzutreiben.

Besonderes Interesse verdient in diesem Zusammenhang die wirtschaftliche Situation der Suchmaschinen. Die meisten User der Search Engines wissen wenig oder nichts über die Unternehmen, die diese anbieten, und die „strategischen Allianzen", die viele von ihnen inzwischen mit großen Softwareherstellern oder Content-Anbietern eingegangen sind. Die Suchmaschinen befinden sich in der Hand von Unternehmen, über deren Besitzverhältnisse und Geschäftspolitik meist ebensowenig bekannt ist wie über ihre Software. Der Verdacht liegt nahe, daß diese „strategischen Allianzen" auch zu einer Bevorzugung der Geschäftspartner durch die Suchmaschinenanbieter führen könnte. Dieser Frage soll in einem eigenen Kapitel nachgegangen werden.

Zuvor will ich kurz die Entwicklung der Suchmaschinen beschreiben, weil ihre historische Entstehung auch ihre technischen Dispositive bestimmt. Anschließend werde ich technisch definieren, was eine Suchmaschine eigentlich ist (und, um es gleich vorwegzunehmen: Yahoo! ist keine), bevor ich auf den technischen Aufbau der Suchmaschinen eingehe.

1.3. Eine kurze Geschichte der Suchmaschinen

Der erste Versuch ein „Inhaltsverzeichnis" des Internets zu erstellen, hieß **Archie.**[3] **Archie** bestand aus einem „Datensammler" *(data gatherer),* der automatisch die Inhaltsverzeichnisse von anonymen FTP-Servern durchsuchte, und einem Retrievalsystem, mit dem die User mit Suchworten nach FTP-Dateien suchen konnten. Der Suchdienst, der 1990 an der McGill University in Kanada entwickelt worden war, gehört spätestens ab 1992 zu den geläufigsten Internet-„Tools". Damals mußten die User zu **Archie** „telnetten", inzwischen ist der Suchdienst auch im WWW und kann dort mit „Forms" bedient werden, (http://www-ns.rutgers.edu/htbin/archie)

[3] Der Name kommt von dem Wort „Archive", bei dem das „v" weggelassen wurde.

Archie war als Suchwerkzeug für FTP-Dateien so erfolgreich, daß er die Mitarbeiter des Rechenzentrums der University of Nevada in Reno 1992 dazu inspirierte, einen ähnlichen Index für Gopher Menues zu entwickeln, das den Namen Veronica [4] trug. Veronica ähnelt in vieler Hinsicht schon den heute gängigen, kommerziellen Suchmaschinen: Das Programm durchsuchte im Monatsrhythmus alle Gopher Menues, die beim „Mother Gopher" an der University of Minnesota angemeldet waren. Suchen konnten mit Hilfe von Bool'schen Operatoren, also dem gleichen AND, OR und NOT, die auch heute noch bei den meisten Suchmaschinen wie **Alta Vista** eingesetzt werden. Und obwohl die Zahl der zu untersuchenden Dokumente für heutige Verhältnisse einigermaßen überschaubar war (im November 1994 verzeichnete **Veronica** 15 Millionen Dokumente aus Gopher-, FTP- und HTML-Space), wurde schon damals beklagt, daß man als User der unüberschaubaren Zahl von Dokumenten und den Suchmethoden von Veronica hilflos ausgeliefert sei.

Der erste Such-Robot für das gerade neu entstehende WorldWideWeb war der **Worldwide Web Wanderer** (http://www.mit.edu/people/mkgray/net/background.html), der von dem MIT-Studenten Mathew Gray im Frühjahr 1993 programmiert wurde. Ursprünglich zählte **The Wanderer** nur WWW-Server, aber einige Monate später fügte Michael L. Maudlin ein „Retrieval Program" namens „**Wandex**" hinzu, um die gesammelten Daten durchsuchen zu können. (Mauldin, ein Computerwissenschaftler an der Carnegie Mellon University, entwickelte übrigens später die Suchmaschine „Lycos" und ist heute „Chief Scientist" bei der Suchmaschine, die inzwischen vom Universitäts-Forschungsprojekt zu einem kommerziellen Unternehmen geworden ist.) **The Wanderer** durchsuchte und katalogisierte von Juni 1993 bis Januar 1996 zweimal pro Jahr das Netz.

Im Oktober 1993 wurde **Aliweb** (http://www.nexor.com/aliweb/ kurz für: Archie-Like Indexing of the Web) entwickelt. **Aliweb** überließ einen Teil der Arbeit bei der Katalogisierung des Internets den Betreibern von WWW-Servern. Diese mußten für ihren Server einen Index erstellen, und dieses bei **Aliweb** anmelden. **Aliweb** selbst war lediglich ein in Perl geschriebenes *Retrieval System,* das die auf diese Weise zusammengestellten Indexe durchsuchte und sich bei seinen Suchen auf die Angaben der Server-Betreiber und der Autoren der Seiten verließ.

Im Dezember 1993 gingen fast gleichzeitig drei neue Suchmaschinen ans Netz: **Jumpstation, WorldWideWeb Worm und RBSE Spider. Jumpstation** und der **WorldWideWeb Worm** waren Suchroboter, die Websites nach Titel und Header **(Jumpstation)** beziehungsweise nach Titel und URL **(WorldWideWeb Worm)** indexierten. Wer mit diesen beiden Tools suchte, bekam eine Liste von „Hits" ohne weitere Bewertung in der Reihenfolge, in der sie in der Datenbank abgespeichert waren. Der **RBSE Spider** und der im April 1994 an der University of Washington gestartete **Webcrawler** (http://www.webcrawler.com/) waren die ersten Suchmaschinen, die nicht bloß eine Aufzählung von gefundenen Dokumenten lieferte, sondern diese auch nach einem „Ranking" sortierte.

[4] Anm: Veronica soll angeblich die Abkürzung von „Very easy rodent-oriented net-wide index to computerized archive" sein, zu deutsch etwa: „Sehr einfacher ungezieferartiger netzweiter Index zu computerisierten Archiven", Sehr viel naheliegender ist allerdings eine andere Erklärung des Namens, nach der Veronica die Freundin des amerikanischen Cartoon-Helden Archie war.

Webcrawler (http://www.webcrawler.com/) ist übrigens die einzige der bisher erwähnten Suchmaschinen, die bis heute überlebt hat, auch wenn sie inzwischen kein Uni-Projekt mehr ist, sondern von der konkurrierenden Suchmaschine **Excite** aufgekauft worden ist und inzwischen - wie „Magellan" - nur noch als „Marke im **Excite** Network" geführt wird. Weil der Traffic, den das beliebte Recherche-Werkzeug anzog, drohte, das Universitätsnetz lahmzulegen, verkaufte die University of Washington **Webcrawler** 1995 an den Onlinedienst America Online (AOL) (http://www.webcrawler.com/Help/AboutWC/WCStory.html). Im März zog **Webcrawler** in seine „neue Heimat" bei AOL um. Der Onlinedienst verkaufte die Suchmaschine im November 1996 wiederum an **Exite** weiter. Bis heute firmiert allerdings Brian Pinkerton (http://corp.excite.com/Company/bpinkerton.html), der den Such-Robot 1994 als Student in einem Informatik-Seminar an der University of Washington geschrieben hatte, bei **Excite** als „Vice President of Engineering".

Fast zur gleichen Zeit arbeitete an der Carnegie Mellon University Leonard Maudlin an einem Spider, der später unter dem Namen **Lycos**[5] (http://www.lycos.com) bekannt wurde. Im Mai 1994 begann er mit der Arbeit an dem Spider, dem er im Juli das Retrieval-System „Pursuit" hinzufügte. Wie **Webcrawler** listete auch **Lycos** seine Suchergebnisse nicht einfach nur auf, sondern sortierte sie nach ihrer Relevanz; im Gegensatz zu **Webcrawler** bewertete **Lycos** nicht nur die Häufigkeit eines Wortes in einem bestimmten Dokument, sondern auch die „word proximity" (die Nähe von mehreren Suchbegriffen zueinander). **Lycos** ging am 20. Juli 1994 online.

Wie viele Internet-Einrichtungen sind also auch die Suchmaschinen, die - wie **Lycos** und **Webcrawler** heute als kommerzielles Unternehmen betrieben werden - ein Ergebnis wissenschaftlicher Vorarbeiten an den Universitäten. Erst 1995, dem Jahr, als das Internet langsam das Bewußtsein einer nicht-akademischen Öffentlichkeit erreichte, gingen die ersten Suchmaschinen ans Netz, die von Unternehmen mit Gewinnabsicht entwickelt wurden: **Infoseek** (http://www.infoseek.com) startete Anfang 1995; **Architex,** heute unter dem Namen **Excite** (http://www.excite.com) bekannt, ging im Oktober 1995 online; **Alta Vista** (http://alta**Vista**.digital.com) startete im Dezember 1995 den regulären Betrieb. Während **Alta Vista** als Projekt des Western Research Lab, einer Forschungsabteilung der Computer-Firma Digital Equipment Corporation (DEC) entstand, war es von Anfang an das „business modell" von **Excite** und **Infoseek,** sich durch Anzeigen zu finanzieren. Inzwischen verkauft auch **Alta Vista** Bannerwerbung auf seinen Seiten. Bis heute kam eine Reihe von anderen kommerziellen Suchmaschinen dazu.

Gegenwärtig gehören Suchmaschinen zu den wenigen kommerziellen Angeboten im Internet, die wirklich Profite machen. Suchmaschinenanbieter wie **Infoseek** oder **Lycos** sind im vergangenem Jahr an die Börse gegangen, und während **Infoseek** noch keine schwarzen Zahlen schreibt, machte **Lycos** im dritten Quartal 1997 - nach einem Jahr an der Börse - erstmals Gewinne. Auch Unternehmen wie **Yahoo!** oder **Alta Vista** verzeichnen nach eigenen Angaben inzwischen gesunde Gewinne.

