Fachhochschule Köln
Cologne University of Applied Sciences
Fakultät für Informations- und Kommunikationswissenschaften
Studiengang Informationswirtschaft

Diplomarbeit
(Version 1.o)

Untersuchung zum Wachstum eines verteilten Index einer Peer-to-Peer-Web-Suchmaschine

Vorgelegt von: Britta Jerichow
Fachhochschule Köln

Datum: 24. Januar 2007

Abstract

Diplomarbeit:

Untersuchung zum Wachstum eines verteilten Index einer Peer-to-Peer-Web-Suchmaschine

Britta Jerichow
Fachhochschule Köln
Studiengang Informationswirtschaft
14. Mai 2007

Das Internet ist heutzutage ein vielgenutztes Informationsmedium. Als Hilfsmittel zum Auffinden von Informationen dienen für 90 Prozent aller Internetuser Suchmaschinen, welche mit komplexen Datensammlungs- bzw. Crawling-, Abspeicherungs- und Suchalgorithmen arbeiten. Die etablierten Suchmaschinen basieren zumeist auf dem Client-Server-Prinzip. In dieser Diplomarbeit wird eine Suchmaschine, die auf einem Peer-to-Peer-Prinzip konzipierte wurde, mit ihrer divergierenden Art der Datensammlung und Datenabspeicherung vorgestellt. Im Fokus steht die Untersuchung zum Wachstum eines verteilten Index einer Peer-to-Peer-Web-Suchmaschine in Abhängigkeit von Userzahl und Zeit. Hier wird überprüft, ob sich das neuartige, userzentrierte Crawlingverfahren zur Generierung eines Index eignet. Die Neuerung bei diesem Verfahren besteht darin, dass der User die Funktion des Crawlers durch einfaches Aufrufen der Webseiten im Browser übernimmt.
Zur Überprüfung der Eignung werden drei Hypothesen formuliert, die anhand der durch einen Versuch erhobenen Daten sowie der theoretisch erarbeiteten Grundlagen geprüft und diskutiert werden. Die Erkenntnisse aus Theorie und Empirie geben Aufschluss über den Zeitraum, die Useranzahl und die zu erreichende Indexgröße. Sie liefern keine konkreten Zahlen, lassen jedoch Rückschlüsse für die Auswahl von Zielgruppe und Marketingmaßnahmen zum Aufbau einer Suchmaschinen-Community zu.

Schlagworte: Internet, Suchmaschine, Peer-to-Peer, P2P, Index, Indexgröße, Crawlingverfahren, userzentriert, Suchverhalten

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AOL	-	America Online
CPU	-	Central Processing Unit
DHT	-	Distributed Hash Table
HTML	-	Hypertext Markup Language
HTTP	-	Hypertext Transfer Protocol
ICQ	-	Akronym für: I seek you
IM	-	Instant Messaging
IP	-	Internet Protocol
ISP	-	Internet Service Provider
MSN	-	Microsoft Network
P2P	-	Peer-to-Peer
PDF	-	Portable Document Format
URL	-	Uniform Resource Locator
VoIP	-	Voice over Internet Protocol
WWW	-	World Wide Web

1 Einleitung

Diese Arbeit untersucht das Wachstum eines verteilten Index einer Peer-to-Peer-Suchmaschine. In der Einleitung werden zunächst die Problemstellung sowie der Kontext und die wissenschaftliche Relevanz der Arbeit beschrieben. Darauf folgt die Vorstellung der Ziele dieser Arbeit. Im Anschluss daran wird in Kapitel 4 die Abgrenzung des Untersuchungsgebietes vorgenommen. Das letzte Kapitel der Einleitung verschafft einen kurzen Überblick über den Aufbau und über den weiteren Verlauf der Arbeit.

