



Weshalb Google AltaVista überflügelte.

Seminararbeit Christoph G Schmutz

November 2003

Prof. Dr. Andreas Meier, Lehrstuhl für Wirtschaftsinformatik

Betreuung durch Andreea Ionas, Diplomassistentin für Wirtschaftsinformatik

Universität Fribourg, Schweiz

Abstract

Der ehemalige Branchenleader unter den Suchmaschinen, AltaVista, wurde um die Jahrtausendwende von einem Forschungsprojekt zweier Studenten abgelöst, der Suchmaschine Google. Dank der konsequenten Ausrichtung dieses Projektes auf "Web-Tauglichkeit", Qualität der Suchresultate und Benutzerfreundlichkeit, sowie der innovativen technischen Umsetzung davon, liegt Google heute hoch in der Gunst der Internet-Nutzer. Dies im Gegensatz zur Konkurrenz AltaVista, welche deutlich abgeschlagen ist.

Keywords: Internet, Suchmaschine, AltaVista, Google, PageRank, Anchortext

1	Einleitung	03
1.1	Problemstellung	03
1.2	Zielsetzung	03
1.3	Vorgehensweise	03
2	Einführung ins Thema	04
2.1	Problematik des Internet	04
2.2	Bestehende Lösungen	04
2.2.1	Kataloge	04
2.2.2	Roboterbasierte Suchmaschinen	04
2.2.3	Metasuchmaschinen	05
3	Ist AltaVista tatsächlich von Google überflügelt worden?	05
3.1	Geschichte	05
3.1.1	AltaVista	05
3.1.2	Google	07
3.2	Anzahl indizierter Seiten	09
3.3	Popularität	10
3.4	Suchanfragen pro Tag	11
3.5	Zusammenfassung	12
4	Technik einer roboterbasierten Suchmaschine	12
4.1	Datenbeschaffung	12
4.1.1	Was finden Robots?	13
4.1.2	Was kann nicht gefunden werden?	13
4.1.3	Was darf nicht gefunden werden?	13
4.2	Indizierung	14
4.2.1	Parser	14
4.2.2	Store Server	15
4.2.3	Lexicon	15
4.2.4	Hit Lists	15
4.2.5	Repository	15
4.2.6	Aktualisierung der Daten	15
4.3	Anfragebearbeitung	16
4.3.1	Anfragemöglichkeiten	16
4.3.2	Darstellung der Suchresultate	17
5	Stärken von Google	17
5.1	Crawler	18
5.2	Repository	18

5.3	Anchor	18
5.4	Barrels	19
5.5	Links	19
5.6	Page Rank	19
5.6.1	Eigenschaften guter Ranking-Algorithmen	19
5.6.2	Definiton des Page Rank-Algorithmus	20
5.6.3	Bewertung des Page Rank-Algorithmus	22
5.7	Searcher	23
5.7.1	Suchvorgang	23
5.7.2	User-Feedback Strategie	23
5.7.3	Graphische Aspekte	23
6	Zusammenfassung	24
7	Literaturverzeichnis	25
8	Quellenangabe der Abbildungen	26
9	Anhang A	27
10	Anhang B	28

1 Einleitung

1.1 Problemstellung

Was vor gut 20 Jahren in den Vereinigten Staaten als militärisches Projekt begann, ist bis heute zur wahren Informationslawine angewachsen; das Internet. Suchmaschinen sind eine der verschiedenen Möglichkeiten, wie das Problem der beliebig vielen Informationen angegangen wird. AltaVista, die lange Zeit populärste Suchmaschine im WorldWideWeb, wurde Ende der 90iger Jahre von Google abgelöst, einem Forschungsprojekt von zwei amerikanischen Studenten. Diese Arbeit beschäftigt sich mit der Frage, welche technischen Aspekte zu diesem „Machtwechsel“ geführt haben.

1.2 Zielsetzung

Es soll einen Überblick vermittelt werden, über die Funktionsweise einer roboterbasierten Suchmaschine, die technischen Errungenschaften auf diesem Gebiet bis heute, sowie das Erfolgsgeheimnis von Google.

1.3 Vorgehensweise

In einem ersten Teil wird eine kurze Einführung ins Thema gegeben, wobei die untersuchte Thematik in einen grösseren Zusammenhang gestellt und gegen andere Gebiete abgegrenzt wird. Der Titel dieser Arbeit impliziert, dass die Suchmaschine AltaVista von Google überrundet worden ist. Dies soll in einem zweiten Schritt überprüft und belegt werden. Anschliessend wird die Funktionsweise von roboterbasierten Suchmaschinen im Allgemeinen erklärt, worauf dann die spezifischen Stärken von Google beleuchtet werden.

2 Einführung ins Thema

2.1 Die Problematik des Internet

Das WorldWideWeb (WWW) wird von Jahr zu Jahr umfangreicher. So hat sich beispielsweise die Anzahl der registrierten Domains mit Endung .ch hat in den letzten Jahren um Faktor 1772 vermehrt. Waren es im Januar 1995 noch bescheidene 307 Domains, so sind es heute (Stand 1. Oktober 2003) 544'228 [Switch 03]. Egal wie stark genau das Wachstum ist, und wie viele Websites es insgesamt tatsächlich gibt, unbestritten ist, das WWW stellt eine enorme Informationsflut zur Verfügung. Eine der Hauptschwierigkeiten im Umgang mit dem WorldWideWeb ist es deshalb, gezielt die gewünschten Informationen im virtuellen Datenschwungel zu finden.

2.2 Bestehende Lösungen

Um Ordnung ins Internet zu bringen gibt es heute verschiedene Lösungsansätze, wovon die wichtigsten hier kurz geschildert werden, damit eine klare Abgrenzung möglich wird.

2.2.1 Kataloge

Ein Ansatz ist es, von Menschenhand Ordnung ins Chaos zu bringen. Hierbei durchsuchen Online-Redaktoren des Anbieters das WorldWideWeb und nehmen qualitativ hochwertige Seiten in den Katalog (Verzeichnis, Directory) auf, welcher normalerweise in verschiedene (thematische) Rubriken unterteilt ist. Die Nutzer besitzen die Möglichkeit, eigene Seiten anzumelden, welche vom Anbieter zwecks Qualitätswahrung kontrolliert werden. Ein bekanntes Beispiel für Verzeichnisse dieser Art ist yahoo.com. Häufig besitzen solche Verzeichnisse auch ein mehrstufiges Rating, womit die einzelnen Seiten, je nach dem von Anbietern oder Nutzern, bewertet werden.

Andere Kataloge werden von den Benutzern selbst erstellt, beziehungsweise von Gruppen von Nutzern. Dabei betreuen einzelne Nutzer auf freiwilliger Basis die verschiedenen Kategorien und sind unter Umständen auch gleichzeitig für die Begutachtung (Qualitätswahrung) verantwortlich, als Beispiel sei hier OpenDirectory [dmoz.org 03] genannt. Hier werden auch Vorschläge für neue Rubriken und Unterrubriken entgegengenommen. [Bekavac 02]

Der Vorteil von Katalogen liegt in der relativ hohen Qualität der aufgeführten Seiten, da „von Mensch“ sortiert. Dafür hat man mit speziellen Suchanfragen meist wenig Erfolg, und muss sich mit – im Verhältnis zu den Suchmaschinen – wenig Treffern zufrieden geben.

2.2.2 Roboterbasierte Suchmaschinen

Die roboterbasierten Suchmaschinen arbeiten nach einem grundsätzlich anderen System. Hier wird das Netz von einem Programm durchsucht, welches von einigen Startseiten aus automatisch den Links folgt und die besuchten Seiten zurückschickt. Die entdeckten Seiten werden komprimiert und in einer Datenbank abgelegt (indiziert). Die Suchanfrage vom Benutzer wird also schlussendlich nur auf diese Datenbank ausgewertet, welche die aufgespürten Seiten enthält.

Die grösseren derartigen Suchmaschinen betreiben häufig neben der Stammseite auch länderspezifische Seiten (www.altavista.com, www.altavista.fr, www.altavista.ch), welche mindestens die Anfrageseite, manchmal auch den Support, speziell in der Landessprache halten, ansonsten aber ein und den gleichen Index abfragen. Im Folgenden ist normalerweise die .com-Version der Suchmaschinen gemeint, wenn die Angabe fehlt.

2.2.3 Metasuchmaschinen

Die Methode der Metasuchmaschinen ist ganz einfach, sie gibt die Suchanfrage an verschiedene roboterbasierte Suchmaschinen weiter, welche diese je einzeln auswerten. Die Resultate werden von der Metasuchmaschine „eingesammelt“ und je nach Maschine dargestellt. Häufig werden die Resultate, egal von welcher Suchmaschine sie kommen, neu gewichtet und geordnet. [Bekavac 02]

3 Wurde AltaVista tatsächlich von Google überflügelt?

3.1 Geschichte

3.1.1 AltaVista

AltaVista, was "Blick von oben" bedeutet, war eines der ersten grösseren Suchwerkzeuge, das im Internet auftauchte. Wissenschaftler der Firma Digital Equipment Corp. suchten nach Möglichkeiten, wie sie die Leistungsfähigkeit ihres neuentwickelten Computers "Alpha 8400 TurboLaser" aufzeigen könnten, und kamen auf die Idee, mit Datenbanken zu arbeiten, einer der häufigsten Computeranwendungen damals. Datenbankanwendungen sollten auf ihrem neuen Computer bedeutend schneller laufen als auf herkömmlichen Systemen. [Hawkins 03] Dazu kam, dass einer der Mitarbeiter eine Software programmiert hatte, um seine umfangreiche E-Mail Korrespondenz zu indizieren und so durchsuchbar zu machen. [Karzaunikat 03] Daraus entwickelte sich die Idee, eine Volltextsuchmaschine für das Internet zu entwickeln, welche sämtliche Wörter auf sämtlichen Seiten indizieren, und so das WorldWideWeb durchsuchbar machen würde. Im Sommer 1995 schickte man zum ersten mal ein Programm namens "Scooter" ins Netz, und war gespannt, was dieses finden würde. Man schätzte das Internet bereits damals auf eine Grösse von etwa 80'000 Server und 30 Millionen Web-Pages. Da der Versuch erfolgreich war und rund 16 Millionen Seiten indiziert wurden, machte man die Suchmaschine, nach abgeschlossener interner Testphase, am 15. Dezember 1995 der Öffentlichkeit zugänglich.



Abbildung 1 Screenshot www.altavista.com, 22. Oktober 1996

Die neue Suchmaschine wurde schnell einmal bekannt und beliebt, sowohl bei Laien, als auch bei Spezialisten. Am ersten Tag zählte man 300'000 Besucher, Ende 1996 bearbeitete AltaVista bereits 19 Millionen Suchanfragen pro Tag und wurde zur führenden Anbieterin von Suchdiensten im Internet (Abbildung 1). Allmählich erweiterte man die Seite aber zum allgemeinen Einstiegsportal (Abbildung 2), der Suchdienst verlor an Bedeutung gegenüber den immer zahlreicher werdenden Zusatzmöglichkeiten wie Gratis-E-Mail, News und ähnliches.