[5] Der Name Lycos kommt von dem lateinischen Namen der „Wolfsspinne": Lycosidae Lycosa. Die Wolfspinne fängt ihre Beute nicht in einem Netz, sondern geht selbst auf die Jagd.

Nicht umsonst hat Microsoft im Oktober 1997 angekündigt, mit einem Projekt namens „Yukon" in diesem vielversprechenden Markt mitverdienen zu wollen. Auch von anderen Anbietern werden fast wöchentlich neue Suchmaschinen in Betrieb genommen. Die meisten der neueren Suchmaschinen versuchen nicht mehr, das ganze Netz zu verzeichnen, sondern beschränken sich auf *Special Interest*-Themen oder Lokalisierungen.

So gibt es mit WWWomen (http://www.wwwomen.com/) eine Suchmaschine für Frauen-Sites; Scifisearch (http://www.scifisearch.com/) sucht nur nach Science Fiction und „paranormalen Phänomenen" und „Filez" (http://www.fdez.com/) nach Computerprogrammen. Auch für Länder und Regionen gibt es eigene Suchmaschinen. Für deutsche und deutschsprachige Sites sind es z.Z. mindestens 22 verschiedene Suchmaschinen und Directories, und es kommen immer noch neue dazu. Diese lokalisierten Maschinen suchen entweder nur innerhalb bestimmter Domains (wie z.B. das holländischen Search.nl (http://www.search.nl)), einige technisch avanciertem Modelle (wie z.B. **Fireball** (http://www.fireball.de) erkennen mit speziellen statistischen Methoden die Sprache, in der ein Dokument verfaßt ist, und können so auch deutschsprachige Dateien verzeichnen, die nicht „.de" (für Deutschland), „.ch" (für die Schweiz) oder „.at" (für Österreich) im Domain Namen haben.

Das „diversifizierteste" Beispiel für eine regionale Suchmaschine, das ich bei meiner Recherche entdeckt habe, ist „Mowhawk Valley Online" (http://www.mvsearch.corn), eine Suchmaschine, die nur nach Homepages aus dieser ländlichen Gegend im amerikanischen Bundesstaat New York sucht.

In Deutschland hat die Suchmaschine Sharelook (http://www.sharelook.de) inzwischen „Regionalausgaben" für die Städte Berlin (http://berlin.sharelook.de/), Bochum (http://bochum.sharelook.de/), Düsseldorf (http://duesseldorf.sharelook.de/), Hamburg (http://hamburg.sharelook.de/, Köln (http://koeln.sharelook.de/), München (http://muenchen.sharelook.de/), Zürich (http://zuerich.sharelook.ch) und Wien (http://wien.sharelook.at).

Fast alle „großen" Suchmaschinen bieten inzwischen über die reine Suchfunktion hinaus *value-added services* an: So liefern einige der Suchmaschinen zum Beispiel Verzeichnisse von Email-Adressen und Telefonnummern an; andere verwandeln sich in der letzten Zeit sogar zunehmend in Quasi-Online-Dienste, in denen (z.T. personalisierbare) Nachrichten, Kleinanzeigen, Stadtpläne, Wettervorhersagen oder Chatrooms angeboten werden, und bei denen man sich oft sogar eine eigene Email-Adresse einrichten lassen kann. Diese redaktionellen und Service-Angebote tragen wiederum zu einer weiteren Stärkung ihrer zentralen Position im Netz bei.

1.4. Was ist eine Suchmaschine?

Schon der Terminus „Suchmaschine" wird oft falsch benutzt, und sowohl für „echte" Suchmaschinen wie auch für reine Netzverzeichnisse (die sogenannten *directories* oder *Indices)* gebraucht. Der Unterschied zwischen diesen beiden Arten von Netzangeboten besteht darin, wie ihre Adressen-Listen zusammengestellt werden.

„Echte" Suchmaschinen wie **Alta Vista** (http://altaVista.digital.com/) oder **HotBot** (http://www.hotbot.com/) suchen sich ihre URLs selbständig zusammen, indem sie das Netz „durchwandern", und ihre „Fundstücke" dann in aufbereiteter Form ihren Usern zur Verfügung stellen.[6]

Zu den *Directories* gehören zum Beispiel Websites wie **Yahoo!** oder **Web.de,** die von einer - menschlichen - Redaktion zusammengestellt werden. Bei der deutschen Version von **Yahoo!** (http://www.yahoo.de) sind zum Beispiel vier „freiberufliche Netzsurfer" damit beschäftigt, im Internet interessante Sites zu suchen, oder diejenigen, die bei ihnen angemeldet werden, zu prüfen und in die Yahoo!-„Ontologie" einzufügen.

Eine Mischung aus Suchmaschine und Index sind die „Hybriden", die beide Funktionen miteinander kombinieren. So bietet beispielsweise **Excite** (http://www.excite.com) neben einer Suchfunktion ausgewählte Sites auch nach „Channels" sortiert an. Und auch **Yahoo!** lenkt seine User automatisch an eine interne Version von **Alta Vista** auf dem **Yahoo!-Server** weiter, wenn das gesuchte Stichwort nicht unter den Kategorien der eigenen Datenbank gefunden wurde.

Alle Suchmaschinen funktionieren grundsätzlich nach dem selben Prinzip, unterscheiden sich aber in signifikanten Details. Das macht es schwierig, abwechselnd mit verschiedenen Maschinen zu suchen: während eine Suchmaschine zum Beispiel Bool'sche Operatoren wie „AND" und „NOT" „versteht", muß man die nächste mit „+" oder „-" „füttern", um dieselben Funktionen auszuführen. Im nächsten Teil wird darum erklärt, welches die Bestandteile sind, aus denen sich alle Suchmaschinen zusammensetzten, bevor ich auf die Unterschiede zwischen den verschiedenen search engines eingehe.

1.5. Woraus besteht eine Suchmaschine?

Alle Suchmaschinen haben drei Elemente, zu denen als erstes der ***Spider*** gehört, der manchmal auch **„Crawler"** oder **„Robot"** genannt wird. Er „durchkriecht" auf der Suche nach Daten das Netz. Die beiden anderen Elemente sind der **Index** und das **Suchmaschinen-Interface.**

Der Spider „wandert" durch das Internet, und „sammelt" dabei Webpages, die er auf seinen Server überträgt. Dort werden sie in den Index eingefügt. Dann folgt er den Links auf der gefundenen Seite weiter zur nächsten Seite. Wegen dieser Funktionsweise braucht ein Spider keine lange Liste von URL's, um seine Suche zu beginnen. Eine einzige Seite mit Hotlinks, von denen er zu anderen Sites weitergeschickt wird, genügt.

Man kann das, was der Spider tut, darum mit der Aufgabe vergleichen, ein Telefonbuch zu schreiben, wenn man nur eine einzige Telefonnummer kennt: man müßte diese Telefonnummer anrufen, den Teilnehmer nach allen Telefonnummern, die er kennt, fragen, diese Nummern anrufen, bei diesen Teilnehmern wiederum alle bekannten

[6] Wegen dieser „Wanderungen" trugen die Suchmaschinen der ersten Generation oft Namen wie **The Wanderer,** später auch die Namen von Spinnenarten (z.B. Lycos oder **Inktomi).**

Telefonnummern erfragen, und so weiter. Theoretisch könnte man mit dieser Methode irgendwann alle Telefonnummern der Welt finden.

So, wie man mit dieser Methode freilich die Telefonnummern von jemand, der sie für sich behält, nicht herausbekommen würde, so findet der Spider von sich aus auch keine Page, die nicht mit anderen verlinkt ist. Anbieter, die nicht wollen, daß ihre Sites gespidert werden (zum Beispiel, weil sich der Inhalt regelmäßig ändert, oder weil sie nicht wollen, daß ihr Server von den Robot-Abfragen überlastet werden, kann das mit einer einfachen Test-Datei erreichen: „robots.txt".

1.5.1. Der Robot-Exclusion-Standard

Wenn ein Robot „gut erzogen" ist, hält er sich an den „Robots-Exclusion-Standard" (http://info.Webcrawler.com/mak/projects/robots/norobots.html). Dieser Standard ist kein „offizielles", von irgendeiner Internet-Institution entwickeltes Gesetz, sondern eine Übereinkunft, die die Mitglieder der „Robot"-Mailing-Liste (http://info.Webcrawler.com/mailing-lists/robots/info.html) am 30. Juni 1994 informell getroffen haben. So eine Übereinkunft war nötig geworden, weil sich 1993 und 1994 Situationen häuften, in denen Server durch ununterbrochene Spider-Zugriffe lahm gelegt wurden, wie es im „Robots Exclusion Standard" heißt:

> „In 1993 and 1994 there have been occasions where robots have visited WWW servers where they weren't welcome for various reasons. Sometimes these reasons were robot specific, e.g. certain robots swamped servers with rapid-fire requests, or retrieved the same files repeatedly. In other situations robots traversed parts of WWW servers that weren't suitable, e.g. very deep virtual trees, duplicated information, temporary information, or cgi-scripts with side-effects (such as voting).
>
> These incidents indicated the need for established mechanisms for WWW servers to indicate to robots which parts of their server should not be accessed. This standard addresses this need with an operational solution,"

Wenn auf einem Server so eine Datei mit Robot-Befehlen existiert, kann man sie finden, indem man an die URL den Dateinamen „robots.txt" anfügt, also z. B. http://www.domainname.com/robots.txt." Wer sich etwa die „robots.txt"-Datei von *Focus* (http://www.focus.de/robots.txt) ansieht, findet sehr detaillierte Anweisungen an den Spider, die ihn davon abhalten sollen, bei einem seiner Besuche den Inhalt von Seiten zu speichern, die regelmäßig aktualisiert werden oder nicht für die Öffentlichkeit bestimmt sind:

```
# robots.txt for http:www.focus.de
# Gibt an, welche Unterverzeichnisse nicht durch Crawler durchsucht werden sollen

User-agent: *
Disallow: /ERRORS/   # Fehler-
Seiten Disallow: /test/    # Test-
Seiten Disallow: /testl/   # Test-
Seiten
```