2 Problemstellung

Die Problemstellung ergibt sich hauptsächlich aus den aktuellen **Problemfeldern** (Kritik- und Angriffspunkte) in der Suchmaschinenbranche. Das für diese Arbeit relevante und in der Öffentlichkeit am häufigsten diskutierte und kritisierte Thema ist die zentrale Architektur der herkömmlichen Suchmaschinen. In dieser Architektur werden die Daten in einem zentralen Index abgespeichert, woraus **komplexe Probleme** resultieren können, die im Folgenden anhand von Beispielen dargestellt werden (vgl. Thiele/Speck 2004 und Wolling 2005: 529).

Das erste Problem der herkömmlichen Suchmaschinen ist die **Entscheidungs- und Selektionsmacht** der Suchmaschinenbetreiber über das Erscheinen von Webangeboten in Suchergebnissen sowie deren Rangreihenfolge (Ranking). Dieses Problem lässt sich anhand eines aktuellen Vorfalls anschaulich verdeutlichen: Die Unternehmen Google und MSN-Suche mussten nach einer erfolgreichen Klage diverser belgischer Zeitungen deren Webseiten vollständig aus dem Google-News-Index sowie aus dem Suchmaschinenindex von der MSN-Suche entfernen. Das Autorenrechte vertretende Unternehmen Copiepresse hatte von beiden Suchmaschinenbetreibern verlangt, alle Artikel der Zeitungen, die Copiepresse angehören, aus dem Google-News-Index zu entfernen, da keine Vereinbarung zwischen Google und den Zeitungen zur Verwendung der Artikel geschlossen wurde (vgl. Cloer 2006).

Die zweite Problematik, die sich aus der zentralen Struktur von herkömmlichen Suchmaschinen ergibt, ist die **Möglichkeit der Zensur** durch den Suchmaschinenbetreiber. So werden in China bspw. von den Suchmaschinen Yahoo, MSN und Google die von der chinesischen Regierung geforderten Zensuren eingehalten (vgl. Rohwedder 2005). In Deutschland werden Seiten mit rechtsradikalen Inhalten von Google zensiert (vgl. Schmitz 2006). Auch wenn in diesen Fällen nationale Gesetze

eine Zensur[1] für eine Marktpräsenz erfordern, wird die starke Macht der Suchmaschinenbetreiber in diesen Handlungen verdeutlicht.

Beim **Spamming**[2], einem weiteren Problem, versuchen Webseitenbetreiter durch spezielle Programmierungen, Webseiten auf die ersten Rangplätze in den Suchergebnislisten zu positionieren und manipulieren auf diese Weise die Rangreihenfolge der Ergebnisse (vgl. Wolling 2005: 529ff.). Dies erscheint oft notwendig, da auf eine Suchanfrage meist mehrere hundert oder sogar tausend Ergebnisse geliefert werden, jedoch 90 Prozent der User in der Regel nur die ersten zehn Ergebnisse sichten (vgl. Machill/Welp 2003: 255ff.). So wurden die BMW Webseiten im Januar 2006 aufgrund von Spamming zu Teilen aus dem Google-Index gelöscht. Erst als die unsichtbare Doorway-Page[3] aus dem Quellcode der Webseite entfernt war, wurde die Wiederaufnahme genehmigt (vgl. Ihlenfeld 2006). Auch wenn in diesem Fall die Handlungsweise des Suchmaschinenbetreibers vertretbar ist, wird hier dessen Macht über die auf seinen Servern gespeicherten Informationen und die hohe Abhängigkeit der Webseitenbetreiber von den Suchmaschinenbetreibern deutlich.

Verdeutlicht wird die Wichtigkeit von Suchmaschinen-Alternativen an der **monopolartigen Marktposition** von Google. Dieses Unternehmen hat weltweit mit 80 Prozent Marktanteil eine marktbeherrschende Stellung eingenommen. Betrachtet man das oben dargestellte Problem der Zensur und den hohen Marktanteil von Google, ließe sich überspitzt behaupten, Google habe Einfluss auf die Informationen, die 80 Prozent der Internetuser über seine Services beziehen. Durch den Aufbau neuer Internetdienste wie Gmail, Google Desktop Search, -Toolbar, -Talk, -Earth, -Maps, -Books, -Scholar sowie diverser anderer Dienste wird das Unternehmen oft als „Wissensmonopol“ bezeichnet. Diese Aussage wird unter anderem dadurch bekräftigt, dass alle Daten an einem Ort (zentral) zusammenlaufen und die Erstellung komplexer und detaillierter Userprofile ermöglichen. Je mehr Google Services ein User nutzt, desto umfangreicher wird sein daraus ableitbares Profil (vgl. Thiele/Speck 2004: 8).