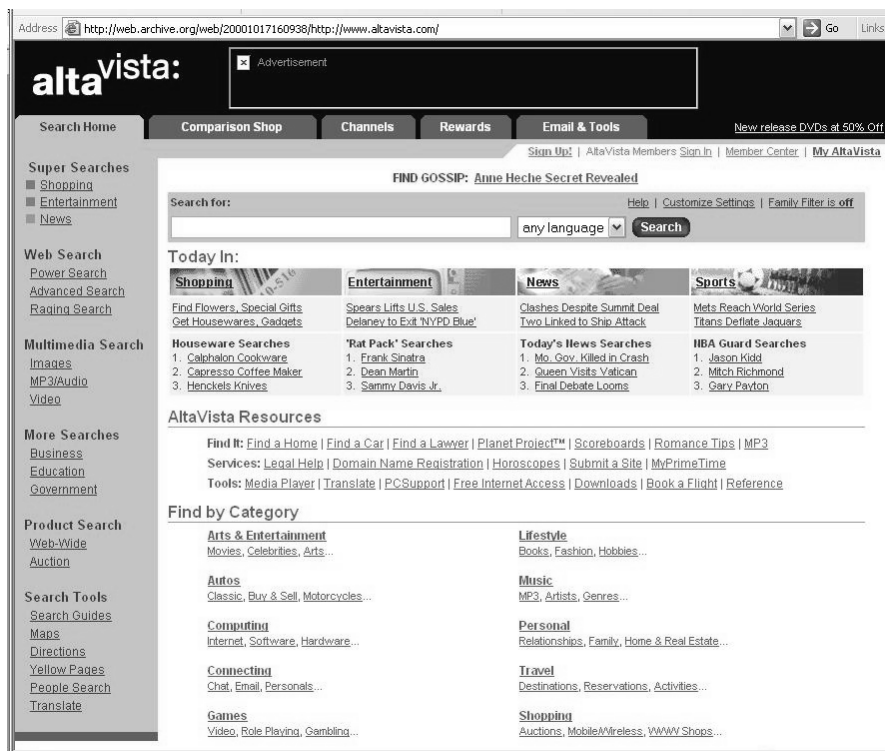


Abbildung 2 Screenshot www.altavista.com, 17. Oktober 2000

Anfang 1998 wurde die Digital Equipment Corporation von Compaq für 9.6 Milliarden US-Dollar übernommen, ein Jahr später die Suchmaschine als The AltaVista Company ausgegliedert. Im Juni 1999 stieg CMGI, eine Internet Investment Firma welche zur Zeit 20% von Lycos besass, mit 83% bei AltaVista ein. Unterdessen hatte die Konkurrenz, bestehend aus Inktomi, Northern Light, Excite und Google, aufgeholt, AltaVista musste Marktanteile abgeben, die Firma war nicht mehr Leader der Branche. Die Jahre 2000 und 2001 waren geprägt vom massiven Stellenabbau und der Neuausrichtung der Unternehmung, welche sich nun wieder auf das Suchangebot konzentrieren wollte, und deshalb die Mitarbeiter für die unzähligen Zusatzdienste entlassen musste. Ende 2002 erhielt die Suchmaschine ein neues Aussehen (Abbildung 3), zahlreiche neue Suchfunktionen wurden implementiert (Bilder, Multimedia) und der erste automatische Internetübersetzungsdienst namens Babel Fish lanciert. Im Februar 2003 ist The AltaVista Company zum Spottpreis von 140 Millionen US-Dollar (drei Jahre vorher wurde der Wert der Unternehmung noch auf 2.3 Milliarden US-Dollar geschätzt) von Overture übernommen worden. [Hawkins 03]



Abbildung 3 Screenshot www.altavista.com, 26. November 2003

3.1.2 Google

Der Name „Google“ stammt von der Bezeichnung „googol“, welche für die Zahl $1 * 10^{100}$ steht, und soll die Vision des Unternehmens unterstreichen, die immense Informationsflut des Internet in den Griff zu kriegen. Ein Hinweis auf die Bedeutung dieser Suchmaschine ist die Tatsache, dass der English Oxford Dicionary das Verb „google“ aufgenommen hat, als Synonym für die Informationssuche im Internet.



Abbildung 4 Screenshot www.google.com, 2. Dezember 1998

Im Januar 1996 entwickelten Lawrence Page und Sergey Brin, Studierende der Universität Stanford, eine neue Suchmaschine namens „BackRub“. Diese verstand es, im Gegensatz zu herkömmlichen Suchwerkzeugen, den eigenen Weg zurückzuverfolgen, den Links auch „rückwärts“ nachzugehen. Daraus entwickelten sie Google, mit dem Ziel, Suchresultate zu generieren, welche zufriedenstellender sein sollten als diejenigen der herkömmlichen Suchmaschinen. Ende 1997 veröffentlichten die beiden ein Papier mit dem Titel „The Anatomy of a Large-Scale Hypertextual Web Search Engine“, welches sich explizit auf die Besonderheiten von Information Retrieval (Aufspürung von Informationen) im WWW bezog und die Beschreibung eines Prototypen für eine „umfassende Suchmaschine“ enthielt. [Brin et Page 98] Dieser Prototyp, Google, mit anfänglich 24 Millionen indizierten Seiten, war der Öffentlichkeit über den Server der Stanford University zugänglich.

In der ersten Hälfte des Jahres 1998 fuhren Page und Brin fort, die Technik ihrer Suchmaschine zu verfeinern und schlugen damals eine Grundrichtung ein, welche bis heute besteht: Sie kauften nicht einzelne, grosse Server, sondern verbanden viele Billig-Computer miteinander auf Basis des Linux-Betriebssystemes. [google.com 03]

Die beiden Informatikstudenten begannen sich nun nach potentiellen Partnern umzusehen, welche ihnen ihre neue Suchtechnologie abkaufen würde. Die Suche nach Investoren gestaltete sich aber, trotz des dotcom-Boom, schwierig. Die führenden Anbieter von Internetportalen erkannten zwar die Stärke des neuen Systemes, aber nicht dessen Potenzial.

Am 7. September 1998 gründeten Page und Brin deshalb ein eigenes Unternehmen mit Namen „Google Inc.“, Sitz in einer Garage in Menlo Park, Kalifornien, einem Startkapital von einer Million US-Dollar und dem ersten angestellten Mitarbeiter, Craig Silverstein, dem Technology Director. Obwohl immer noch in der Betaversion (Abbildung 4), bearbeitete google.com bereits 10'000 Suchanfragen pro Tag, und gewann langsam an Rang und Namen.

Im Februar 1999, die Anzahl der Suchanfragen pro Tag hatte sich um Faktor 50 erhöht, war die Garage zu klein geworden, man zog mit nun acht Angestellten nach Palo Alto um, notabene demselben Ort wo AltaVista zuhause ist. Ein halbes Jahr stand der nächste Umzug vor der Türe, nach Mountain View, California, dem heutigen Sitz der Firma. Der

steile Aufstieg hielt an, Google wurde verschiedentlich ausgezeichnet, erschien in Bestenlisten und entwickelte sich vom Geheimtipp zum Branchenleader.

Im Juni 2000 war die Datenbank der Suchmaschine auf neu eine Milliarde indizierter Seiten angestiegen, den damaligen Höchstwert. Eine Partnerschaft mit yahoo.com im selben Monat unterstrich den endgültigen Wandel der Firma vom Forschungsprojekt Studierender, zu einer ernstzunehmenden Unternehmung, welche 18 Millionen Suchanfragen pro Tag bearbeitete.

Ende 2000 belief sich dieser Wert bereits auf 100 Millionen, wie die Gründer in ihrem Grundlagenpapier vorausgesehen hatten. [Brin et Page 98] Während all der Jahre konzentrierte man sich auf die Suchfunktion und verzichtete auf jedwelchen Ballast auf der Seite. „Die Google Startseite kann weder E-Mails verschicken noch Kaffee kochen.“ [Schöch 01] Dies ist bis heute so geblieben (Abbildung 5), obwohl unterdessen weitere Partnerschaften abgeschlossen, zusätzliche Suchdienste (Bilder, Diskussionsforen, eigenes Verzeichnis, News) implementiert und verschiedene Tools zum Einbinden von Google in verschiedene Software entwickelt wurden.



Abbildung 5 Screenshot www.google.com, 2. Dezember 1998

Mit Froogle brachte das Unternehmen zudem im Jahr 2002 eine Produktesuchmaschine auf den Markt. Aktuell liest man von einer möglichen Übernahme der Firma durch Microsoft, sowie dem allfälligen Börsengang des schwarze Zahlen schreibenden Unternehmens. [Lindner 03]

3.2 Anzahl indizierter Seiten

Folgende Statistiken veranschaulichen die Entwicklung der Grösse des Index der verschiedenen Suchmaschinen, von 1995 bis 2003. Die dabei verwendeten Zahlen stammen von den Betreibern der Internetsuchdienste selbst, aufgeführt sind nur solche Dienste, welche heute noch aktiv sind. [Sullivan-1 03]

Weshalb Google AltaVista überflügelte.

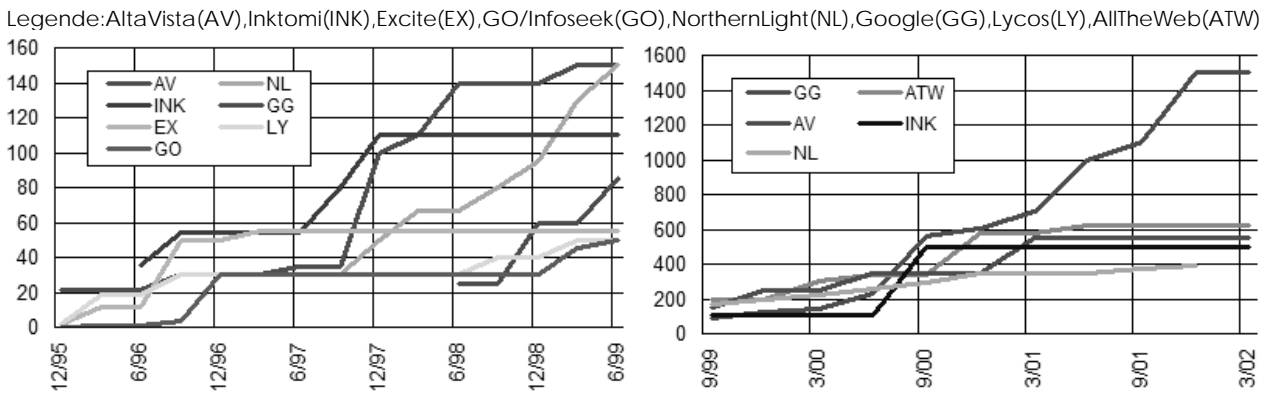


Abbildung 6 Anzahl indizierter Dokumente (in Millionen) Abbildung 7 Anzahl indizierter Dokumente (in Mio)

Im Dezember 1995 startet AltaVista mit dem deutlich grössten Index von gut 20 Millionen Seiten (Abbildung 6). Es folgt ein Wettbewerb mit den anderen Anbietern, in welchem das Unternehmen aus PaloAlto von Mitte 1998, beim Einstieg von Google, bis Mitte 1999, dem Gleichziehen von NorthernLight, klar in Führung liegt (ebenfalls Abbildung 6). Im Verlauf des Jahres 2001 setzt sich Google deutlich von der Konkurrenz ab und indexiert im März 2002 mehr als doppelt so viele Seiten, die Übrerrundung von AltaVista in dieser Kategorie findet zwischen März und September 2000 statt (Abbildung 7).