Disallow: /test2/ # Test-Seiten
Disallow: /test3/ # Test-Seiten
Disallow: /test4/ # Test-Seiten
Disallow: /test5/ # Test-Seiten
Disallow: /test6/ # Test-Seiten
Disallow: /Test/ # Test-Seiten
Disallow:/cgi-bin/ # Scripts
Disallow: /GLOBPICS/ # allg. Grafiken
Disallow:/G/GP/GPA/ # Politik-News
Disallow:/G/GV/GVA/ # Vermischtes-News
Disallow: /G/GS/GSA/ # Sport-News
Disallow: /G/GW/GWA/ # Wirtschaft-News
Disallow: /G/GN/GNA/ # Special-Event-News
Disallow: /G/GZ/GZA/ # Special-Event-News
Disallow:/G/GT/GTA/ # Special-Event Splitter
Disallow: /G/GX/GXA/ # Special-Event-News
Disallow: /DA/DANEWS/# News Finanzen
Disallow: /DB/DBNEWS/# News Job+Karriere
Disallow: /DC/DCNEWS/ # News Technik+PC
Disallow: /DD/DDNEWS/# News Medien+Netz
Disallow: /DI/DINEWS/ # News Politik
Disallow: /DJ/DJNEWS/# News Immobilien
Disallow: /A/a.htm # ehemals Willkommensseite
Disallow: /DG/ # ehemals FOCUS-TV
Disallow: /*GA/* # ehemals News
Disallow: /GB/ # ehemals News
Disallow: /DC/DCQ/ # ehemals Handy-Service
Disallow: /T/TE/TEE/ # ehemals FocusTV Rep.
Disallow: /T/TE/TEF/ # ehemals FocusTV Rep.
Disallow: /T/TE/TEG/ # ehemals FocusTV Rep.
Disallow: /T/TE/TEH/ # ehemals FocusTV Rep.
Disallow: /T/TE/TEJ/ # ehemals FocusTV Rep.
Disallow: /DC/DCU/ # ehemals Raubkopien
Disallow: /DD/DDE/ # ehemals Zahlenbingo im WWW
Disallow: /DD/DDS/ # ehemals Internet-Laeden
Disallow: /DD/DDL/ # ehemals Internetsucht
Disallow: /DA/DAI/ # ehemals Finanzen Immobilien Checklisten
Disallow: /DA/DAI/DAIA/ # ehemals Finanzen Immobilien Checklisten
Disallow: /DA/DAI/DAIB/ # ehemals Finanzen Immobilien Checklisten
Disallow: /DA/DAY/ # ehemals Kreditkarten
Disallow: /DA/DAN/ # ehemals Anlagetips

So ausführlich müssen die Befehle in „robots.txt" jedoch keinesfalls sein. Viele Websites (etwa die Site der Focus-Konkurrenz *Spiegel* (http://www.spiegel.de) haben überhaupt kein „robots.txt", und auch zum Beispiel bei Microsoft ist der Text recht sparsam:

 # robots.txt for http://www.microsoft.com/

 # do not delete this file, contact Mark Ingalls for edits ! ! ! !

User-agent: *
Disallow: /isapi/ # keep robots out of the executable tree
Disallow: /scripts/

Hat der Spider eine Seite und ihre URL einmal entdeckt, kehrt er in regelmäßigen Abständen zu ihr zurück, und prüft, ob sich etwas verändert hat Dabei sollte man sich die „Netzwanderungen" des Spiders nicht wie eine wirkliche, „räumliche" Reise durchs Netz vorstellen; der Spider ist vielmehr eine Art automatisierter Browser, der sich von einem zentralen Server aus selbständig durchs Netz klickt.

Meist durchstöbert mehr als nur ein Spider das Netz, bei *Lycos* sind es nach Firmenangaben z.B. 20 automatisierte Software-Programme, die sich gleichzeitig durchs Netz bewegen. Um den Server nicht zu überlasten, „klicken" sie sich dabei allerdings langsamer durch eine gefundene Site als es ihnen technisch möglich ist, d.h. nicht schneller als mit einem Seitenaufruf alle 30 Sekunden.

Ein entscheidender Unterschied zwischen den verschiedenen Spider-Programmen ist ihre Vorgehensweise, wenn sie eine Site gefunden haben: die sogenannte *traversal strategy.* [7] Hierbei wird unterschieden zwischen *Depth-first search, Breadth-first search* und *Random search.* Diese verschiedenen *traversal strategies* ergeben sich aus der Arbeitsweise des Spiders. Weil er jede Seite auf seinen Server lädt, um sie dort zu indexieren, aber nicht alle miteinander durch Links verbundenen Seiten auf einmal laden kann, muß er eine Art „Angriffsplan" entwickeln. Dabei geht er im Grunde genauso vor, wie es ein Mensch tun würde, der manuell Dokumente oder Websites durchsucht:

> „Die eine Methode ist es, mit einer beliebigen Homepage anzufangen, ihren Text aufzunehmen, und alle URL's, zu denen es von dieser Seite aus Links gibt, auf eine immer länger werdende Liste zu schreiben. Dann würde man dasselbe mit der ersten URL auf der Liste machen. Nachdem man diese URL aufgenommen hat, könnte man sie von der Liste streichen. Diese Art von Algorithmus nennt man eine *depth-first search,* weil man sich dabei weiter und weiter von der ursprünglichen URL entfernt, bis man zu einer Seite kommt, die keine Links hat. Dann kehrt der Spider zu der Ausgangsseite zurück und folgt von da aus der nächsten Link-Kette....
>
> Man kann an diesem Algorithmus auch eine Kleinigkeit ändern: neu entdeckte URL's werden nun an das Ende der Liste, und nicht mehr an ihren Anfang gesetzt. Diese simple

[7] Bei dieser Beschreibung beziehe ich mich auf: **Dreilinger, Daniel:** Internet Search Engines, Spiders, and Meta-Search Engines, in: **Williams, Joseph:** Bots and other Internet Beasties, Indianapolis 1996 (sams.net), S. 237 - 241

> Änderung nennt man *breadth-first search...* Der Effekt davon ist, daß man zuerst alle Seiten abruft, die einen Link von der Ausgangs-Homepage entfernt sind, dann alle Seiten, die zwei Links entfernt sind und so weiter."[8]

AltaVista arbeitet zum Beispiel mit einem *Breadth-first-System,* während Lycos nach dem *Depth-first-System* Daten sammelt. Der *Random Search* ist eine Kombination dieser beiden Systeme, bei dem der Robot nach einem Zufallsprinzip URLs von der Liste aufruft - mal von den vorderen, mal von den hinteren Rängen.

Mit allen drei Systemen müßte ein Spider auf seinen Reisen durch das Netz theoretisch früher oder später zu jeder verlinkten Seite kommen. In der Praxis kann es allerdings sehr lange dauern, bis ein Spider eine bestimmte Site entdeckt hat. Um das Programm schneller auf die eigene Homepage zu „locken", kann man darum bei vielen Suchmaschinen selbst URLs anmelden. Diese URL wird dann automatisch in die Liste der Seiten aufgenommen, die der Spider „besuchen" will. Bei Excite kann man sogar „in Echtzeit" zusehen, wie der Spider die eingegebene Site aufruft, und dann nach kurzer Wartezeit meldet, daß er die Seite binnen kurzem in seinen Index aufnehmen wird.

Dieser Index (oder Katalog) ist das zweite Element der Suchmaschine. An ihn liefert der Spider seine Suchergebnisse zurück. Bei manchen Suchmaschinen erhält der Index eine vollständige Kopie aller Seiten, die der Spider bei seinen Reisen durchs Netz gefunden hat (z.B. Alta Vista), andere speichern nur die ersten hundert Worte jeder gefundenen Page (z.B. Infoseek).

Wenn sich die Seite ändert, sollte der Spider das bei seinem nächsten Besuch merken, und die veränderten Informationen an den Index weitergeben. Allerdings kann zwischen dem Spiderbesuch auf der einen Seite und deren Indexierung eine gewisse Zeit vergehen. So lange eine Seite nicht im Index ist, kann sie über die Suchmaschine auch nicht gefunden werden: Wie lange es dauert, bis ein Spider eine URL besucht hat und diese anschließend im Index auftaucht, darüber gehen die Angaben von Anbietern und eigene Erfahrungen weit auseinander.

Während zum Beispiel AltaVista angibt, daß selbst angemeldete URLs in wenigen Tagen „gespidert" und dann beim nächsten Update im AltaVista-Index verzeichnet werden, zeigen einige eigene Experimente mit eingegebenen URLs, daß es manchmal mehrere Wochen dauern kann, bis diese von der Suchmaschine ausgegeben werden. Hat der AltaVista-Spider mit dem Namen „Scooter" eine URL allerdings einmal gefunden, kehrt er in regelmäßigen Abständen von vier bis sechs Wochen zu dieser zurück, um sie zu überprüfen.

Das Suchmaschinen-Interface ist das dritte Element der Suchmaschine. Wenn ein User eine Anfrage eingibt, arbeitet sich dieses Programm durch die Millionen von Seiten im Index der Maschine, sucht nach Treffern und gibt ihnen ein „Ranking". Dieses Programm entscheidet, wie und an welcher Stelle eine Seite bei einer Suche „ausgespuckt" wird. Wer verstanden hat, wie dieses Programm funktioniert, kann

[8] **Dreilinger, Daniel:** Internet Search Engines, Spiders, and Meta-Search Engines, in: **Williams, Joseph:** Bots and other Internet Beasties, Indianapolis 1996 (sams.net), S. 239f - meine Übersetzung

Suchmaschinen nicht nur selbst effektiver benutzten, sondern sich auch erklären, nach welchen Kriterien die oft willkürlich erscheinenden Bewertungen zustande kommen.