Die öffentliche Downloadfreigabe des AOL-Datenfiles reiht sich in die folge kritischer Ereignisse ein. Das Datenfile enthielt Suchanfragen von rund 500.000 Usern über

---

1 Die Open Net Initiative untersucht weltweit das Ausmaß und die Auswirkungen der Internetzensur. Auf deren Homepage findet sich eine Karte, die den Zensurgrad in sämtlichen Staaten der Welt zeigt. Online unter: http://www.opennet.net/map/index2.html (Abruf: 28.09.2006).

2 Als bekannteste Techniken lassen sich Keyword Stuffing, Linkspamming, Doorway-Pages, Hidden Links und Invisible Text nennen.

3 Doorway Pages sind hoch optimierte Seiten, deren alleiniger Zweck es ist, bei den Suchmaschinen angemeldet und dort gut platziert zu werden.

einen Zeitraum von 3 Monaten. Da AOL bei Suchanfragen auf den Google-Index zurückgreift, gerieten somit Google-Daten an die Öffentlichkeit. Diese zusammengefassten Daten ließen aufgrund nur leichter Anonymisierung, Rückschlüsse auf einzelne User zu (vgl. Wilkens 2006). Anhand des geschilderten Falles stellt sich die **Frage nach Datenschutz**, der ein enorm sensibles Thema in der Diskussion über Suchmaschinen darstellt.[4]

Die immer kürzer werdende Halbwertszeit von Wissen und die damit verflochtene Informationsflut stellen die nächste große Herausforderung an die Suchmaschinenbetreiber. So hat sich die Informationsmenge zwischen dem 18. und 19. Jahrhundert verdoppelt und zwischen dem 19. und 20. Jahrhundert sogar verzehnfacht (vgl. Speck 2004[5]). Speck vermutet sogar, dass sich die Informationsmenge ab 2050 jeden Tag verdoppeln wird (vgl. ebd). Suchmaschinen dienen dazu diese wachsende Informationsmenge zu erschließen. Hinzu kommt, dass parallel zur Informationsflut auch der Bedarf an Speicherplatz für den Suchmaschinenindex steigt, womit wiederum die personellen und finanziellen Ressourcen steigen. So werden bereits heute 450.000 Server (vgl. Mehta 2006 und Gilder 2006) benötigt, um den Suchmaschinenservice von Google instand halten zu können.

Das **Wachstum der Informationen und Daten im Internet** und die Anzahl der für Suchmaschinen zu erfassenden Webseiten stellt ein weiteres Problem dar (vgl. Wolling 2005). Die Anzahl der Webseiten wächst schneller, als die Suchmaschinen in der Lage sind, diese zu erfassen. Aus der großen Informationsmenge ergeben sich zwei weitere Probleme: Zum einen die **Vollständigkeit** der täglich hinzukommenden Webseiten und zum anderen die **Aktualität**[6] der im Index gespeicherten Daten. Hier können Defizite zur Unzufriedenheit bei den Usern führen.

Des Weiteren existieren Webangebote, die aufgrund der Arbeitsweise von Suchmaschinen nicht in den Index dieser Suchmaschinen aufgenommen werden können. Ist eine Webseite nicht mit einer anderen verlinkt, und wurde sie nicht von ihrem Web-

---

4 Auf der 28. Internationalen Konferenz der Datenschutzbeauftragten (3. November 2006) verabschiedeten die versammelten Datenschutzbeauftragten eine Entschließung zum "Datenschutz bei Suchmaschinen", weitere Informationen unter: http://www.datenschutz-berlin.de/doc /int/konf/28/Entschliessung%20zum%20Datenschutz%20bei%20Suchmaschinen.pdf (Abruf: 11.11.2006).