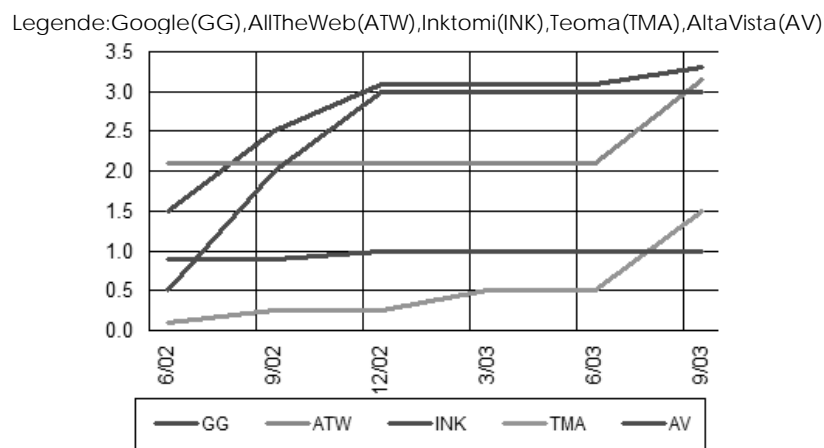


Abbildung 8 Anzahl indizierter Dokumente (in Milliarden)

Heute präsentiert sich die Situation so, dass Google mit einem Index von 3.308 Milliarden Seiten¹ weiterhin an der Spitze liegt, die Konkurrenz aber, insbesondere AltaVista, praktisch gleichaufgezogen ist (Abbildung 8).

3.3 Popularität

Eine Studie, welche die Gewohnheiten im Bezug auf die Informationsbeschaffung mit Suchmaschinen untersucht, ergab, dass im August 2003 für 32% der Suchanfragen Google benutzt wurde und für 1% AltaVista² (Abbildung 9). Zu beachten ist, dass sowohl Yahoo, als auch AOL ihre Anfragen an den Index von Google stellen, so dass der Anteil von Google eigentlich um einiges höher ist.

¹ 3'307'998'701, Stand 15.November 2003, <http://www.google.ch>

² Aufteilung der Suchanfragen (inklusive Multimedia- und Bildsuche) von einer Million US-Bürger auf die verschiedenen Suchmaschinen im August 2003. [Sullivan-2 03]

Weshalb Google AltaVista überflügelte.

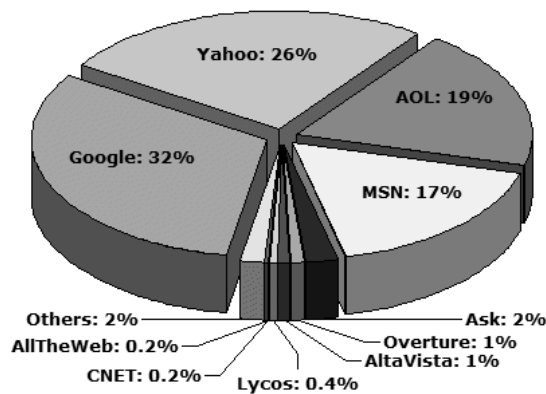


Abbildung 9 Verteilung der Suchanfragen

Legende: Google(GG), AOLSearch(AOL), Yahoo(YH), MSNSearch(MSN), AskJeeves(AJ), InfoSpace(IS)
AltaVista(AV), Overture(OVR), Netscape(NS), Earthlink(ELNK), Looksmart(LS), Lycos(LY)

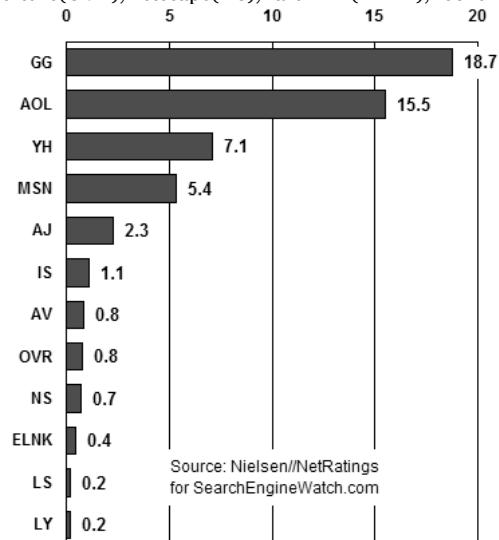


Abbildung 10 Suchmaschinenbenutzung in Millionen Stunden

Eine ähnlich angelegte Untersuchung mass unter anderem, wie lange auf den einzelnen Seiten gesucht wurde. Im Januar 2003 stellten die Testpersonen insgesamt 18.7 Millionen Stunden lang Suchanfragen auf der Website von Google, im Gegensatz zu 800'000 Stunden bei AltaVista³ (Abbildung 10).

3.4 Suchanfragen pro Tag

Die Anzahl der Suchanfragen pro Tag ergibt einen weiteren Hinweis auf die Gewichtigkeit einer Suchmaschine. Die entsprechenden Zahlen (Stand: Februar 2003) sind eindeutig: 250 Millionen Suchanfragen an Google (inklusive der Partnerseiten, also z.B. Yahoo, AOL), 18 Millionen an AltaVista. [Sullivan-3 03]

³ Die Zahlen beziehen sich auf das Suchverhalten im Internet von 60'000 amerikanischen Internetnutzern im Januar 2003. [Sullivan-4 03]

3.5 Zusammenfassung

Während die Geschichte der beiden Unternehmen bereits einige Anhaltspunkte geben, sprechen die Zahlen eine deutliche Sprache. Die Akzeptanz und Bekanntheit von AltaVista war beim Markteintritt von Google sicherlich um ein Erkleckliches über dem Wert des Neulings, was aber heute (November 2003) bei weitem nicht mehr der Fall ist; AltaVista weist eine deutlich tiefere Zugriffs-, sowie Popularitätsrate aus. Die Annahme, dass Google AltaVista überflügelt hat, erweist sich also als richtig, mit Ausnahme der Grösse des Indexes, hier hat AltaVista Google wieder eingeholt.

4 Technik einer roboterbasierten Suchmaschine

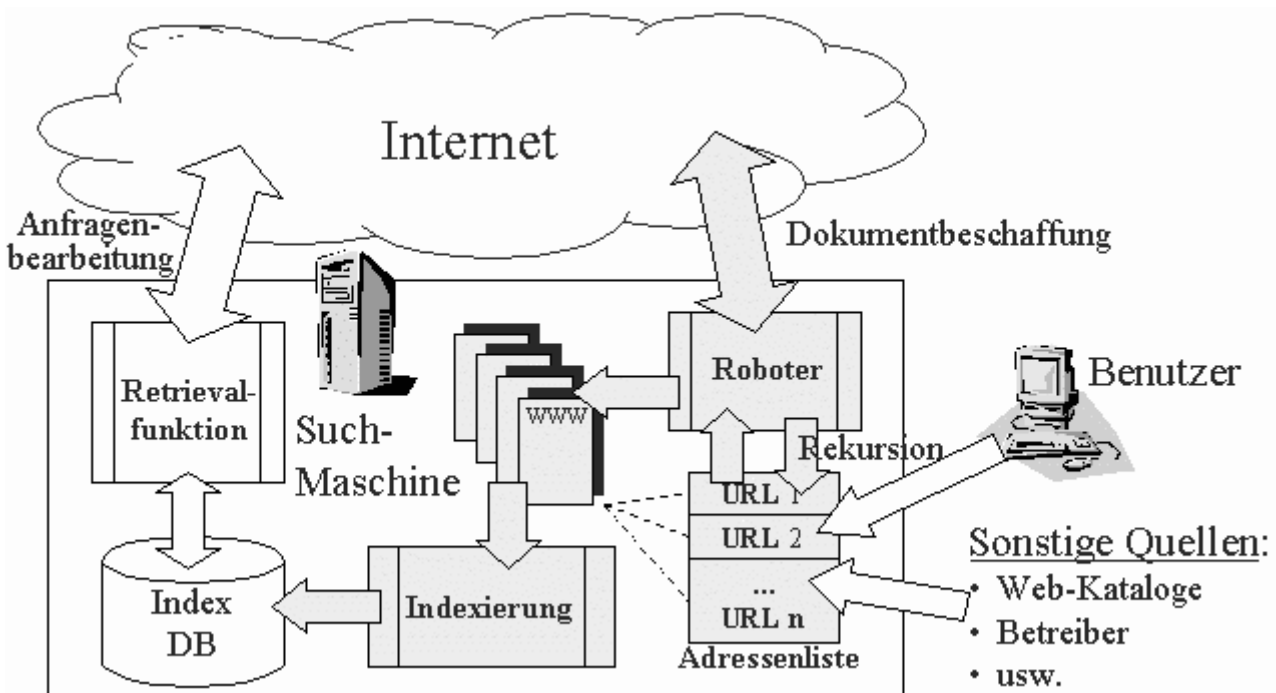


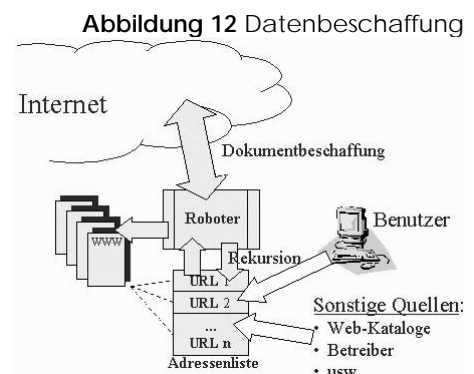
Abbildung 11 Architektur einer roboterbasierten Suchmaschine

Die drei wichtigsten Bestandteile einer roboterbasierten Suchmaschine sind die Datenbeschaffung (Akquisition), deren Indizierung, sowie die Anfragebearbeitung.

4.1 Datenbeschaffung

Ein Robot - auch Bot, Spider, Agent, Crawler, Web Wanderer und Worm genannt - ist ein Programm, welches per Hypertext Transfer Protocol (HTTP) automatisch Seiten aus dem Internet abfragt und deren Verweisen folgt. Die Betreiber von Suchmaschinen benutzen Robots um Datenmaterial für den Index zu erhalten.

Basierend auf einer Liste von Startadressen (Uniform Resource Locator, URL) verlangt der Robot die erste Seite vom entsprechenden Server, analysiert und



speichert sie. Anschliessend folgt er den Links, welche auf der gespeicherten Seite zu finden sind, und folgt so allen Links, bis diese ins Leere führen, eine Seite nicht verfügbar ist, oder eine Seite keine ausgehenden Links mehr aufweist. Dabei sind gleichzeitig hunderte von Verbindungen zu den entfernten Servern offen, damit unvermeidbare Wartezeiten (Antwortverzögerung des angefragten Servers) optimal genutzt werden können. Ergibt die Anfrage nach einer Seite eine Fehlermeldung (zum Beispiel HTTP-Fehler 404 - File not found), so registriert der Robot, weshalb er nicht auf diese Seite zugreifen konnte. [Koster 03, Bekavac 02] Den gesamten Vorgang, bei dem häufig mehrere Robots parallel arbeiten, nennt man auch Harvesting (Ernten, Sammeln). [Karzaunikat-2 03]

Er läuft ununterbrochen ab, damit die Datenbestände stetig vergrössert, und aktualisiert werden können. Bei den meisten Suchmaschinen können auch die Benutzer Internetadressen als Startpunkte angeben, und des weiteren wird häufig ein Katalog (siehe 2.2.1) als Adressenlieferant benutzt.