Um die Seiten entsprechend der User-Anfrage zu bewerten, bedienen sich die Suchmaschinen einer Reihe von Regeln. Am wichtigsten ist es für die Suchmaschine, an welcher Stelle sich der Suchbegriff oder die Suchbegriffe auf der Seite befinden und wie häufig diese vorkommen. Der wichtigste „Standort" für das Suchwort ist die Titelzeile. Wenn ein Suchwort zwischen den HTML-Tags <titel> und </titel> steht, dann betrachten die meisten Suchmaschine dieses Dokument als relevanter als eins, in dem das Suchwort erst im <body>-Text vorkommt. Auch wenn ein Suchwort im eigentlichen Text weit oben steht (z.B. in einer Überschrift oder im ersten Satz), bewertet die Suchmaschine dieses Dokument ebenfalls als relevanter als eine Datei, bei der das Suchwort erst später im Text erscheint.

Die Häufigkeit, mit der das Suchwort im Text vorkommt, ist ein anderer Faktor, von dem das Ranking abhängt: je häufiger ein Wort im Verhältnis zu den übrigen Worten auf einer Seite vorkommt, desto wichtiger muß es sein. **Das bedeutet freilich auch, daß die Suchmaschinen quasi „automatisch" kurze Texte für wichtiger ansehen als lange.** Eine Seite, auf der der Text nur aus dem Suchwort besteht, „zählt" mehr, als ein langer Text, auch wenn in ihm das Suchwort immer wieder vorkommt.

Nach diesem Prinzip arbeiten alle Suchmaschinen. Der Unterschied zwischen den verschiedenen Maschinen liegt in den Details. Das beginnt damit, daß einige Search Engines mehr Webseiten verzeichnen als andere. Einige Suchmaschinen indexieren diese Seiten auch häufiger als andere, weil ihr Spider häufiger die gesammelten Links überprüft. Deswegen sucht jede Suchmaschine in einer anderen Sammlung von Seiten als die Konkurrenz.

Bei einigen Suchmaschinen wird außerdem die „Beliebtheit" von Seiten in die Bewertung miteinbezogen: **WebCrawler** und **Lycos** prüfen anhand ihrer Datenbank auch, wie viele Links es auf diese Site gibt, und beziehen die Anzahl der Links in ihre Bewertung der Seite mit ein: je mehr Links es auf eine Seite gibt, desto „beliebter" und folglich „besser" muß sie auch sein. Wie ich zeigen werde, trägt dieser Mechanismus zu einer Hierarchisierung des Internet bei, weil sie Sites, zu denen viele Links führen, bevorzugt, und dadurch populäre Seiten noch populärer werden, während Sites, zu denen weniger Links führen, auch in der Hierarchie der Suchmaschine noch weiter nach unten rutschen.

Einige der Hybrid-Suchmaschinen beurteilen außerdem Seiten, die in ihrem von Redakteuren zusammengestellten Index stehen, als relevanter ein als andere Seiten - wenn eine Homepage gut genug ist, um ein Review zu bekommen, dann ist sie wahrscheinlich auch relevanter als eine Site, die nicht besprochen worden ist.

Viele Webdesigner glauben, daß die Angaben in den Meta-Tags die wichtigsten Angaben für eine Suchmaschine sind und daß man diese sogar regelrecht steuern könnte, wenn man in den Meta-Tags die richtigen Suchbegriffe versteckt hat Das stimmt nicht. **HotBot und Infoseek** bewerten die Stichworte in den Metatags einer Seite geringfügig höher als die Worte im Text. Aber **Excite** liest sie zum Beispiel gar nicht, und die

Erfahrung zeigt, daß auch HTML-Seiten, die überhaupt keine Meta-Tags haben, gut bei den Suchmaschinen platziert sein können.

In letzter Zeit sind einige Suchmaschinen dazu übergegangen, Sites für „Spamming" zu bestrafen, in dem sie diese ganz aus dem Index verbannen. Als Spamming betrachtet man es zum Beispiel bei **Vista,** wenn ein Wort auf einer Seite häufig wiederholt wird. Gerade die Anbieter von Sex-Sites verbergen in den Meta-Tags gerne Hunderte von Suchbegriffen. Einige Suchmaschinen-Anbieter haben darum ihre Index-Software so programmiert, daß sie Stichworte in den Metatags ignoriert, die öfter als dreimal vorkommen.

Als WWW-„Spam" gilt es aber vor allem, wenn man auf einer Site „unsichtbare" Worte unterbringt, indem diese mit der gleichen Farbe wie die Hintergrundfarbe geschrieben werden. Auf vielen Porno-Siten finden sich auf den Eröffnungsseiten die üblichen Wortketten („sex women porn" etc), die zum Beispiel in rosa auf einem rosa Hintergrund stehen. Während der normale Surfer diese Worte nicht sieht, liest die Suchmaschine sie wie normal sichtbaren Text.

Infoseek schließt Seiten, bei denen die Schriftfarbe die gleiche wie die Hintergrundfarbe ist, inzwischen aus dem Index aus; man wird erst wieder aufgenommen, wenn man die Seite umgestaltet und die Wiederaufnahme per Email „beantragt" hat, und die Seite von Mitarbeitern von **Infoseek** überprüft worden ist. Wer nochmal bei dem selben „Vergehen" ertappt wird, wird für immer aus dem Verzeichnis der Suchmaschine ausgeschlossen.

Zwei *features* von WWW-Sites machen den Robots der Suchmaschinen besondere Probleme: Frames und CGI-Code. Daten, die in Frames stehen, finden normale Suchroboter nicht. Wie ein Netzsurfer, der einen alten Mosaic-Browser benutzt, finden die Roboter nur Informationen, die nicht in Frames „verborgen" sind. Den Links, die zum Beispiel von einem "Inhaltsverzeichnis-Frame" in eine Site führen, können sie nicht folgen. Das bedeutet für die vielen WWW-Angebote, deren Inhalt nur in Frames untergebracht ist, daß der Robot von der gesamten Site nur folgende Worte speichert: „Sorry! You need a frames-compatible browser to view this site." Wer trotzdem will, daß der Text, der im Frame steht, gefunden werden soll, der muß den Text der Site nochmal zwischen den sogenannten <no frames>-Meta-Tags unterbringen.

Ein anderes Problem für die Robots der Suchmaschinen sind Common Gateway Interface(CGI)-Scripts, die verwendet werden, damit der User bei einer Site nach Daten suchen kann. Die gesuchten Daten werden dann auf „dynamischen Seiten" angezeigt, d.h. sie sind eigens auf die Anfrage hin generiert worden. Für Suchmaschinen sind diese *ad hoc* erstellten Dokumente, die nur wegen der Robot-Anfrage entstanden sind, uninteressant.

Um zu verhindern, daß der Robot tausende von Seiten sammelt, die er selbst generiert hat, hat zum Beispiel Louis Monier von **Vista** seinen Robot so programmiert, daß er keine URLs mit einem Fragezeichen einsammelt, weil eine mit einem CGI-Skript generierte Seite immer dieses Satzzeichen enthält: „a crude way of avoiding cgi scripts", wie er dieses Verfahren in einer Email an den Autor nannte. Durch

diese Funktion werden Suchmaschinen auch beispielsweise davon abgehalten, sich gegenseitig zu durchsuchen, weil auch die Anfragen an die meisten Suchmaschinen mit einem CGI-Skript abgewickelt wird.[9]

1.6. Zusammenfassung

All diese technischen Details sind beim ersten Lesen wahrscheinlich verwirrend. Zusammenfassend kann man sagen, daß die Anbieter von Suchmaschinen diese so gut wie überhaupt nicht bewußt „steuern", und bestimmte Sites willkürlich bevorzugen oder benachteiligen. Es sind eher die „einprogrammierten" Paradigmen der Robots und der Index-Software, die dazu führen, daß eine Site in der Suchhierarchie an einer bestimmten Stelle auftaucht. Problematisch erscheint an diesen Paradigmen aus der Sicht der User und Content Provider

1. daß kurze Texte gegenüber langen bevorzugt werden,

2. daß die <titel>- und die Meta-Tags, die vielen Usern, die eine HTML-Seite gestalten, wahrscheinlich überhaupt nicht bekannt sind, eine so entscheidende Rolle bei der Platzierung in einigen Suchmaschinen spielen (die gängigen HTML-Editoren generieren in
der Regel die Meta-Tags, ohne daß der User darauf Einfluß hat, wenn er sich nicht mit dem Source-Code herumplagen will), und

3. daß einige Suchmaschinen (wie z.B. **Excite)** die angebliche „Popularität" einer Site dazu benutzten, ihre Relevanz zu bestimmen.

Allerdings muß man sich immer wieder ins Gedächtnis rufen, daß keiner der oben angeführten problematischen *Programm-Features* bei allen Suchmaschinen anzutreffen ist, sondern sich alle Search Engines in entscheidenden Details unterscheiden.

Inzwischen haben einige Suchmaschinen damit begonnen, „Spamming"-Technik durch Modifikation ihrer Robot- und Index-Programme zu bekämpfen. Dies erscheint mir aber nicht als unberechtigte Einflußnahme, sondern eher als notwendige „Verteidigungsmaßnahmen" gegen die zum Teil tatsächlich reichlich dreisten Versuche von Content-Providern, ihre Sites zu promoten.

Daß einige Suchmaschinen inzwischen allerdings die Homepages von Servern wie „Ourworld" von CompuServe, „Tripod" und „Geocities", auf denen man sich umsonst eine Homepage anlegen kann, ignorieren (wie ich es im Kapitel über **Alta Vista** ausführlich beschreiben werde), ist allerdings eine nicht gerechtfertigte Einflußnahme auf die Auswahl von URLs, die diese Suchmaschinen liefern - und zwar eine, die die Suchmaschinen selbst um einige der ungewöhnlichsten und „Internet-typischsten" Sites in ihrer Datenbank bringt.