5 weitere Untersuchungen zur Informationsmenge von Varian und Lyman unter: http://www.sims. berkeley.edu/how-much-info-2003 (Abruf: 15.12.2006).

6 Zu dieser Thematik wurde aktuell eine Studie von Bar-Yossef und Gurevich 2006 anhand der drei global Player der Suchmaschinenbranche durchgeführt. Patzwaldt hat die Ergebnisse übersichtlich in seinem Blog zusammengefasst. Online unter: http://www.at-web.de/blog/20060930/google-hat-die-meisten-leichen-im-keller.htm (Abruf: 02.10.2006) Auch Lewandowski führte eine Aktualitätsstudie der Suchmaschinen durch. Vgl. Lewandowski 2006a.

seiten-Hersteller bei der Suchmaschine angemeldet, bleibt sie unentdeckt im sogenannten **Deep-** oder **Invisible Web** liegen.

3 Kontext und wissenschaftliche Relevanz der Arbeit

Den Kontext dieser Arbeit bildet die Entwicklung einer alternativen Suchmaschine auf Basis der Peer-to-Peer-Technologie. Ein Beweggrund für die Entwicklung war die zunehmende Kritik an den etablierten Suchmaschinen. Das Moor'sche Gesetz[7] und die darin formulierte, schnelle technische Weiterentwicklung, die besonders signifikant in der Computer- und Telekommunikationsbranche zu beobachten ist, bildet die Basis für immer leistungsfähigere Computer-Ressourcen und eine verbesserte Grundlage für den Einsatz von Peer-to-Peer-Technologien.

Die Idee einer Peer-to-Peer-Suchmaschine ist nicht neu. Ein Open-Source-Projekt namens Yacy - **Y**et **A**nother **Cy**berspace - wurde in der Fachpresse als zukunftsorientiert beschrieben und existiert bereits auf dem Markt (vgl. Sietmann 2005: 52 und Schwärze 2004: 23). Die Antwortzeiten liegen hier jedoch durchschnittlich bei sechs Sekunden und sind somit noch weit entfernt von denen der Marktführer, welche Zeiten von unter einer Sekunde realisieren.[8] Aus einem Eintrag in dem Blog www.netzpolitik.org am 07. Sep. 2006[9] geht hervor, dass von den Usern ein ähnliches System mit besserer Performance „sehnsüchtig" gewünscht wird. Sander-Breuerman, Vorsitzender des Suma e.V.[10], fordert für die Zukunft der Suchmaschinen „etwas, das die Pluralität der Gesellschaft widerspiegelt" (vgl. Panzer 2006). Auf der dritten Tagung des Suma e.V. äußerte er außerdem die Forderung nach dezentralen Suchmaschinenstrukturen, die sich nicht dominieren lassen. Er zieht eine Analogie zur „Wikipediasierung" der Suchmaschinen. Zudem strebt Europa eine von den USA unabhängige Suchmaschine an. So entstand das Projekt Quaero. Dieses Projekt verdeutlicht, dass bereits auf Bundesebene Maßnahmen getroffen wurden, um dem Wissensmonopol von Suchmaschinen wie Google entgegen zu wirken.

Um eine weitere Alternative zu den herkömmlichen Suchmaschinen bieten zu können, wurde die **Peer-to-Peer-Suchmaschine „Faroo"** entwickelt. Durch die

---

7 Vgl. http://de.wikipedia.org/wiki/Mooresches_Gesetz (Abruf: 12.12.2006).

8 Vgl. Patzwaldt 2006. Patzwaldt, Klaus: einer der bekanntesten deutschen Suchmaschinenexperten und Betreiber des bekannten Webportals @-web.de in dem er zu aktuellen Themen der Suchmaschinenbranche berichtet.