4.1.1 Was finden Robots?

In erster Linie „finden“ die Robots Dokumente welche mit der Hypertext Markup Language (HTML) geschrieben werden (entspricht den „normalen“ Seiten im WorldWideWeb), es gibt aber auch andere Dienste, welche zum Teil erfasst werden. So gibt es Suchdienste, welche die FTP-Verzeichnisse durcharbeiten, wobei meist nur Pfadnamen und Textdateien (Volltext) untersucht werden können. [Bekavac 02]

4.1.2 Was kann nicht gefunden werden?

Probleme gibt es mit Seiten, welche mit Frames (Rahmen) arbeiten, da die Robots fast ausschliesslich Verweisen folgen, welche durch den normalen HTML-Befehl () definiert sind und keine Methode vorhanden ist, die verschiedenen Teile eines „Frame-Konstrukt“ einander richtig zuzuordnen. [Eipert 00, Bekavac 02]

Weiter werden Seiten nicht gefunden, auf welche kein Link verweist und für die gleichzeitig keine Anmeldung bei Suchmaschinen vorgenommen wird, sogenannte Insel-Seiten.

Neu erstellte Dokumente werden erst erfasst, wenn der Robot diese auf seinem regelmässigen „Rundgang“ entdeckt und indiziert. Diese sogenannten „time-lags“ entstehen auch zwischen Anmeldung bei der Suchmaschine und der Bearbeitung der Anmeldung.

Dynamische Dokumente, welche mit dem Common Gateway Interface (CGI), Java oder JavaSkript aufgrund von Formularanfragen generiert werden, sind ebenfalls für Robots unzugänglich, da sie keine Möglichkeit haben sinnvolle Formulareinträge automatisch vorzunehmen. [Bekavac 02] In der Fachliteratur werden diese Seiten als „Deep Web“, oder „Invisible Web“ bezeichnet, da sie als relativ zahlreich gelten, und im Prinzip unsichtbar sind. Stellt man beispielsweise eine Anfrage an den Webkatalog einer Bibliothek, so wird für jeden Benutzer speziell eine Seite mit den Resultaten generiert. Auf dem angefragten Server befinden sich nur die Informationen in Datenbanken, sowie die Werkzeuge für die Abfragen, so dass man eine praktisch unendliche Anzahl möglicher Webseiten zusammenstellen könnte, aber auf dem Server trotzdem keine einzige Website „sichtbar“ ist.

4.1.3 Was darf nicht gefunden werden?

Natürlich gibt es auch Internetseiten mit Daten, welche nicht der breiten Öffentlichkeit zur Verfügung gestellt werden sollen, und keinesfalls in den Index einer Datenbank geraten dürfen. Für solche Fälle wurde der „Robot Exclusion Standart“ entwickelt, ein von

den meisten Robots eingehaltener Standard, wonach auf jedem Server im Wurzelverzeichnis zuerst ein Textdokument namens „robots.txt“ bearbeitet wird (Beispiel: www.unifr.ch/robots.txt). Dieses enthält Anweisungen, welche Seiten und Verzeichnisse auf dem betreffenden Server von welchen „Spürhunden“ indiziert werden dürfen, und was zu ignorieren ist.

Nun ist nicht jeder Anbieter einer Website gleichzeitig der Besitzer des Webservers, daher, es gibt mehrere Homepages auf demselben Server, aber nur eine „robots.txt“-Datei. Um einzelne Seiten vor dem Zugriff der Crawler zu schützen, gibt es die Möglichkeit, dies im jeweiligen Header der HTML-Struktur zu deklarieren (dazu Anhang A), was allerdings nur von wenigen Robots unterstützt wird.

Des Weiteren tabu sind durch Passworte, Firewall oder Registrierung zugriffsgeschützte Gebiete, so sämtliche E-Mail-, E-Banking und anderen privaten Onlinekonten.

4.2 Indizierung

Die Qualität der Suchergebnisse einer Suchmaschine sind stark davon abhängig, welche Güte der Index aufweist, daher in welcher Weise indiziert wurde. Abbildung 14 zeigt ein etwas detaillierteres Schema, des heute üblichen Vorganges, und ist angelehnt an die Struktur von Google.

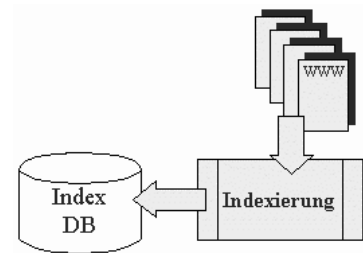


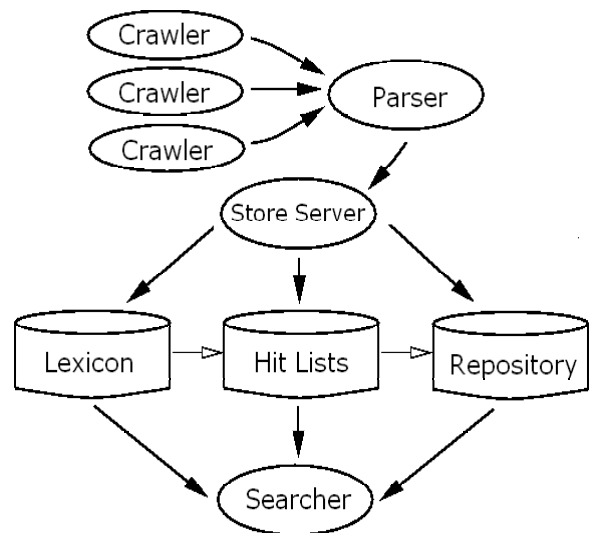
Abbildung 13 Indizierung

Abbildung 14 Struktur der Indizierungswerkzeuge

4.2.1 Parser

Der Parser ("Zerteiler") ist ein Programm, welches die syntaktische Analyse der von den Robots (Crawler) gespeicherten Dokumente übernimmt, das heißt, er untersucht den Quelltext und "liest" diesen gewissermassen (ähnlich einem Browser). Er erstellt einen Ableitungsbaum, in welchem allen Daten des Dokumentes ihr Wert zugeordnet wird. Das heißt, er erkennt die HTML-Struktur und identifiziert URL, Metadaten, Bodytext, Überschriften, Kommentare und ähnliches als solches. In anderen Worten: Jedes Wort wird in vordefinierte "Schachteln" verpackt. Zusätzlich analysiert der Parser den Quelltext

auf Fehler, welche auch aufgrund der vielen verschiedenen HTML-Versionen, welche im WWW vertreten sind, relativ häufig sind. Der resultierende Ableitungsbaum wird an den Store Server weitergeleitet, welcher nun die Daten in der ihm bekannten Variante (dem Ableitungsbaum) erhält. Mit dem Bild der Schachteln ausgedrückt: Es gibt keine Daten mehr, welche lose herumschwirren. Alles ist mit dem passenden Etikett versehen, in der passenden Schachtel eingeordnet. Je nach Parser werden zudem hier die als Suchbegriffe identifizierten Wörter mit verschiedenen möglichen Methoden auf ihren Stamm zurückgeführt⁴. [Wichmann 99, Schulzki 00]



⁴ Wird die Suchanfrage ebenfalls auf den Wortstamm reduziert, so erhält man, gibt man als Suchbegriff die Pluralform eines Wortes ein, nicht nur alle Seiten, welche den Begriff im Plural enthalten, sondern ebenso alle, welche den Singular (daher den Wortstamm) enthalten. [Schulzki 00]

4.2.2 Store Server

Der Store Server entnimmt dem vom Parser erhaltenen Ableitungsbaum die für die Suche wichtigen Daten, und verteilt sie: Die für die Suche relevanten Begriffe werden im Lexicon abgelegt, die diese Begriffe beschreibenden Daten in den Hit Lists, die komprimierten Websites im Repository und externe Links werden wieder den Robots gespielt.

Das System mit Lexicon, Hit Lists und Repository funktioniert nach dem Prinzip des „invertierten Index“. [Schulzki 00]

Die eine Möglichkeit einen Index aufzubauen besteht darin, zu jedem Dokument anzugeben, welche Suchbegriffe darin enthalten sind. Die Idee des invertierten Index ist, dass man die Suchbegriffe auflistet und dazu angibt, in welchen Dokumenten diese überall zu finden sind und welche Funktion sie besitzen.

4.2.3 Lexicon

Im Lexikon wird jeder Begriff gespeichert, der irgendwo auf einer der vom Robot gesammelten Seiten vorkommt, abzüglich der sogenannten Stoppwörter (Artikel, und, oder, sonstige Füllwörter). Es gibt jedoch Suchmaschinen (AltaVista), welche eine komplette Liste von Wörtern führt, so dass beispielsweise eine Suche nach den beiden Begriffen „das“ und „Hund“ ein anderes Resultat ergibt, als eine Suche nach „der“ und „Hund“. Für jeden Begriff aus dem Lexicon gibt es einen spezifischen Datensatz in den Hit Lists, mit welchem er „verbunden“ ist.

4.2.4 Hit Lists

In den Hit Lists sind die beschreibenden Daten abgelegt, das heisst, zu jedem Begriff ist angegeben, in welchem Dokument des Repository er vorkommt und welche Funktion ihm im Dokument selbst zukommt.

So wird registriert, an welcher Stelle im Dokument der Term erscheint, ob in der Adresse, im Titel, im Header, am Anfang oder am Ende des Body. Im Header der HTML-Struktur gibt es die sogenannten Metadaten, welche der Urheber einer Seite selbst eintragen kann, so zum Beispiel den Titel, Stichworte, eine kurze Beschreibung des Inhaltes, Angaben zum Autor, Erstellungsdatum und ähnliches (dazu Anhang A). Indiziert wird also auch, wo genau im Header sich ein Wort befindet.

Auch die Umgebung, also welche Worte in der Nähe eines Begriffes sind und welche nicht, wird aufgenommen, wie auch die Häufigkeit eines Begriffes im Verhältnis zur Gesamtlänge des Textes. [Schulzki 00]

Einige Suchmaschinen (z.B. Google) registrieren hier auch, ob ein Term Teil eines Linktextes zu einem Bild, einer anderen Seite oder einer Stelle im selben Dokument ist (auch Anchor Text genannt). [Brin et Page 98]

4.2.5 Repository

Im Repository werden sämtliche von den Robots (Crawler) aufgespürten, und dem Parser komprimierten Dokumente abgelegt. Via Hit Lists sind alle Suchbegriffe aus dem Lexicon mit denjenigen Seiten in welchen sie auftreten verbunden. Diese Suchbegriffe „zeigen“ gewissermassen auf das entsprechende Dokument. Diese (vektoriellen) Verknüpfungen sind für den Suchvorgang von erheblicher Bedeutung.

4.2.6 Aktualisierung der Daten

Um den Index aktuell zu halten, müssen die Daten während des Betriebs stets wieder auf Existenz und neue Inhalte überprüft werden, damit allfällige Änderungen aufgenommen werden können. Der Aktualisierungsvorgang entspricht in groben Zügen demjenigen der Indizierung. Schon vorhandene Daten werden „überschrieben“, neue hinzugefügt und

Seiten welche nach mehrmaligem Anfragen nicht verfügbar sind, werden gelöscht. [Bekavac 02]

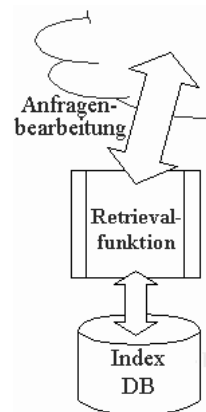
Die Aktualisierungsfrequenz ist von Suchmaschine zu Suchmaschine verschieden, und liegt zwischen einigen Tagen und mehreren Wochen. [Karzaunikat-2 03] Dadurch entsteht das Problem der toten Links (auch "Dangled Links"), das heisst, dass Suchresultate Dokumente angeben, welche unterdessen eine andere Adresse haben, oder überhaupt nicht mehr existieren. So kann es auch geschehen, dass eine Seite zwar noch die gleiche URL hat, aber der Inhalt total verändert wurde, so dass die Suchterme, welche im Index zu dieser Seite gespeichert sind, gar nicht mehr zutreffen. Schätzungsweise 10 – 15% des Index einer Suchmaschine sind Dangled Links. [Wichmann 99]

Eine höhere Aktualisierungsfrequenz ist schwierig zu erreichen, da die Ressourcen, die Grösse und das Wachstum des WorldWideWeb, hier einschränken. Deshalb aktualisieren die meisten Suchmaschinen nicht alle Seiten gleich häufig, sondern versuchen anhand verschiedener Kriterien (PageRang, Häufigkeit des Auftretens in Resultatlisten, etc) zu differenzieren, so dass Seiten mit häufig wechselndem Inhalt fleissiger von den Robots „besucht“ werden.