Diese Entwicklung kann dazu führen, daß bestimmte Inhalte von Suchmaschinen nicht mehr oder wenigstens nicht mehr so leicht gefunden werden können. Sehr kritisch ist in diesem Zusammenhang zu sehen, daß alle großen Suchmaschinen inzwischen

[9] **Dreilinger, Daniel:** a.a.O., S. 242f

„strategische Partnerschaften" mit anderen Sites geschlossen haben. Hier könnten wirtschaftliche Interessen zu weiteren „Selbstbeschneidungen" in den Datenbanken der Suchmaschinen führen. Über diese Entwicklung mehr im folgenden Kapitel.

1.7. Strategische Allianzen

Die Qualität einer Suchmaschine allein entscheidet noch nicht über ihren Erfolg. Wichtig ist es für die Anbieter von Suchmaschinen auch, ihre Site an den wichtigsten Ort im Netz zu „platzieren". Dazu haben alle Anbieter von Search Engines mit Softwarefirmen und anderen Websites „strategische Allianzen" vereinbart, bei denen zum Teil auch erhebliche Summen investiert werden, um den Traffic auf der eigenen Site zu erhöhen und diese dadurch wiederum für Werbekunden attraktiver zu machen.

Besondere Bedeutung kommt dabei der Kooperation mit den beiden großen Browser-Herstellern zu: Netscape und Microsoft. Außer **Magellan** haben zum Beispiel alle Suchmaschinen-Anbieter in irgendeiner Form Geschäftsbeziehungen mit Netscape, dem Produzenten des immer noch beliebtesten WWW-Browsers. Wer bei der Netscape-Software auf den „Net Search"-Button klickt, wird mit der Net Search Seite (http://home.netscape.com/home/internet-search.html) auf der Site von Netscape verbunden, von der aus man eine Anfrage in eine Reihe von Suchmaschinen eingeben kann.

Die Platzierung kostet die Anbieter von Suchmaschinen eine Menge Geld: Die vier Suchmaschinen, die bei Netscape dort als „premier providers" aufgeführt werden **(Lycos, Yahoo!, Infoseek, Excite)** haben dafür bezahlt, daß ein Link zu ihrem Angebot an dieser Stelle steht. (Mit einem "Customize"-Button kann der User eine fünfte Suchmaschine zu den vier „premiere providers" hinzufügen.) Auch die anderen Suchmaschinen, die als „marquee provider" aufgeführt werden, mußten für ihre Auflistung an dieser prominenten Stelle bezahlen.

Wie begehrt diese Platzierungen sind, kann man unter anderem daran sehen, daß die auf vier Suchmaschinen beschränkte Position als „premier provider" bis April 1998 ausgebucht ist. Schon eine Erwähnung als „Distinguished Provider" (ein unscheinbarer Text-Only-Link) kostet 30.000 Dollar pro Monat (http://home.netscape.com/ads/index.html#search). Was Netscape für die Platzierungen als „premiere provider" verlangt, ist ein Geschäftsgeheimnis der Unternehmens; es wird geschätzt, daß diese Position Lycos und Co 1996 fünf Millionen Dollar wert war. Was Netscape 1997 für dieselbe Platzierung verlangt hat, ist Gegendstand von zahlreichen Spekulationen; es wird aber vermutet, daß sich dieser Preis inzwischen auch nach dem Traffic richtet, den eine Suchmaschine über die Netscape-Homepage anzieht.

Um keinen der vier "premier providers" zu benachteiligen, wird der User nach dem Zufallsprinzip jedes Mal zu einer anderen Suchseite befördert, wenn er auf den „Net Search"-Button klickt: Auf jeder dieser Seiten sind jeweils groß die Namen aller vier „premier providers" angegeben, außerdem jedoch eine Suchmaske von jeweils einer der vier verschiedenen Suchmaschinen. Wer in dieses Formular einen Suchbegriff eingibt, wird zu der Site des jeweiligen Unternehmens weiterbefördert.

In der Fachpresse ist behauptet worden, daß dieses „Rotationsprinzip" nicht mehr vollkommen beliebig ist. Bei einem Test, bei dem 30 Mal hintereinander der „Net Search"-Knopf angeglickt wurde, fand „Search Engine Watch" (http://www.searchenginewatch.com/) folgende Verteilung:

Excite	11	35%
Infoseek	9	29%
Lycos	8	26%
Yahoo	3	10%

Auch die Platzierung auf der „Find it Fast Page" von Microsoft (http://home.microsoft.com/access/allinone.asp) ist kostenpflichtig. Hier sind die Suchmaschinen, die an der Top-Position rotieren, **Excite, AOL Netfind, Infoseek, Lycos** und **Yahoo!**

Die strategischen Allianzen mit Microsoft und Netscape sind wahrscheinlich die wichtigsten der Suchmaschinen, die einzigen sind es aber nicht: so haben **Infoseek, Vista, LookSmart** und **Lycos** mit dem Online-Dienst **CompuServe und Excite mit America Online und PointCast** ähnliche Verträge, die den Suchmaschinen eine hervorgehobenen Position auf den jeweiligen Homepages dieser Unternehmen sichern.

Die Platzierung auf den wichtigen Sites von Software-Herstellern und Online-Diensten werden einstweilen durch die „Gesetze der freien Marktwirtschaft" geregelt. Die Suchmaschinen, die es sich nicht leisten können oder wollen, für diese Platzierung Millionensummen zu bezahlen (wie z.B. **Search.com** oder **Northern Lights)** werden durch die Abwesenheit auf diesen entscheidenden Sites benachteiligt und dürften sich in geringeren „Click-Rates" ausdrücken.

Man kann allerdings darüber spekulieren, inwiefern die „strategischen Allianzen" zwischen Suchmaschinen und den Anbietern wichtiger Sites in Zukunft Einfluß auf die Selektion von „wichtigen" und „unwichtigen" Sites nehmen könnte. Microsoft hat zum Beispiel angekündigt, daß bei seinem Suchmaschinen-Projekt „Yukon" die Auswahl der Sites in der Datenbank stärker „gefiltert" werden soll. Ob das bedeutet, daß lediglich Redundanzen und Dubletten in der Datenbank vermieden werden sollen, oder ob Angebote von Microsoft oder von anderen Unternehmen bevorzugt werden, bleibt anzuwarten.

Open Text, die einzige Suchmaschine, die das Experiment wagte, Anzeigenkunden bessere Plätze in der Suchhierarchie zuzuweisen, hat sich mit dieser Geschäftstaktik selbst vom Markt der Suchmaschinen katapultiert: die User ignorierten die **Open Text,** als sich herausstellte, daß die Suchergebnisse manipuliert waren.

2. Das Beispiel: Alta Vista

Alta Vista ist die einzige Suchmaschine, über die es bereits ein eigenes Buch gibt: „The Alta**Vista** Search Revolution" von Richard Seltzer, Eric J. Ray und Deborah

S. Ray, in dem auch die Entstehung des Unternehmens ausführlich beschrieben wird.[10] Das Kapitel über die Unternehmensgeschichte ist sogar in deutsch auf der Homepage von **Alta Vista** zu finden (http://**altaVista**.telia.com/av/tmpl/de/news/story.html), weswegen sie hier nur kurz zusammengefasst werden soll:

Laut der offiziellen Version der Unternehmens-Saga, die auf der Alta**Vista**-Homepage nachzulesen ist, war die Entwicklung der Suchmaschine Produkt eines zufälligen Kantinengesprächs einiger Angestellten von Digital, die sich darüber Gedanken machten, ob eine Volltext-Suche des Internets möglich sei.

Nach dieser Version ist **Vista** entstanden, um der Netzöffentlichkeit die Leitungsfähigkeit ihres Chips „Alpha 8400" zu demonstrieren. Will man Louis Monier, dem wichtigsten Entwickler der Suchmaschine glauben, dann war die Entstehung von **Alta Vista** allerdings eher eine Art Zufallstreffer. In einer Email an den Autor schrieb er auf die Frage, ob er sich vorstellen könne, daß Digital das **AltaVista**-Projekt aus Kostengründen einstellen könnte:

> „AV as a showcase for Digital hardware is more an "after-the-fact rationalization" than an intent. It happened, and then management looked for a reason to justify its existence. It is a constant source of very positive PR, a showcase for hardware and other technical stuff, a source of "Internet good will", a very visible presence for Digital, and finally a source of revenues. The chances of Digital closing it down are very remote. (;-))"

Die Idee zur Entwicklung einer Suchmaschine entstand im April 1995. Am 4. Juli dieses Jahres (dem amerikanischen Nationalfeiertag!) schickte Monier den Suchrobot „Scooter", den er selbst programmiert hatte, zum ersten Mal los. Zu dieser Zeit arbeitete er zusammen mit einer Arbeitsgruppe im Forschungszentrum von Digital an einer Indexierungssoftware und einem Frontend für die Benutzer. Die Indexierungssoftware „Indexer" war eine Weiterentwicklung eines Suchprogramms für Email und Newsgroup-Postings, das von Digital schon früher entwickelt worden war, das *Frontend* wurde eigens für die Suchmaschine geschrieben.

Am 15. Dezember ging **Alta Vista** nach mehreren Testläufen offiziell ans Netz, und entwickelte sich schnell zu einer der beliebtesten Suchmaschinen, weil seine Datenbank bei vielen Netzsurfern als die vollständigste bei allen Suchmaschinen gilt, obwohl die Größe von **Vista** bei Suchmaschinen-Spezialisten umstritten ist

Vista ist ein echter Volltext-Index. Im Gegensatz zu einigen anderen der großen Suchmaschinen (wie zum Beispiel **Lycos),** speichert **Vista** wirklich jedes Wort einer gefundenen Seite ab. Dazu gehören neben dem eigentlichen Text eines Dokuments auch der Quellcode, also zum Beispiel die „Tags" und „Metatags". Die große Textmenge, die **Vista** sammelt, ist sowohl die Stärke wie die Schwäche dieser Suchmaschine: Sie liefert auch zu den abwegigsten Suchwörtern noch jede Menge „Treffer".