9 Vgl. 2. Kommentar unter: http://netzpolitik.org/2006/blockt-google-privacy-enhancing-tools/ (Abruf: 07.10.2006).

10 Gemeinnütziger Verein zur Förderung der Suchmaschinen-Technologie und des freien Wissenszugangs, weitere Informationen: http://suma-ev.de/index.html (Abruf: 15.12.2006).

Umsetzung der Peer-to-Peer-Technologie erhalten Index und Crawler von Faroo einen dezentralen Charakter. Somit lässt sich die Möglichkeit der zentralen Selektionsmacht, der zentralen Zensur und dem Spamming sowie Datenschutz in der oben erläuterten Art und Weise, beheben.

Das Wachstum des Index einer neuen Suchmaschine ist essentiell. Um die Suchanfragen der User beantworten zu können, wird ein Index mit einer umfangreichen Menge an indexierten Webseiten und Worten benötigt. Bevor Faroo den Suchmaschinenmarkt betritt, werden im Voraus Untersuchungen zum Index-Wachstum angestellt, um mögliche Herausforderungen bereits vor Markteintritt möglichst genau einschätzen zu können.

Der Zugang zu Informationen ist ein sensibles Thema in der Informationsgesellschaft, welches seit der Einführung des Internets immer mehr an Bedeutung gewinnt (vgl. Grietje 2005 und Bernhardt 2003: 319). Die wissenschaftliche Relevanz dieser Arbeit für den Studiengang Informationswirtschaft liegt in der Informationsbeschaffung. Das Internet dient sowohl konventionellen als auch professionellen Internetusern gleichermaßen zumindest als Einstiegsquelle für eine Informationsrecherche[11]. Es besteht also ein großes Interesse, die enorm heterogene Informationsvielfalt auch für Information Professionals nach Kriterien zu erschließen, die einen möglichst hohen, zuverlässigen und vor allem objektiven Informationsgehalt ermöglichen. Eine wissenschaftliche Relevanz für den Forschungsbereich ergibt sich aus dem Kontext dieser Arbeit. Da die dargestellte Technologie der Datensammlung im WWW eine Neuerung darstellt, wurden bisher weder Forschungen durchgeführt noch Ergebnisse publiziert. Bisherige Untersuchungen waren dagegen eher auf die Recherchestrategien der User ausgerichtet (vgl. Machill/Welp 2003).

4 Thema und Ziel der Arbeit

Die Auseinandersetzung mit dem Internet, die vielfältige, umfangreiche und dynamische Informationsmenge, die sich laut verschiedener Marktstudien nur durch Suchmaschinen auffinden lässt sowie die technischen Eigenschaften von Peer-to-Peer-Technologien sind Grundlagen dieser Arbeit. Auf dieser Basis wird die Peer-to-Peer-Suchmaschine „Faroo“ mit ihrer neuartigen Technik zur Datensammlung im Internet, der Indexgenerierung und -speicherung beschrieben.

---

[11] Es ist hinzuzufügen, dass zusätzlich ein bewusster Umgang sowie Medienkompetenz zum Beurteilen und Selektieren der Informationen im Internet erforderlich ist.

Ziel dieser Arbeit ist die Prüfung des neuartigen Datensammlungsverfahrens auf seine Eignung zur Generierung eines Suchmaschinenindex. Dieses neue Datensammlungsverfahren soll anhand eines Versuchs zum Wachstum eines verteilten Index einer Peer-to-Peer-Suchmaschine eruiert werden. Fokussiert wird dabei die Abhängigkeit des Indexwachstums von den Variablen Userzahl und Zeit. Ergänzend werden weitere Aspekte des neuen Datensammlungsverfahrens beleuchtet die den Umfang der im Index gespeicherten Informationen beeinflussen.

Die abschließende kritische Diskussion der Versuchsergebnisse gibt Aufschluss über die Eignung des Indexierungsverfahrens, eine Größenordnung der Anzahl an Usern und den nötigen Zeitraum zum Aufbau eines konkurrenzfähigen Suchmaschinenindex.