Abbildung 15 Anfragebearbeitung

4.3 Anfragebearbeitung

Die Anfragebearbeitung stellt die Schnittstelle zwischen Benutzer und Suchmaschine dar. Per Formular wird eine Suchanfrage formuliert, welche die Suchmaschine mit den indizierten Daten (Lexicon, Hit Lists) vergleicht. Treffer werden auf einer neu generierten Seite, mit zusätzlichen Informationen zu den einzelnen Dokumenten aus dem Repository, nach Relevanz geordnet dargestellt. [Wichmann 99]



4.3.1 Anfragemöglichkeiten

Der Benutzer hat je nach Anbieter viele verschiedene Möglichkeiten seine Suchanfrage zu gestalten. Grundsätzlich bieten die meisten Suchmaschinen zwei verschiedene Formulare (Modi) zur Eingabe der Anfrage an: Einerseits eine einfache Suche (Simple Search), was eine triviale Stichwortsuche bezeichnet, sowie andererseits eine erweiterte (Extended-, Advanced-, Power Search), welche die Angabe zusätzlicher Kriterien und Operatoren ermöglicht.

So ist es möglich das Suchgebiet einzugrenzen, das heisst Beschränkung auf eine Sprache, ein Land (Länderendungen in der URL), ein Medium (Bild, Audio, Film), ein Format (Portable Dokument Format PDF, Word, Excel, etc), einen Server oder auch auf ein bestimmtes Alter der Dokumente.

Einzelne Suchbegriffe können mit den Boole'schen Operatoren miteinander verknüpft werden, indem die drei Begriffe „AND“, „OR“, sowie „AND NOT“ dazwischen gesetzt werden. „AND“ bedingt alle so verbundenen Worte auf einer Seite, „OR“ mindestens eines davon und „AND NOT“ schliesst Seiten aus, welche den bezeichneten Begriff enthalten. Zudem ist manchmal auch eine „NEAR“-Anweisung möglich, welche eine gewisse Nähe zwischen Wörtern verlangt.

Setzt man ein Pluszeichen (+) unmittelbar vor einen Term, so muss dieser zwingend im Dokument erscheinen (erzwingt auch Stoppwörter). Wird ein Minus (-) vorangestellt, werden Seiten ausgeschlossen, welche den betreffenden Ausdruck enthalten. Wird weder ein Boole'scher Operator, noch ein Plus- oder Minuszeichen verwendet, so werden mehrere Begriffe in einer Anfrage wie mit einem „OR“ dazwischen behandelt.

Möchte man eine bestimmte Phrase finden, so kann man die entsprechenden Wörter auch in Anführungszeichen setzen, was den gleichen Effekt hat, wie eine Verknüpfung mit

„AND“. Zudem können diese Phrasen mit den Plus- und Minuszeichen kombiniert werden.

Eine weitere Variation ist die Trunkierung (auch Wildcard genannt), welche durch einen Stern (*) als Platzhalter ausdrückt, wo beliebige Buchstaben vorkommen dürfen. Wird jedoch von praktisch keiner Suchmaschine unterstützt.

Setzt man „anchor:“ (AltaVista) vor einen Suchterm, so werden als Resultat Seiten geliefert, welche diesen Begriff in einem Verweistext enthalten haben. Es gibt noch mehrere solcher HTML-Anweisungen mit welchen man ganz gezielt die indizierten beschreibenden Daten ausnützen kann (dazu Anhang B).

Schliesslich gibt es auch noch die Möglichkeit in der Resultatmenge einer vorgängigen Anfrage zu suchen, um so die Ergebnisse zu verfeinern. [Barker 03; Fischer 99]

4.3.2 Darstellung der Suchresultate

Trifft eine Anfrage bei einer Suchmaschine ein, so gleicht diese sie mit dem Index (Lexicon, Hit Lists) ab und leitet eine Liste mit den Links zu passenden Seiten an den Benutzer zurück. Da nicht alle Resultate gleich relevant sind, werden die Verweise von der Suchmaschine nach Wichtigkeit geordnet dargestellt (Ranking). Vergewärtigt man sich nun, dass Anfragen immense Mengen an Links zu passenden Seiten zurückgeben, und Benutzer nicht Stunden damit verbringen möchten, diese Listen durchzuarbeiten, folgt daraus, dass die Gewichtung der Suchergebnisse eminent wichtig ist für den Erfolg einer Suchmaschine. Besonders bedeutend ist, dass die besten Resultate ganz am Anfang der Resultatliste erscheinen.

Um eine einzelne Seite bezüglich der Relevanz für eine bestimmte Suchanfrage zu bewerten, werden die beschreibenden Daten (4.2.4), welche in den Hit Lists indiziert sind gewichtet. So definiert man beispielsweise, dass ein Wort wichtiger für eine Seite ist, wenn es im Titel, in einer Überschrift oder im Text häufig vorkommt. So kann eine eindeutige Rangliste erstellt werden. Problematisch ist bei diesem Bewertungssystem, dass es relativ anfällig für Tricksereien und Betrügereien ist.

Ein neuerer Ansatz (Google, PageRank) macht sich die Struktur des Internets zunutze, indem er die Linkstruktur untersucht. Ein Link von Seite A zu Seite B wird als Votum von A für B verstanden, wonach also auf eine wichtige Seite viele Links zeigen, auf eine unwichtige wenige. Die Verweise von einer wichtigeren Seite werden höher gewichtet als diejenigen von unwichtigeren. So kann eine Rangfolge erstellt werden, welche davon ausgeht, dass eine Seite mit zahlreichen, auf sich gerichteten Links stets relevanter ist als eine Seite mit wenigen solchen Verweisen. [Brin et Page 98]

5 Stärken von Google

Das Google stark ist, wurde bereits gezeigt. Nun sollen anhand der Architektur von Google (Abbildung 16) die spezifischen Abweichungen zum allgemeinen Modell (aus Punkt 4) erläutert werden, unter der Annahme, dass sich die Suchmaschine dank ihnen durchgesetzt hat. Auf die Eigenheiten von AltaVista wird hier nicht eingegangen, da diese nicht zum selben Erfolg geführt haben.

Zudem soll an dieser Stelle auch auf die Auswertung von Untersuchungen bezüglich der Qualität der Suchresultate verzichtet werden, unter der Annahme, dass der einzelne Nutzer versucht seinen Nutzen zu maximieren, und so automatisch die Suchmaschine mit den (subjektiv) besten Resultaten auswählt. Somit zeigt die aggregierte subjektive Nachfrage der Benutzer automatisch den Anbieter mit den objektiv besten Suchresultaten,

womit die Punkte 3.3 (Popularität), sowie 3.4 (Suchanfragen pro Tag) als genügend aussagekräftig eingeschätzt werden.

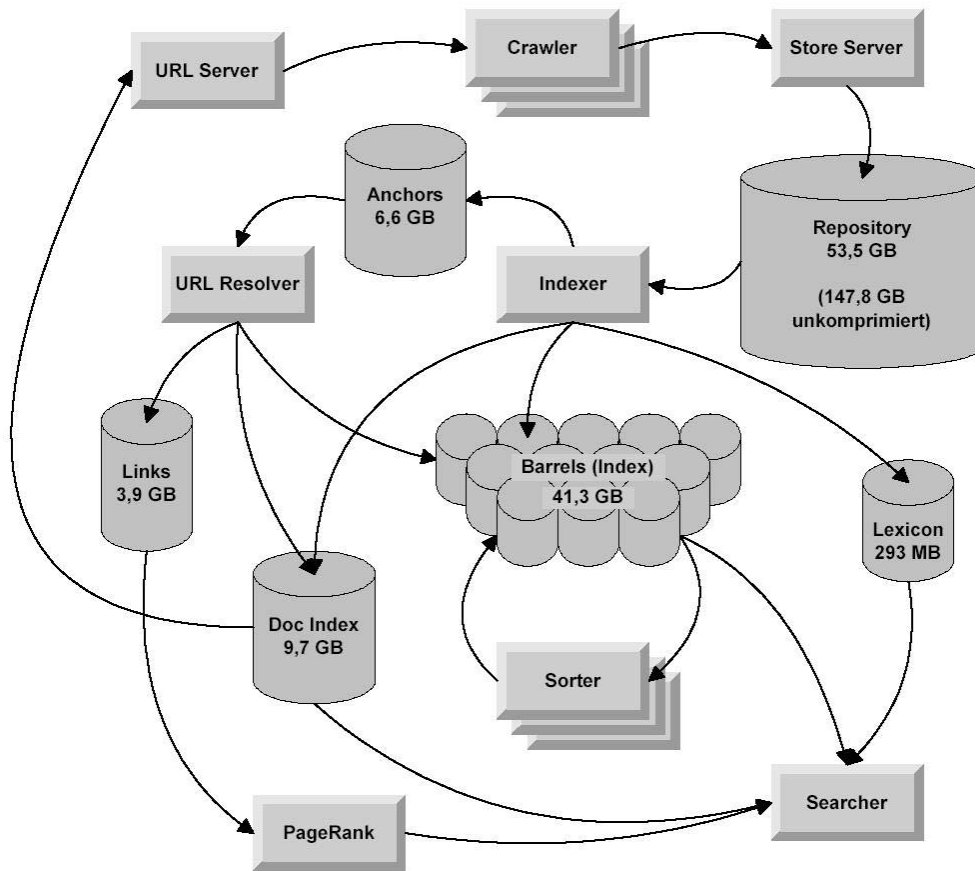


Abbildung 16 Systemarchitektur von Google (Stand Zahlen: Ende 1997)

5.1 Crawler

Wie unter Punkt 3.2 (Anzahl indizierter Seiten) gezeigt besitzt Google den grössten Index, wenn auch nur noch knapp. Dies lässt auf leistungsfähige Crawler schliessen. Zudem sind diese in der Lage neben den "normalen" HTML-Seiten auch andere Formate wie Adobe Acrobat PDF (.pdf), Adobe Postscript (.ps), Microsoft Word (.doc), Microsoft Excel (.xls), Microsoft Powerpoint (.ppt) und Rich Text Format (.rtf) zu verwalten (Stand: 26. November 2003).

5.2 Repository

Das Repository enthält den komprimierten, aber vollständigen HTML-Text, eine eindeutige docID (Nummer, welche ein Dokument im ganzen System eindeutig identifiziert), sowie die URL zu jeder einzelnen Seite. Heute kann bei Google zu jedem Resultat auch die im Repository zuletzt gespeicherte Variante der entsprechenden Seite abgerufen werden, was sehr nützlich sein kann, beispielsweise bei Dangled Links.