[10] **Seltzer, Richard; Eric J. Ray; Deborah S Ray:** The AltaVista Search Revolution, Berkeley 1997 (Osborne McGraw-Hill)

Das ist oft nützlich, gelegentlich aber auch lästig, wenn man zu einem beliebigen Thema Informationen sucht und plötzlich Millionen von URLs geliefert bekommt. Zwar bietet **Vista** auch die Möglichkeit einer Suche, die stärker „refined" ist und (zum Beispiel durch Bool'sche Befehle) präzisere Differenzierung möglich macht. Aber die Suchergebnisse, die man bekommt, wenn man mit einem einfachen Stichwort sucht, dürfte wohl auch den passioniertesten Netzsurfer überfordern. „Mehrwert" in Form von Nachrichten, Chats oder anderen Content liefert die **Vista** nicht. Allerdings gibt es neuerdings die Beta-Version einer Übersetzungssoftware namens „Babelfish" auf der Site der Suchmaschine.

Vista hat einen Index mit über 100 Millionen WWW-Seiten, die von einer Million Servern zusammen gesammelt wurden; außerdem kann man auch in vier Millionen Usenet-Postings aus 13 000 Newsgroups suchen. Doch trotz dieser Größe, gab es Anfang 1997 in den USA eine Debatte über die tatsächliche Größe von **Vista,** die durch die gesamte Computer-Branchenpresse ging.

Der Streit begann mit einem harmlosen Artikel (http://www5.zdnet.com/anchordesk/story/story__768.html) in ZDNet (http://www5.zdnet.com/), der Tips gab, wie man seine eigene Homepage populärer machen könne. John Pike, der Webmaster der Federation of American Scientists Web Site, reagierte auf diesen Artikel mit einem Leserbrief (http://www5.zdnet.com/anchordesk/talkback/talkback__l 1638.html), als er feststellte, daß 600 der über 6.000 Seiten von seiner Website bei **Alta Vista** verzeichnet sind. In einer Reaktion auf seine Kritik antwortete Louis Monier von **Vista** umgehend (http://www5.zdnet.com/anchordesk/talkback/talkback_13066.html):

> „The claim that Alta**Vista** indexes 'the whole Web' was never made by myself or (I hope) any Alta**Vista** employee. The truth is that no search engine indexes the whole Web. The concept of 'the size of the Web' in itself is flawed, as there are many sites virtually infinite in size: dynamically generated documents, personalized news pages and shopping baskets using cookies, robot traps, scripts, the list goes on. Also unless one spends a lot of effort cleaning it up (we do), an index holds a lot of pages unlikely to ever be retrieved, like multiple copies of the same page and access logs. Size alone is a poor measure of usefulness."

Bei Alta Vista, die lange damit geworben haben, die Suchmaschine mit den größten Index zu sein, wird also ebenfalls eingeschränkt, welche Seiten aufgenommen werden, und welche nicht. Wie diese Praxis genau funktioniert, wurde im April 1997 auf der I-Advertising Mailing List, einer Liste über Werbung im Internet, genauer bekannt: Den Anstoß dazu gab Chris Longley, der seine Homepage auf dem Tripod-Server bei **Vista** anmelden wollte. *Tripod* bietet - wie *Geocities* und der Online-Dienst *CompuServe* - gratis Serverplatz an, auf dem man sich eine eigene Homepage einrichten kann. Als Longley seine Seiten anmelden wollte, bekam er die Meldung: "Too many URLs submitted from that site."

Eine Nachfrage bei **Alta Vista** ergab, daß die Suchmaschine nicht länger Seiten von Servern wie *Tripod* oder *Geocities* aufnimmt, weil auf ihnen viele Spammer sogenannte, Jumppages" einrichten. "Jumppages" sind Webseiten, die so geschrieben

wurden, daß sie bei Suchmaschinen besonders gute Rankings bekommen. Spammer legen oft viele verschiedene Seiten auf solchen öffentlichen Servern ab, und melden alle bei den Suchmaschinen in der Hoffnung an, daß eine besonders hoch in deren Hierarchie auftaucht. Von diesem „Jumppages" kommt man dann auf die eigentliche Werbeseite.

Ich gehe deshalb so ausführlich auf diese Beispiele ein, um zu zeigen, daß selbst **Vista,** die bei vielen als vollständigste Suchmaschine der Welt gilt, keineswegs ein auch nur annähernd vollständiges Verzeichnis von Netzseiten bietet. Das hat für Louis Monier einen ganz einfachen Grund: „Nobody can afford enough hardware to index the entire Web AND serve it back to the entire planet." (http ://www5.zdnet.com/anchordesk/talkback/talkback_ 13066.html)

Während diese materielle Einschränkung für alle Suchmaschinen gilt, gibt es bei **Vista** einige „Spezialitäten", die zu der besonderen Zusammenstellung der Hits beitragen: Das beginnt mit dem Robot, dem sein Schöpfer Monier den Namen „Scooter" gegeben hat.

2.1. Der Robot

Dieser Robot wurde so programmiert, daß er bei seinen Suchen nach dem "Breadth First"-Prinzip verfährt, d.h. daß er neu gefundene Links an das Ende einer Liste der zu untersuchenden Links stellt. Die praktische Konsequenz davon ist, daß „Scooter" eine neu gefundene Site nicht sofort vollständig durchsucht, sondern nur langsam und Schritt für Schritt erfaßt.

„Scooter" geht so vor, um den Server, den er untersucht, nicht zu überlasten. Zu der Geschwindigkeit, mit der sich der Robot durch eine Homepage bewegt, schreibt Monier in einer Email an den Autor:

> „The only rule the spider follows is to space out the visits: never fetch more than one page at the same time, see how long it takes to get a page and don't contact the server again for some multiplier of that duration. Simple rule but it works very well."

Wenn „Scooter" eine URL einmal entdeckt hat, kehrt er regelmäßig zu ihr zurück. Wie die Logfiles des Servers der Projektgruppe Kulturraum Internet (http://duplox.wz-berlin.de) zeigen, taucht „scooter.pa-x.dec.com" (wie die IP-Adresse von „Scooter" lautet) relativ regelmäßig einmal im Monat auf, und „fetcht" dann in zwei Wochen im Durchschnitt etwa 1200 der 1400 Seiten auf diesem Server. Das Tempo der Zugriffe variiert: manchmal greift er im Minutenrhythmus zu, manchmal läßt er sich bis zu 24 Stunden Zeit zwischen zwei Hits.

2.2. Der Index

Im Gegensatz zu anderen Suchmaschinen (wie z.B. **Excite)** speichert **Alta Vista** die Seiten, die in den Index aufgenommen werden, wirklich vollständig. Außer dem vollständigen <body>-Text gehören dazu auch

Titel, Header, Inhalt der Meta-Tags, die Größe in Bytes, das Datum an dem die Site zuletzt gespidert wurde, alle Links, der Anchor-Text der Links und andere HTML-Tags wie Applets, der Host, Image-Files etc. Eine eigene Summary generiert **Alia Vista** nicht.

Dadurch erlaubt es dem geduldigen Surfer wirklich gründlichste Suchen. Ich habe mit **Alta Vista** bei der Suche mit meinem Namen als Stichwort schon Seiten gefunden, auf denen mein Name (oder Teile davon) nur einmal oder ganz am Ende einer Seite stehen. In punkto Vollständigkeit der gefundenen Seiten steht **Alta Vista** also keiner anderen Suchmaschine nach.

2.3. Die Suchmaschinen-Software

Wie werden die verschiedenen Sites, die bei **AltaVista** gespeichert werden, bewertet, wenn ein User ein oder mehrere Suchworte eingibt? Wie bereits erwähnt, wird die Position des Suchworts im Dokument bewertet, bei mehreren Worten auch, wie nahe diese in den gefundenen Dokument beieinander stehen. Im Unterschied zu anderen Suchmaschinen kann der User bei **Vista** beim „Refined Search" auch die Relevanz eines Wortes bei der Suche selbst bestimmen. So kann man zum Beispiel durch ein Pluszeichen vor einem Suchwort der Suchmaschine bedeuten, daß dieses Wort besonders wichtig ist. Durch solche Befehle kann der User also selbst die Ergebnisse der Suche steuern.

Außerdem sind bei **Alta Vista** inzwischen nicht nur Suchen nach Stichwörtern, sondern auch nach verschiedenen Dokumententypen (bei WWW-Dokumenten u.a. Bilder (.gif, .jpg), Java, Host, Top Level Domain; bei UseNet-Postings: From, Subject, Newsgroup, Summary, Keyword) möglich. Weil auch der verborgene „Anchor-Text" von **Vista** gespeichert wird, ist es auch möglich, herauszufinden, von wo es Links auf eine bestimmte Site gibt.

„Trunkierungen" (z.B. „Macinto*" als Suchwort) gibt es bei **Vista** im Gegensatz zu anderen Suchmaschinen genauso wenig wie „Stemming" (die Suchmaschine findet Worte, die den selben Wortstamm wie das Suchwort haben, z.B. „working" mit dem Suchwort „work"). Umlauten oder Buchstaben mit Apostroph, die nicht im ASCH-Alphabet vorkommen, werden durch Buchstaben aus dem „Character Sets Latin-1" ersetzt (z.B. ä durch a etc.).

Laut Louis Monier gibt es bei **Alta Vista** auch keine Stopworte (also Worte wie „the", „is" oder „and" etc., die so häufig sind, daß die Suchmaschine sie nicht mehr berücksichtigt):

> **„Vista** does not really use stop words. If a word is very common in the index, and you use ranked (simple) search, the "weight" of this word falls to zero. In other terms, containing the word "the" is not a very good differentiator for a document. But these words are indexed and can be used either by forcing simple search (with +the), or by using advanced search (the AND King), or phrases "to be or not to be". Which word is ignored during ranking simply depends on its frequency and varies from one day to the next"

Wer ein besonders häufiges Wort - wie ein Hilfsverb oder eine Präposition - als Suchwort bei **Vista** eingibt, bekommt bei einer einfachen Suche also die Meldung: „No documents match the query", kann aber mit der „refined"-Suche trotzdem Dokumente mit diesen Worten suchen.

Die vielen Möglichkeiten, mit denen man bei **Vista** seine Suche verfeinern kann, machen die Suchmaschine jedoch zu einem nach wie vor sehr brauchbaren Rechercheinstrument. Wer seine Suche richtig einzuschränken versteht, kann die tausende von Hits, die man mit einem einzigen, normalen Suchwort bekommt, schnell auf eine überschaubarere Zahl eingrenzen. Das gesamte Internet wird von **Vista** freilich nicht erfaßt, aber das wäre wohl auch eine unmögliche Aufgabe.

Vista fehlen dagegen einige Eigenschaften, die andere Suchmaschinen wenn schon nicht besser, so doch übersichtlicher machen. Während **Vista** als Stichwörtern immer die ersten zwanzig Wörter eines Dokuments angibt (die bei vielen Webpages leider Navigationsbegriffe wie,,Home", Links" oder „Back" sind), fassen andere Suchmaschinen die gefundenen Hits oft sinnvoller zusammen: bei **Excite** werden zum Beispiel die Inhalte der Meta-Tags angezeigt, wenn es welche gibt; bei **Lycos** werden aus jedem Dokument die wichtigsten Begriffe extrahiert, ein Verfahren, das freilich etwas zweifelhaft erscheint, wenn man weiß, daß Lycos nur die ersten 200 Worte von jedem gefundenen Dokument speichert.

Aber wie inzwischen klar sein dürfte, gibt es die „beste" Suchmaschine überhaupt nicht, sondern nur verschiedene Methoden, das Datenwirrwarr des Internets durchsuchbar zu machen. Letztlich kommt es weniger auf die technische Ausstattung einer Suchmaschine an, auch nicht auf möglichst große Datenbestände, sondern auf die Fähigkeit des Users, souverän mit den zur Verfügung stehenden Suchoptionen umzugehen.

3. Zusammenfassung

Ausgangspunkt dieser Arbeit war die Frage, inwiefern Suchmaschinen zu einer Zentralisierung des Internets beitragen und ob trotz ihrer scheinbar „blinden" Technologie die Möglichkeit des Mißbrauchs oder der Instrumentalisierung zum Beispiel im Interesse von Wirtschaftsunternehmen besteht. So haben sich diese Annahmen nicht bestätigen lassen.

Problematisch erscheint aber vor allem, daß manche Suchmaschinen (wie **Lycos)** auch die Zahl von Links auf eine bestimmte Site als Maßstab für ihre Gewichtung nehmen. Indem sie die Bedeutung einer Seite nach ihrer angeblichen Popularität messen (welche sich durch die Zahl der Links auf diese Seite ausdrückt), tragen sie zu einer Zentralisierung des Internets auf angeblich besonders beliebte Sites bei. Auch der Ausschluß bestimmter Server (wie *Geocities* oder *Tripod)* trägt zur Verringerung der Suchergebnisse ebenso bei wie zur Verarmung der Datenbanken der Search Engines bei, weil so auch besonders ungewöhnliche, „Internet-typische" Sites ausgeschlossen werden.

Weitere problematische Details bei der Programmierung einiger Suchmaschinen habe ich in der Zusammenfassung des Kapitelabschnitts 1.5. beschrieben.

Eine weitere Frage, die dieses Papier beantworten sollte, lautete: „Können sich die Geschäftspartner von Suchmaschinen-Anbietern gute Positionen im Such-Index erkaufen?" Die Tatsache, daß die Suchmaschine **Excite** eigene „Channels" bei den Suchergebnissen bevorzugt, macht solche Überlegungen noch plausibler: Wer bei **Excite** Suchworte eingibt, die Rubriken-Titel im Excite-Index (z.B. „Sports", „Shopping" oder „Lifestyle") sind, bekommt als erstes Links zu diesen Rubriken serviert, bevor er die URLs von anderen Sites angezeigt bekommt.

Die oben formulierte Vermutung ließ sich allerdings ebenfalls nicht uneingeschränkt bestätigen: Zwar beurteilen einige Suchmaschinen, Sites, die sie auch in ihrem Index haben, als relevanter als andere, aber eine Bevorzugung von Geschäftspartnern konnte nicht festgestellt werden. Zumindest ganz primitive Methoden (zum Beispiel der Einkauf von einem Unternehmen in die Hierarchie einer Suchmaschine) scheint es nicht zu geben. Der einzige Anbieter einer Suchmaschine, der seine Werbekunden durch bevorzugte Platzierung in dem Ranking seines Dienstes „belohnen" wollte, war **Open Text,** die dafür mit dem vollständigen Verlust ihrer Akzeptanz bei den Usern bezahlen mußten. Das Unternehmen ist inzwischen aus dem Suchmaschinen-Markt ausgestiegen.

Während diese Art von Manipulation offensichtlich nicht üblich ist, lassen die Betreiber von Suchmaschinen ihren Robots und Index-Systemen auch nicht vollkommen „freien Lauf", wie die Entfernung von Spam-Seiten bei **InfoSeek** und **Vista** und der Ausschluß von bestimmten Servern bei **Vista** zeigt. Außerdem schränken die Suchmaschinen durch ihre Vorgehensweise auch die Möglichkeiten bei der Gestaltung ein. Zu diesen Einschränkungen gehört zum Beispiel, daß sie vorwiegend nach Text suchen, Bilder hingegen schlecht indexieren können. Außerdem gibt die Programmierung von Suchmaschinen wie Lycos Seiten mit kurzen Texten den Vorzug vor längeren Texten.

Auch ohne plumpe Tricks verstärken die Suchmaschinen auf jeden Fall eine Tendenz zur Zentralisierung, die freilich schon der Hypertext-Logik zu eigen ist. Hypertext ist nämlich keineswegs das Gewebe ohne Zentrum, als das er gerne dargestellt wird. Wie Rainer Rilling in seinem Aufsatz, Auf dem Weg in die Cyberdemokratie" (http://staff-www.uni-marburg.de/-riUingr/bdweb/texte/cyberdemokratie-text.html) gezeigt hat, können auch Links zu einer Lenkung und Zentralisierung von Aufmerksamkeit und Sichtbarkeit führen:

> „Links strukturieren die Verteilung von Sichtbarkeit, Aufmerksamkeit und schließlich Anerkennung im Informationsraum. Das Web generiert den eigenartigen, systemspezifischen Zwang, Kenntnis vorhandener Präsenzen durch Links auszuweisen, somit das Bemühen, in einem Raum eigene Zentralität zu demonstrieren, dessen einfachste Grundstruktur eben durch das Verhältnis von Zentrum und Peripherie gebildet wird. Die grassierenden Hotlists, die es in anderen Medien so eben nicht gibt... stehen für diesen Imperativ. Nur wer

> Verweiskompetenz demonstriert, verhält sich programmgerecht, systemspezifisch, informationsraumgerecht...
>
> Netzreputation - oder soziales Netzkapital - entsteht durch kompetente Verweise auf andere/s und Verweise anderer auf sich selbst. Reputation und Zentralität durch Hypertextverweise hängen auf durchaus vertraute wechselseitige Weise miteinander zusammen: Reputation schafft Zentralität, Zentralität generiert Reputation...
>
> Der Hyptertextmechanismus ist nichts anderes als ein äußerst zwingender Imperativ, Peripherie, Marginalität oder, politisch formuliert, potentiellen Dissens zugunsten von Zentralität oder, politisch formuliert, Mainstream zu verlassen,"

In diesem Prozeß spielen die Suchmaschinen eine wichtige Rolle. Nicht nur, daß sie selbst zentrale „Eingangstore" zum Internet sind, eine Position, die sie durch strategische Allianzen mit anderen stark frequentierten Websites (wie zum Beispiel denen von Netscape, Microsoft oder AOL) und „Web-Napping" noch verstärken. Durch die Klassifikation ihrer Suchergebnisse nach angeblicher Popularität verstärken sie den „Drang zum Zentrum", der dem Hypertext schon durch seine eigene Logik inhärent ist, noch.

Dennoch: trotz dieser Einschränkungen hat der User der Search Engines einen relativ breiten Handlungsspielraum - sowohl bei der Informationssuche wie auch beim Plazieren seiner Seiten unter den Suchergebnissen der Search Engines. Wer die Suchmaschinen als „Hierarchisierer" des Internets verurteilt, macht es sich zu einfach und übersieht sowohl die Sachzwänge, denen die Suchmaschinen ausgesetzt sind, als auch die Tatsache, das sich auch ohne die Suchmaschinen im Internet eine gewisse Hierarchie zwischen „zentralen" und „peripheren" Sites herausgebildet hätte. Wer eine der „großen" Suchmaschinen wie **Alta Vista** richtig benutzten gelernt hat, kann mit ihr auch abwegiger und weniger zentrale Sites entdecken.

Auch die Anbieter von Inhalten können dazu beitragen, daß ihre Sites leichter gefunden werden können - eine Möglichkeit, die viele content provider nur nachlässig oder gar nicht nutzen. Wer etwa eine Seite mit einem ungewöhnlichen Thema richtig programmiert (zum Beispiel durch den gezielten Einsatz von Metatags), kann sie nach wie vor an zentraler Stelle der Suchhierarchien positionieren. Und wer es versäumt, seine Site bei den führenden Suchmaschinen anzumelden, muß sich nicht wundern, wenn es sehr lange dauert, bis diese von selbst auf die eigene Site aufmerksam werden.