5.3 Anchor

Der Linktext wird normalerweise als Information über die Seite auf welcher er zu finden ist interpretiert. Google geht einen Schritt weiter, indem ein Link als eine qualitative,

positive Aussage über eine andere Seite angesehen wird, weshalb dieser Text als hochwertige Information zur Seite auf welche er verweist zu betrachten ist.

Dies bringt einige Vorteile. So beschreibt ein solcher "Ankertext" eine fremde Seite häufig besser als diese sich selbst, da der Autor des Linktextes dem Besucher seiner Seite kurz und knapp mitteilen möchte, was sich hinter dem Link verbirgt, was genau dem Wunsch des Benutzers einer Suchmaschine entspricht.

Zudem ermöglicht dieses System Seiten mit Inhaltsangaben zu indizieren, welche vom Typ her nicht indizierbar sind (Punkt 4.2.1), sowie Seiten welche noch nicht besucht werden konnten. Dies führt zu einer massiven Vergrößerung des Indexes. Als Nachteil ist anzuführen, dass so natürlich auch Seiten in den Index geraten können, welche nie existiert haben. Durch das Ranking der Resultate sollte dieses Problem jedoch eher selten auftreten.

In der Anchor-Datenbank ist jeder gefundene Link gespeichert, mit Ankertext, sowie Angaben, auf welcher Seite er sich befindet und wohin er verweist. Zudem wird der Ankertext in den Index aufgenommen, mit der docID der referenzierten Seite.

5.4 Barrels

Der Indexer übernimmt die Funktion des Parser, indem er die Seiten zu Wortlisten kürzt und diese ins Lexicon speichert, wo jedes Wort eine wordID (Ziffer, welche jedes Wort im ganzen System eindeutig identifiziert) erhält. Dann wird ein sogenannter Forward Index erstellt, welcher in den Barrels gespeichert wird. Der Forward Index ist nach docID geordnet, zu welchen jeweils die Verweise zu den passenden Suchtermen angegeben ist (wordID). Diese eher kleinen, aber zahlreichen Speichereinheiten sollen einen hohen Grad an Parallelität und Prozess-Lokalität ermöglichen. So wird der Forward Index vom Sorter stetig bearbeitet, um daraus einen nach wordID's geordneten Inverted Index zu erstellen, welcher ebenfalls in den Barrels abgelegt wird. Zudem werden durch den Sorter auch ständig Platz- und Zugriffsoptimierungen vorgenommen.

5.5 Links

Der URL Resolver, welcher aus den Angaben in der Anchor-Datenbank die absoluten URL's berechnet, speichert diese in die Links-Datenbank, wo alle docID's in ein Verhältnis bezüglich der Verweise gesetzt werden. Dies geschieht in Form einer riesigen quadratischen Matrix, aus welcher der PageRank berechnet werden kann (dazu Punkt 5.6.2). [Brin et Page 98]

5.6 PageRank

Mit dem PageRank erhält jedes Dokument einen Rang, so dass man eine Rangliste aller Seiten im WorldWideWeb erstellen könnte. Dieser Ranking-Algorithmus stellt ein Herzstück von Google dar. Dabei wird das WorldWideWeb als riesigen gerichteten Graphen angesehen, wobei die Seiten Knotenpunkte, die Verweise die Kanten darstellen.

5.6.1 Eigenschaften guter Ranking-Algorithmen

Es war das erklärte Ziel der wissenschaftlichen Arbeit, aus der Google entstanden ist, „gute“ Ranking-Algorithmen zu finden und in einer realistischen Anwendung zu erproben. [Schöch 01] Nun soll zuerst definiert werden, was unter einem „guten“ Ranking-Algorithmus zu verstehen ist. Die folgenden vier Eigenschaften sind [Schöch 01] entnommen.

Geschwindigkeit: Die Antwortzeit einer Suchmaschine ist – zusammen mit einem guten Ranking – eines der wichtigsten Kriterien für die Nutzer-Akzeptanz. Daher müssen

zeitaufwendige Berechnungen offline im Voraus vorgenommen werden. Alle Algorithmen, die online bei der Bearbeitung einer Suchanfrage laufen, müssen extrem schnell sein.

Skalierbarkeit: Das WWW übertrifft schon jetzt im Umfang alle praktisch relevanten Datenbanken, und die Anzahl der Dokumente im Web verdoppelt sich etwa alle 3–6 Monate. Ähnliches gilt für die Nutzerzahlen des WWW. Damit eine Suchmaschine in Zukunft noch nutzbar bleibt, müssen die verwendeten Algorithmen extrem gut skalieren.

Spamresistenz: Es gibt bei Suchmaschinen immer zwei Gruppen von Interessenten: Die Suchenden und die Gefundenen. Für viele Websites ist die Anzahl der Besucher gleichbedeutend mit Umsatz und Geschäft. Daher setzen die Betreiber dieser Websites alles daran, die Ranking-Algorithmen der großen Suchmaschinen gut kennen zu lernen und ihre Seiten darauf zu optimieren. Für Spitzenpositionen in der Trefferliste sind oft auch sehr merkwürdige Mittel und Wege recht – klassische Beispiele sind "Text in Hintergrundfarbe" oder auch spezielle "Brückenseiten", die menschliche Internetnutzer nicht zu Gesicht bekommen. Im Extremfall führt das zum sogenannten Index-Spamming (Überflutung des Index mit irrelevanten Seiten durch gezielte Manipulation des Suchalgorithmus): Die "Treffer" einer Suchmaschine werden unbrauchbar, weil nicht wirklich relevante Seiten als erstes aufgeführt werden, sondern solche, die den Index am erfolgreichsten manipuliert haben. Um das zu verhindern, sollte ein guter Ranking-Algorithmus schwer zu manipulieren, also spamresistent, sein.

Plausibilität: Das einzige, was für den Anwender letztlich zählt, ist die subjektive Zufriedenheit. Um das zu erreichen, müssen die Prinzipien, nach denen eine Suchmaschine das Ranking der Treffer durchführt, dem Anwender plausibel und sinnvoll erscheinen. Ein theoretisch perfekt durchdachtes Ranking nützt nichts, wenn der Anwender das Ergebnis nicht nachvollziehen kann.

5.6.2 Definition des PageRank-Algorithmus

Der Notation von [Page et al 98] folgend bezeichnen wir mit u eine beliebige Webseite, F_u steht für die Menge aller Seiten auf welche von u aus ein Link führt (Forward Links), und B_u für die Menge der Seiten welche auf u verweisen (Backward Links). $N_u = |F_u|$ sei die Anzahl Forward Links auf der Seite u und c ein Normalisierungsfaktor, um das Total der PageRanks konstant zu halten. $R(u)$ stellt eine vereinfachte Version des PageRanks dar.

$$R(u) = \frac{c \sum_{v \in B_u} R(v)}{N_u}$$

Abbildung 17 Vereinfachte Definition des PageRank

Zur Veranschaulichung der Formel soll Abbildung 18 dienen, welche ein „Miniweb“ mit stabiler PageRank-Bewertung darstellt. In diesem Beispiel ist $c=1$, in Realität ist aber $c<1$ weil es viele Seiten gibt, welche keine Forward Links haben, und deren Gewichtung dem System verloren geht. Die konkreten PageRanks berechnet man, indem man den Seiten

zuerst einen Startwert zuweist (zum Beispiel $\frac{1}{\text{Anzahl zu bewertende Seiten}}$) um anschliessend mit mehreren Iterationen, einen akzeptablen Näherungswert zu erhalten⁵.

⁵ Unter der Adresse <http://www.markhorrell.com/seo/pagerank.asp> steht ein „PageRank Calculator“ zur Verfügung, welcher das entsprechende Verfahren schön veranschaulicht.

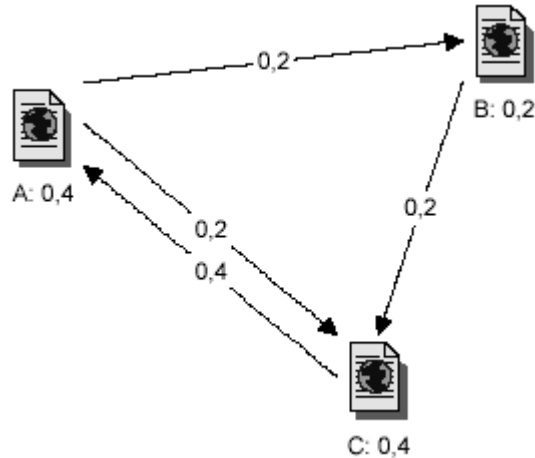


Abbildung 18 Stark vereinfachtes Modell des WorldWideWeb

Man kann die Link-Struktur des Internet auch als riesige quadratische Matrix darstellen, wobei jede Linie, sowie jede Spalte eine Seite darstellt (Abbildung 19). Die einzelnen Werte sind entweder gleich Null wenn es keinen Link gibt, oder aber $1/N_u$ falls es einen Verweis gibt. Dies ist die Matrix, welche in Links (Abbildung 16) gespeichert ist.

	A	B	C
A	0	0	$\frac{1}{1}$
B	$\frac{1}{2}$	0	0
C	$\frac{1}{2}$	$\frac{1}{1}$	0

$$= \mathbf{K} = \begin{pmatrix} 0 & 0 & \frac{1}{1} \\ \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{1} & 0 \end{pmatrix} = \text{Matrix A}$$

Abbildung 19 Beispiel für Darstellung des WWW als Matrix

Der PageRank entspricht dem dominanten Eigenvektor der Matrix A. In [Page et al 98] ist die mathematische Beweisführung dazu zu finden, auf welche hier aber verzichtet werden soll.

Diese vereinfachte Definition von PageRank enthält noch keine Massnahmen, um sogenannte Rank Sink auszuschliessen. Hierbei handelt es sich um Seiten, welche untereinander einen geschlossenen Kreislauf bilden und zwar einen von aussen eingehenden Link, aber keinen nach ausserhalb des Kreislaufes gehenden Forward Link aufweisen (Abbildung 20).



Abbildung 20 Sink Rank

Dies führt mit jeder Iteration zu einer Akkumulierung des PageRanks dieser Seiten, jedoch zu keiner Weitergabe der Gewichtung. Um diesen Effekt auszugleichen wird bei jeder

Iteration allen Seiten ein "Bonus" zugeschlagen, so dass zwar der PageRank in solchen Rank Sink's bei jeder Iteration steigt, dies nun aber alle andern Seiten auch tun, so dass die Verhältnisse wieder einigermaßen stimmen. Dieser Bonus wird Rank Source genannt, und ist als Vektor zu verstehen.

Unter Berücksichtigung dieses zusätzlichen Vektors, E genannt, kann der PageRank $R'(u)$ folgendermassen definiert werden:

$$R'(u) = \frac{\sum_{v \in B_u} R'(v)}{N_v} + c \cdot E$$

Abbildung 21 Vollständige Definition des PageRank

Die Rechtfertigung des PageRank ist das sogenannte Random Surfer Model. Dieses geht davon aus, dass ein Surfer zufällig von einer Seite auf die nächste Seite kommt und die Wahrscheinlichkeit, welchen Link er wählt für jeden Link einer Seite gleich ist (PageRank wird auf Anzahl Links pro Seite gleichmässig verteilt). Wird der PageRank als Wahrscheinlichkeitsverteilung über das ganze Web angesehen, so entspricht er der Wahrscheinlichkeit, mit welcher der Random Surfer sich auf einer bestimmten Seite des Web befindet, wobei der "Bonus"-Vektor E (Rank Source) der Wahrscheinlichkeit entspricht, mit der der Surfer aufhört den Links zu folgen und eine neue, zufällige, Startadresse wählt.

Ein weiteres Problem sind die Dangled Links, also Seiten mit Links welche ins Nichts führen. Diese werden entfernt und nachdem für alle anderen Seiten der PageRank berechnet wurde, wieder eingefügt, so dass der entsprechende Page Rank aus den anderen Werten approximiert werden kann. [Brin et Page 98]

5.6.3 Bewertung des PageRank-Algorithmus

Die Bewertung von PageRank erfolgt aufgrund der Kriterien aus 5.6.1 und folgt grösstenteils [Schöch 01].

Geschwindigkeit: Da der PageRank einer Seite unabhängig ist von einer konkreten Suchanfrage, kann er im Voraus offline berechnet werden. Insofern spielt die Effizienz des Algorithmus eigentlich eine untergeordnete Rolle. Dass die Implementierung von [Page et al 98] schnell genug ist für den praktischen Einsatz, wird durch Zahlen belegt: Auf einer Datenbasis von 25 Millionen erfassten Seiten und insgesamt 75 Millionen indizierter URL's (Stand Ende 1997) ist es möglich, den gesamten Prozess zur hinreichenden Approximierung des PageRank in 5 Stunden durchzuführen, obwohl nur die Hälfte des Graphen im Hauptspeicher gehalten werden kann.

Skalierbarkeit: Eine genaue Berechnung des dominanten Eigenvektors im n-dimensionalen Raum (n=Anzahl Webseiten) mit $n \approx 25 \cdot 10^6$ ist in akzeptabler Zeit nicht realisierbar. Man begnügt sich daher mit einer hinreichend genauen Approximation. Als "hinreichend genau" sehen [Page et al 98] eine Gesamtabweichung pro Iteration von weniger als 100 an. Das entspricht durchschnittlich weniger als $4 \cdot 10^{-6}$ pro Seite bei einem durchschnittlichen PageRank von eins. Um diese Genauigkeit zu erreichen, sind nach Angaben der Autoren bei 161 Millionen URL's 45 Iterationen notwendig, für doppelt so viele URL's werden 52 notwendige Iterationen angegeben. [...] Daher kann man davon ausgehen, dass der PageRank-Algorithmus auch auf extrem großen Datenmengen sehr gut skaliert.

Spamresistenz: Der PageRank einer einzelnen Webseite wird theoretisch durch den PageRank jeder anderen Seite im betrachteten Graphen beeinflusst. Dadurch ist es für

einen einzelnen Webseiten-Betreiber sehr schwierig, den PageRank bestimmter Seiten zu manipulieren.

Plausibilität: Der Erfolg gibt Google recht: Offenbar werden die von Google erzeugten Suchergebnisse von vielen Anwendern als zutreffend und hilfreich empfunden. Die User-Feedback Strategie (dazu Punkt 5.7.2) hilft dabei, diesen Aspekt des Rankings zu bewerten und zu verbessern.

5.7 Searcher

5.7.1 Suchvorgang

Als erstes verwandelt Google die Suchanfrage in eine Wortliste, welche vom System gelesen werden kann, und weist den Wörtern die entsprechenden wordID's zu. Anschliessend wird der Forward Index, welcher auf die verschiedenen Barrels verteilt ist, durchsucht, und mit den wordID's verglichen. Für die gefundenen Dokumente wird eruiert, wieviele Treffer in welcher Kategorie zu verbuchen sind (Beispiel: Ein Wort der Suchanfrage einmal im Titel, dreimal in den Meta tags, 15 mal fett geschrieben im Body, vier mal im Body grösser geschrieben, sowie zwei Wörter der Suchanfrage zwanzig mal nahe beieinander im Body). Jede Kategorie besitzt ein festes Gewicht, womit die Treffer verrechnet werden, was schlussendlich einen ersten Rang für jedes Resultat ergibt. Anschliessend wird der Rang einer jeden Seite mit dem jeweiligen PageRank verrechnet, womit die endgültige Sortierung entsteht, in welcher die Resultate für den Benutzer aufgelistet werden. [Brin et Page 98]

5.7.3 User-Feedback Strategie

Die Bewertung der einzelnen Kategorien (5.7.1) stellt ein relativ schwieriges Unterfangen dar. Auswirkungen auf die Qualität der Suchresultate aufgrund von Veränderung dieser Parameter sind schwer nachzuvollziehen. Damit die Betreiber von Google trotzdem eine Ahnung haben, wie sich die Veränderungen einzelner Parameter auf die subjektive Anwenderzufriedenheit bezüglich der Suchresultate auswirkt, besitzen einige vertrauenswürdige Benutzer die Möglichkeit, die erhaltenen Resultate manuell zu bewerten. Diese Feedbacks werden regelmässig gespeichert und können bei Änderung der Parameter Hinweise darauf geben, wie diese sich auf die Suchergebnisse ausgewirkt haben. [Brin et Page 98]

5.7.2 Graphische Aspekte

Eine Stärke Googles bezüglich der Benutzerschnittstelle liegt in der kargen graphischen Gestaltung, wobei trotzdem die für eine Internetrecherche nötigen Informationen und Werkzeuge bereitgestellt werden. Besonders in Zeiten als die Verbindung ins Internet noch hauptsächlich über die Telephonleitung und mit analogen Modems zustande kam, war dies ein gewichtiger Vorteil der Suchmaschine von Google, da hier verhältnismässig wenig Nerven und Geld durch langsame Ladezeiten, bedingt durch unnötig viele Graphiken, verloren gingen.

Neben der ansprechenden Gestaltung der Suchresultate (Beispiel: Seiten der gleichen Homepage werden als solches gekennzeichnet) wurde auch ein "I'm Feeling Lucky"-Button in der Google-Seite implementiert, welcher einen direkt zum höchstbewerteten Resultat führt.

Wie die Abbildungen am Anfang gezeigt haben, hat sich AltaVista von der optischen Gestaltung her Google angepasst (Abbildungen 3 & 5), womit sich auch hier die Spezialität von Google durchgesetzt hat.

6 Zusammenfassung

In den Bereichen Popularität und Suchanfragen pro Tag ist AltaVista von Google ganz klar überflügelt worden, hingegen die Grösse des Index ist praktisch gleich. Zum Erfolg von Google geführt haben also gute Suchresultate, da die einzelnen Benutzer ihren Nutzen maximieren und so die Suchmaschine mit den subjektiv besten Resultaten wählen, die Aggregation der Nachfrage ergibt die meistgenutzte, und somit am meisten subjektiv hoch eingeschätzte Suchmaschine.

Den Ausschlag für den Erfolg von Google gegeben hat in erster Linie die konsequente Ausrichtung des Projektes auf "Web-Tauglichkeit", Qualität der Suchresultate und Benutzerfreundlichkeit. So ist das System dank Parallelität der Prozesse schnell, und sehr gut dem Wachstum des WorldWideWeb anpassbar, dank Aufteilung des Speichers auf viele kleine Einheiten, konsequenter Arbeit mit Linux und Anwendung eines gut skalierbaren Ranking-Algorithmus. Weiter nutzt Google mit dem patentierten Ranking-Algorithmus PageRank, sowie dem Anchortext-Verfahren, in optimaler Weise die Hypertext-Struktur des Web aus, was zu qualitativ sehr guten Resultaten führt. Überdies gibt es mit den archivierten Seiten, sowie den vielen, neben HTML unterstützten Formaten, einige äusserst nützliche und unter den Suchmaschinen exklusive Tools, welche, wie auch die zweckmässig gestaltete Webseite, die Benutzer ansprechen.

Noch weisen müssen wird sich, wie lange sich Google als Branchenführer wird halten können, da unterdessen viele Suchmaschinen die Errungenschaften von Google nachahmen. Dies als typische Reaktion eines Marktes, auf eine bahnbrechende (patentierete) technologische Erfindung, welche nach einer gewissen Anpassungszeit Anbieter enger Substitute auf den Plan ruft (Produktdifferenzierung [Varian 01]).

Ein Anhalten der Vormachtstellung Googles würde also die kontinuierliche Weiterentwicklung der eigenen Suchtechnologie bedingen.

7 Literaturverzeichnis

- [switch.ch 03] Domain-Namen unter .ch
<http://www.switch.ch/de/id/stat/stat-ch.html>
Zuletzt abgerufen: 25. November 2003
- [dmoz.org 03] Open Directory Project
<http://www.dmoz.org/>
Zuletzt abgerufen: 25. November 2003
- [Bekavac 02] Prof. Dr. Bernard Bekavac, Methoden und Verfahren von Suchdiensten im WWW/Internet, 2002
http://www.inf-wiss.uni-konstanz.de/suche/tutorial/such_tutorial_advanced.html
Version vom 30. Oktober 2002
Zuletzt abgerufen: 25. November 2003
- [Hawkins 03] Clive Hawkins, A brief history of the AltaVista search engine.
<http://www.websearchworkshop.co.uk/altavista-history.htm>
Zuletzt abgerufen: 25. November 2003
- [Karzaunikat-1 03] Stefan Karzaunikat, Die AltaVista Geschichte, 2003
<http://www.suchfibel.de/8geschichten/>
Version vom 12. September 2003
Zuletzt abgerufen: 25. November 2003
- [Brin et Page 98] Sergey Brin and Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, 1998
<http://www-db.stanford.edu/~backrub/google.html>
Zuletzt abgerufen: 25. November 2003
- [google.com 03] Google History
<http://www.google.com/corporate/history.html>
Zuletzt abgerufen: 25. November 2003
- [Schöch 01] Volker C. Schöch, Die Suchmaschine Google, Arbeit zum Seminar „Algorithmen für das WWW“, 19. Juni 2001
<http://www.tau-web.de/home/interests/uni/google.pdf>
Zuletzt abgerufen: 26. November 2003
- [Lindner 03] Google läßt weiter auf den Börsengang warten, Artikel von Roland Lindner, Frankfurter Allgemeine Zeitung, 29. September 2003, Nr. 226, Seite 18
- [Sullivan-1 03] Danny Sullivan, Search Engine Sizes, 2. September 2003
<http://www.searchenginewatch.com/reports/article.php/2156481>
Zuletzt abgerufen: 25. November 2003
- [Sullivan-2 03] Danny Sullivan, comScore Media Metrix Search Engine Ratings, 28. Oktober 2003
<http://searchenginewatch.com/reports/article.php/2156431>
Zuletzt abgerufen: 25. November 2003
- [Sullivan-3 03] Danny Sullivan, Searches Per Day, 25. Februar 2003
<http://www.searchenginewatch.com/reports/article.php/2156461>
Zuletzt abgerufen: 25. November 2003
- [Sullivan-4 03] Danny Sullivan, Nielsen NetRatings Search Engine Ratings, 25. Februar 2003
<http://searchenginewatch.com/reports/article.php/2156451>
Zuletzt abgerufen: 25. November 2003
- [Koster 03] Martijn Koster, The Web Robots FAQ
<http://www.robotstxt.org/wc/faq.html>
Zuletzt abgerufen: 25. November 2003
- [Karzaunikat-2 03] Stefan Karzaunikat, Informationen sammeln, 2003
<http://www.suchfibel.de/5technik/sammeln.htm>
Version vom 12. September 2003
Zuletzt abgerufen: 25. November 2003