4. Anhang

4.1. Technischen Spezifizierungen von Alta Vista

(Zusammengestellt aus Angaben der Unternehmen, eigene Recherchen und den Angaben des EU-Forschungsprojekts „Development of a European Service for Information on Research and Education" (DESIRE), Stand: Oktober 1995) (http://www.ub.lu.se/desire/)

URL: http://altavista.digital.com
Umfang: 100 Millionen WWW-Seiten auf einer Million Servern, 4 Millionen Usenet-Postings aus 13 000 Newsgroups
Anbieter: Digital Equipment Corporation
Unternehmens-URL: http://www.digital.com
Mirrorsites: in den USA keine
Häufigkeit von Updates: bis zu zehn Millionen Hits pro Tag
Ratings, review, „added value": keine
Kosten für den User: keine
Hits: 4,7 Millionen pro Tag
Strategische Allianzen: Netscape, Microsoft, Yahoo!

Regionalisierungen
Alta Vista Telia (Alta Vista Telia)
(http://www.altavista.telia.com)
Existiert seit dem 1. Dezember 1996, mit Interface in verschiedenen europäischen Landessprachen, Sitz in Schweden
Alta Vista Latin America/Alta Vista Southern Europe (Alta Vista Magallanes)
http ://www. altavista.magallanes.net
Interface in Spanisch, Portugiesisch und Englisch, Sitz in Madrid
AltaVista Australasia (Alta Vista Australasia)
http://www.altavista.yellowpages.com.au
für Australien und Neuseeland, Sitz in Australien
Alta Vista Asia
http://altavista.skali.com.my
Mirror des amerikanischen „Originals", keine asiatischen Sprachen /Schriften möglich, Sitz in Malaysia

Unternehmensgeschichte
Alta Vista wurde im Sommer 1995 im Research Laboratory der kalifornischen Hardwarefirma Digital unter der Leitung von Louis Monier entwickelt. Der Web-Crawler „Scooter", den Monier programmiert hatte, wurde am 4. Juli 1994 in Testbetrieb genommen, im Herbst entstand aus den Dokumenten, die „Scooter" gefunden hatte, eine Datenbank mit 16 Millionen Einträgen. Am 15. Dezember 1995 „eröffnete" Alta Vista seine Site und hatte nach drei Wochen bereits zwei Millionen Hits pro Tag.

Robot
Typ: „Scooter", folgt „Robots Exclusion Standard"
Methode: automatische Robot Suche, Anbieter können ihre Site bei Alta **Vista** anmelden, aber nicht abmelden
Depth First/Breadth First: Breadth First
Einsatzgebiet: WorldWideWeb, UseNet
Wie oft kehrt der Robot zu einmal gefundenen URL's zurück? alle vier bis sechs Wochen

Index-System
Software „Indexer" von Digital
Was wird in den Index aufgenommen?
Volltext: Ja
Titel: Ja
Header: Ja
Meta-Tags: Ja
Größe in Bytes: Ja
Datum: Ja
Links: Ja
Anchor-Text von Links: Ja
Andere HTML-Tags: Ja, Applets, Host, Image
Summary: Nein

Datenbank-System
Software: „Indexer" von Digital
Boolean: Ja
Nesting (Klammern): Ja
Bewertungsmethode: Bewertet wird die Position des Suchworts im Dokument, bei mehreren Worten, wie nahe diese beieinander stehen. Außerdem kann der User die Relevanz eines Wortes bei der Suche selbst bestimmen.
Trunkierungen: Nein
Stemming: ?
Character Sets: Latin-1, ersetzt Buchstaben, die im englischen Alphabet nicht vorkommen, durch ihre englischen Äquivalente (z.B. ä durch a etc.)

Was kann man suchen?
Suche nach Dokumenten-Typ: Bilder (.gif, .jpg), Java, Host, Top Level Domain; bei UseNet-Postings: From, Subject, Newsgroup, Summary, Keyword,
URL: Ja
Titel: Ja
Volltext: Ja
Links von anderen Sites: Nein
Zusammenfassung: Nein
Anchor-Text, Links: Ja
Concept Search: Nein

4.2. Suchmaschinen Glossar

Boolean search
Suche, bei der es dem User möglich ist, Dokumente, die bestimmte Stichworte haben, ein- oder auszuschließen. Dafür werden „Operatoren" wie AND, NOT und OR verwendet.

Concept search
Suche nach Dokumenten, die mit einem Suchwort in Zusammenhang stehen, ohne dieses notwendigerweise zu enthalten (z.B. bei HotBot).

Full-text index
Index, der jedes Wort eines Dokuments (einschließlich der „Tag"-Befehle und „Stop-Words") indexiert.

Fuzzy search
Auf „Fuzzy Logic" beruhende Suchmethode, die auch dann Resultate liefert, wenn Worte nur unvollständig oder falsch geschrieben eingegeben werden.

Index
Durchsuchbarer Katalog von Dokumenten, die der Robot zusammengetragen hat.

Key word search
Suche nach einem oder mehreren Stichworten, die von einem User vorgegeben wurden.

Phrase search
Suche nach dem gleichen Satz oder der gleichen Wortkombination, die der User eingegeben hat.

Proximity search
Suche, bei der die vom User gesuchten Wörter möglichst nah beieinander stehen sollten.

Query-By-Example
Suche, bei der der User die Suchmaschine mehrere, einem bestimmten Dokument ähnliche Dokumente finden soll. Auch "find similar" genannt.

Suchmaschine
Software, die einen Index durchsucht und „Treffer" zusammenstellt. „Suchmaschine" wird oft als gleichbedeutend mit Index und Spider angesehen, obwohl dies zwei andere Komponenten sind, die mit der Suchmaschine zusammen funktionieren.

Spider
Auch: Crawler oder Robot. Die Software, die ein Dokument durchsucht, in einen Index einfügt und - Links folgend - weitere Dokumente findet

Stemming
Suche nach dem Wortstamm, d.h. daß die Software auch nach dem Verb „work" sucht, wenn man "working" eingibt

Stop words
Konjunktionen, Präpositionen, Artikel und andere besonders häufig vorkommendes Worte, wie z.B „und", „oder", „mit", werden häufig von Suchmaschinen ausgeschlossen. Die Suchmaschine „stoppt" dann nicht bei diesen Wörtern, sondern überspringt sie.

Thesaurus
Liste von Synonymen, die eine Suchmaschine benutzt, um Treffer für ein bestimmtes Wort zu finden, das im Dokument selbst nicht auftaucht.

Trunkierung
Möglichkeit, nur mit den ersten Buchstaben eines Wortes (z.B.Macinto*) zu suchen.

Webnapping
Viele Suchmaschinen bieten Surfern ein Stück HTML-Code für ihre Homepage an, der auf dieser ein kleines Eingabefenster erscheinen läßt, mit dem man in ihrer Search Engine suchen kann.

4.3. Literatur

4.3.1. Gedruckte Publikationen

Helmers, Sabine; Ute Hoffmann; Jeanette Hofmann: Netzkultur und Netzwerkorganisation. Das Projekt „Interaktionsraum Internet", WZB Discussion Papers FS II 96-103, Wissenschaftszentrum, http://duplox.wz-berlin.de/docs/

Glossbrenner, Alfred and Emily: Search Engines for the WorldWideWeb, Berkeley, CA 1997 (Peachpit Press)

Seltzer, Richard; Eric J. Ray; Deborah S. **Ray:** The Alta**Vista** Search Revolution, Berkeley 1997 (Osborne McGraw-Hill)

Steinberg, Steve: Seek and Ye Shall Find (Maybe), Wired 5 (1996), S. 108- 114, 172-182 http://www.wired.com/wired/4.05/indexing/

Williams, Joseph: Bots and other Internet Beasties, Indianapolis 1996 (sams.net)

4.3.2. Online-Publikationen

Aigrain, Philippe: Attention, Media, Value and Economics http://www.firstmonday.dk/issues/issue2_9/aigrain/index.html

Brake, David: Lost in Cyberspace http://www.newscientist.com/keysites/networld/lost.html

Moody, Glyn: Searching the Web for Gigabucks http://www.newscientist.com/keysites/networld/webworm.html

Rilling, Rainer: Auf dem Weg zur Cyberdemokratie? http://staff-www.uni-marburg.de/-rillingr/bdweb/texte/cyberdemokratie-text.html

Winkler, Hartmut: Suchmaschinen - Metamedien im Internet?, Telepolis, http://www.heise.de/tp/deutsch/inhalt/te/135/1 .html

4.4. Hotlinks

Search Engine Bibliography
http://www2.hawaii.edu/~rpeterso/bibliog.htm
Englischsprachige Bibliographie zu Suchmaschinen

EU-Forschungsprojekt „Development of a European Service for Information on Research and Education" (DESIRE)
http://www.ub.lu.se/desire/
Minutiöse Studie über die verschiedenen Suchmaschinen mit sehr detaillierten Informationen über die technische Ausstattung der einzelnen Sites und ihrer Software

Search Engine Watch
http://www.searchenginewatch.com
Sehr gut gemachte private Homepage über Suchmaschinen, die leicht verständlich erklärt, wie Suchmaschinen funktionieren, wie man seine eigene Homepage platziert und wie die ökonomischen Hintergründe sind

Die kleine Suchfibel.
http://www.karzauninkat.com/suchfibel/
Deutschsprachige Site, die ebenfalls sehr gut verständlich erklärt, wie Suchmaschinen funktionieren.

Search Engine Tutorial - Web Design & Promotion
http://www.northernwebs.com/set/set02.html
Richtig anmelden bei Suchmaschinen - und was man dabei erleben kann.

Search Insider
http://www.searchinsider.com
Schön gestaltete Site, die allerdings eher durch opulente Grafiken als durch ihren Inhalt besticht

SearchHelp - Finding Stuff on the Web
http://www.searchhelp.com/
Site mit Suchhilfen - noch im Aufbau

The Art Of Business Web Site Promotion
http://deadlock.com/proinote/
Tips für Anmeldung von kommerziellen Sites bei Suchmaschinen

The Spider's Apprentice - Tips on Searching the Web
http://www.monash.com/spidap.html
Wie benutzt man Suchmaschinen am effizientesten?