- [Eipert 00] Eduard Eipert, Suchmaschinen und ihr technischer Aufbau, 13. Juli 2000
http://www-ra.informatik.uni-tuebingen.de/lehre/ss00/pro_internet_ausarbeitung/proseminar_eipert_ss2000.pdf
Zuletzt abgerufen: 25. November 2003
- [Wichmann 99] André Wichmann, Aufbau und Techniken von Suchmaschinen für das WWW, 1. Juni 1999
<http://www-student.informatik.uni-bonn.de/~wichmann/writings/webcrawler/>
Zuletzt abgerufen: 25. November 2003
- [Schulzki 00] Bert Schulzki, Textbasiertes Ranking, 19. Mai 2000
<http://www.informatik.hu-berlin.de/~schulzki/sm/>
Zuletzt abgerufen: 25. November 2003
- [Barker 03] Joe Barker, Search Engines. Finding Information on the Internet: A Tutorial, 2003
<http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/SearchEngines.html>
Version vom 20. Juni 2003
Zuletzt abgerufen: 25. November 2003
- [Müller 99] Lukas Müller, Bearbeitung von [Bekavac 02], August 1999
<http://www.lupi.ch/Internet/Suche/suche.htm>
Zuletzt abgerufen: 26. November 2003
- [Page et al 98] Lawrence Page, Sergey Brin, Rajeev Motwani und Terry Winograd, The PageRank Citation Ranking: Bringing Order to the Web, 29. Januar 1998
<http://dbpubs.stanford.edu:8090/pub/1999-66>
Zuletzt abgerufen: 26. November 2003
- [Varian 01] Hal R. Varian, Grundzüge der Mikroökonomik, 5. Auflage, München/Wien, 2001, Seite 435

8 Quellenangabe der Abbildungen

- Abbildung 1-5 Screenshot, Internet Archive Wayback Machine, <http://www.archive.org>, zuletzt abgerufen am 26.11.2003
- Abbildung 6–10 Search Engine Watch
<http://www.searchenginewatch.com>, zuletzt abgerufen am 26.11.2003
- Abbildung 11–13,15 Methoden und Verfahren von Suchdiensten im WWW/Internet, Bernard Bekavac
http://www.inf-wiss.uni-konstanz.de/suche/tutorial/such_tutorial_advanced.html, zuletzt abgerufen am 26.11.2003
- Abbildung 14 Aufbau und Techniken von Suchmaschinen für das WWW, André Wichmann
<http://www-student.informatik.uni-bonn.de/~wichmann/writings/webcrawlers>, zuletzt abgerufen am 26.11.2003
- Abbildung 16,18,20 Die Suchmaschine Google, Volker C. Schöch
<http://www.tau-web.de/home/interests/uni/google.pdf>, zuletzt abgerufen am 26.11.2003

9 Anhang A

Metatag (in HTML Notation)	Bedeutung / Verwendung
<pre><meta name="keywords" content="Stichworte"></pre> bzw. <pre><meta http-equiv="keywords" content="Stichworte"></pre>	Stichworte, die den Inhalt des Dokuments möglichst eindeutig und unterscheidbar charakterisieren.
<pre><meta name="description" content="eine kurze Inhaltszusammenfassung"></pre>	Eine kurze und prägnante Inhaltszusammenfassung, die auch für Menschen gut lesbar ist, da der Inhalt dieses Tags von einigen Suchdiensten beim Suchergebnis mit angezeigt wird. Wichtig vor allem bei Verwendung von Frames, Javascript und überwiegend nicht-indexierbarem medialen Anteil im Dokument.
<pre><meta name="abstract" content="Stichworte"></pre>	dito.
<pre><meta name="author" content="Name"></pre>	Soll den Autoren des Dokuments benennen. Nützlich ist die Ergänzung weiterer Angaben wie Organisation und Ort.
<pre><meta name="copyright" content="Name"></pre>	Kennzeichnung des Inhabers der Urheberrechte am Dokument bzw. dessen Inhalt.
<pre><meta name="date" content="jjjj-mm-ttThh:mm:ss+hh:mm"></pre>	Angabe von Datum und Uhrzeit der Erstellung oder Veröffentlichung des Dokuments. Dieses muss einer speziellen Syntax folgen, wie nebenstehend angedeutet: Das "T" (für Time) ist ein feststehendes Schlüsselwort zur Trennung von Datum und Uhrzeit, die Stunden und Minuten-Angabe nach dem "+" betrifft die Zeitabweichung gegenüber der Greenwich-Zeit.
<pre><meta name="generator" content="Software-Werkzeug"></pre>	Hier wird bei Erzeugung von HTML-Code durch Generatoren der Name des Software-Werkzeuges automatisch eingetragen. Prinzipiell wären aber auch andersartige, manuell vorgenommene Eintragungen gültig.
<pre><meta name="publisher" content="Name"></pre>	Eintrag der veröffentlichenden Person, Organisation.
<pre><meta http-equiv="Reply-to" Content="mailto:Email@Adresse.de"></pre>	Angabe der Email-Adresse für Mitteilung von Problemen, Fehlern usw.
<pre><meta name="robots" content="Attributwert"></pre>	Die auch als Robot-Exclusion-Tag bekannte Angabe kann das Verhalten der Suchmaschinen im Umgang mit dem HTML-Dokument bestimmen. Als Attributwert kommen folgende Möglichkeiten in Betracht: noindex - Dokument soll nicht indexiert werden. index - Dokument soll indexiert werden. nofollow - Es sollen keine abgehenden Links verfolgt werden, die Indexierung des aktuellen Dokuments ist allerdings erlaubt. follow - Das Dokument soll indexiert werden und abgehenden Links kann durch Crawling nachgegangen werden. all - Entspricht index und follow.
<pre><meta name="revisit-after" content="Anzahl Tage"></pre>	Dieses Metatag soll den Crawler einer Suchmaschine veranlassen, in der angegebenen Anzahl Tagen diese Seite erneut aufzusuchen.
<pre><meta name="page-topic" content="Stichworte"></pre>	Hier können Angaben zum Themenbereich, auf den sich das Dokument bezieht, gemacht werden.
<pre><meta name="page-type" content="Stichworte"></pre>	Durch dieses Tag kann die Ressourcenart des Dokuments bzw. dessen Darstellungsform angegeben werden, z.B. Grafik, Linkliste, Eingabemaske.

Quelle: http://www.inf-wiss.uni-konstanz.de/suche/tutorial/such_tutorial_advanced.html zuletzt abgerufen am 30.10.03

10 Anhang B

AND	Findet Dokumente mit allen angegebenen Wörtern oder Phrasen. Erdnuss AND Öl findet Dokumente, die sowohl das Wort Erdnuss als auch Öl enthalten.
OR	Findet Dokumente, die mindestens eines der gesuchten Wörter oder Phrasen enthalten. Erdnuss OR Öl findet Dokumente, die entweder das Wort Erdnuss oder Öl enthalten. Die gefundenen Dokumente können auch beide Begriffe enthalten, müssen aber nicht.
AND NOT	Schließt Dokumente aus, die das angegebene Wort oder die Phrase enthalten. Erdnuss AND NOT Öl findet alle Dokumente, die das Wort Erdnuss enthalten, nicht aber den Begriff Öl. NOT muss immer zusammen mit einem anderen Operator, wie etwa AND, verwendet werden. AltaVista kann die Angabe 'Erdnuss NOT Öl' nicht verarbeiten. Geben Sie also Erdnuss AND NOT Öl ein..
NEAR	Sucht Dokumente, die beide angegebenen Begriffe oder Phrasen im Umfeld von 10 Wörtern aufweisen. Erdnuss NEAR Öl würde alle Dokumente finden, in denen kurz nacheinander von Erdnussöl, nicht aber von anderen Ölsorten die Rede ist.
*	Das Sternchen ist ein Platzhalter, der von jedem beliebigen Buchstaben besetzt werden kann. Bass* würde Dokumente mit Bass, Besses und Basset finden. Beachten Sie, dass sie mindestens drei Buchstaben vor * setzen müssen. Sie können das * auch in die Mitte des Wortes setzen. Das ist nützlich, wenn Sie nicht wissen, wie man ein Wort ganz genau schreibt. Spag*etti würde Dokumente finden, die die Worte Spaghetti und Spagetti enthalten.
()	Verwenden Sie Klammern, um komplexe Boolesche Ausdrücke zusammenzufassen. Die Eingabe (Erdnuss AND Öl) AND (Gelee OR Marmelade) findet Dokumente, die die Wörter 'Erdnussöl und Gelee' oder 'Erdnussöl und Marmelade' enthalten - oder beide Kombinationen.
anchor:text	Findet Seiten, die das angegebene Wort oder die Phrase in einem Hyperlink enthalten. anchor:Job +programmieren würde Seiten mit dem Wort Job in einem Link und mit dem Wort programmieren im Inhalt einer Seite finden. Setzen Sie keine Leerzeichen vor oder nach dem Doppelpunkt. Sie müssen das Stichwort wiederholen, um mehrere Wörter oder Phrasen zu finden; anchor:Job OR anchor:Karriere würde zum Beispiel Seiten mit Anker finden, die entweder das Wort Job oder das Wort Karriere enthalten.
applet:class	Findet Seiten, die bestimmte Java-Applets enthalten. Applet:Morph findet Seiten, die Applets mit der Bezeichnung Morph enthalten.
object:class	Findet Seiten, die bestimmte Objekte enthalten, die durch ein anderes Programm erstellt wurden (z.B. ein Flash-Objekt). Object:Money findet Seiten, die Objekte mit der Bezeichnung Geld enthalten.
domain:domainname	Findet Seiten, die sich innerhalb einer bestimmten Domäne befinden. Während die Eingabe domain:uk Seiten aus Großbritannien findet, ergibt domain:com eine Suche nach kommerziellen Sites.
host:hostname	Findet Seiten eines bestimmten Computers. Die Sucheingabe host:www.shopping.com würde Seiten auf dem Shopping.com -Computer finden, während host:dilbert.unitedmedia.com Seiten auf dem Computer mit dem Namen Dilbert und der Host-Adresse unitedmedia.com finden würde.
image:filename	Findet Seiten mit einem bestimmten Bilddatei-Namen. Die Eingabe image:Strände findet Seiten, die Bilddateien mit der Bezeichnung Strände enthalten.

Weshalb Google AltaVista überflügelte.

like:URLtext	Findet Seiten, die einer bestimmten URL ähnlich oder verwandt sind. Die Eingabe like:www.abebooks.com findet Websites, über die gebrauchte und seltene Bücher verkauft werden, so ähnlich wie auf der the www.abebooks-Site. like:sfpl.lib.ca.us/ findet öffentliche und universitäre Bibliotheken. Die Eingabe like:http://www.indiaxs.com/ dagegen findet Sites über die Kultur des Indischen Subkontinents.
link:URLtext	Findet Seiten, die über eine spezielle URL mit einem Link zu einer Seite führen. Die Eingabe link:www.myway.com findet alle Seiten, die mit myway.com verknüpft sind.
text:text	Findet Seiten, die jeglichen Text innerhalb einer Seite außer in Bild-Tags, Links, oder URLs enthalten. Die Suche nach text:Schulabschluss würde alle Seiten finden, die den Begriff Schulabschluss enthalten.
title:text	Findet Seiten, die bestimmte Wörter oder Phrasen im Titel der Seite findet (erscheint in der Titelleiste der meisten Browser). Die Suche nach title:Sonnenuntergang würde demnach Seiten finden, die den Begriff Sonnenuntergang in ihrem Titel enthalten.
url:text	Findet Seiten mit einem spezifischen Wort oder einer Phrase in der URL. Die Eingabe url:Garten findet alle Seiten auf allen Servern, die den Begriff Garten irgendwo in Host-Name, Pfad oder Dateiname enthalten.

Quelle: http://www.altavista.com/help/adv_search/syntax, zuletzt abgerufen am 26. November 2003