5 Abgrenzung des Untersuchungsgebietes

Im Rahmen dieser Arbeit stehen der technische Aufbau (speziell die dezentrale Netztopologie), die dezentrale Umsetzung eines verteilten Index und das dezentrale Crawlingverfahren im Fokus und bilden den Leitfaden für den weiteren Verlauf. Auf das neuartige Rankingverfahren Faroo's wird im Rahmen dieser Arbeit nicht detailliert eingegangen. Das Hauptziel dieser Arbeit beläuft sich primär auf die Datensammlungsmethode und die sich daraus ergebenen Schlussfolgerungen.

6 Aufbau der Arbeit

Die vorliegende Arbeit ist in vier Hauptteile gegliedert. Nach der **Einleitung** beschreibt der zweite Teil die **Theoretischen Grundlagen**, die zum besseren Verständnis der Zusammenhänge zwischen der Technikstruktur, den Netzwerken und deren Effekten sowie der Internetnutzung beitragen.
Im Anschluss an den Theoretischen Teil folgt der **Empirische Teil**. Die empirische Erhebung der Daten zum Wachstum des Index in Abhängigkeit von Zeit und Userzahl wird anhand eines Versuchs durchgeführt. Die Untersuchungsmethode, die Versuchsplanung, der Aufbau und die Auswertung werden hier detailliert dargestellt.
Das **Fazit** gibt als abschließender Hauptteil eine Zusammenfassung der gewonnenen Erkenntnisse und zeigt den Handlungsbedarf auf.

Im Text werden sowohl die männliche als auch die weibliche Schreibform verwendet. Selbstverständlich sollen damit aber, soweit nicht anders erwähnt, beide Geschlechter angesprochen werden. Zum Verständnis dieser Arbeit werden Internet- und informationswissenschaftliche Grundkenntnisse vorausgesetzt.

Ebenfalls wird die grundlegende Funktionsweise einer Suchmaschine - im Sinne dieser Arbeit - als Basiswissen zum Verständnis vorausgesetzt. Das Bewusstsein, dass Suchmaschinen das Internet nicht erst im Moment einer Suchanfrage durchsuchen, sondern auf einen vorher generierten Index zurückgreifen, gilt in diesem Kontext als unabdingbar (vgl. Karzauninkat/Alby 2006: 21).

I Theoretischer Teil und Grundlagen

1 Internet

Der Begriff „Internet“ steht für die englischen Begriffe „Interconnected Networks“ und ist ein elektronisches, dezentral organisiertes Netzwerk aus Computern mit vielen voneinander unabhängigen Netzwerken.[12] Es dient der Kommunikation und dem Austausch von Informationen. Das World Wide Web (WWW) wird oft mit dem Internet gleichgesetzt, tatsächlich ist es jedoch nur einer von mehreren Diensten, die das Internet bietet. Das WWW ist ein Dienst der sich durch hohe Benutzerfreundlichkeit sowie die Einbindung multimedialer Elemente in Webseiten auszeichnet. Ein Beispiel für einen nicht WWW-Dienst ist der E-Mail-Dienst (vgl. Strauch/Kuhlen/Laisiepen 2004a: 64,133). Im weiteren Verlauf dieser Arbeit werden die Begriffe „Internet“, „WWW“ sowie die abgekürzte Form „Web“ synonym verwendet, solange eine andere Bedeutung nicht ausdrücklich dargestellt wird.

1.1 Inhalte und deren Struktur im Internet

Grundsätzlich lässt sich das Web in zwei Bereiche einteilen: das „Oberflächen-Web“ und das „Deep-Web“ (oder „Invisible-Web“). Im Oberflächen-Web liegen digitale Dokumente, Webseiten und Dokumente, die fest verlinkt sind und für die dem User keine Kosten bei deren Betrachtung entstehen (vgl. Stock/Lewandowski 2005: 1). Zu den im Deep-Web liegenden Informationen gehören an das Internet angeschlossene Datenbanken. Diese stellen erst auf eine Suchanfrage des Users Ergebnislisten aus Datenbanken als dynamische Webseite zusammen (vgl. Garbe 2001: 511) und können sowohl kostenfrei genutzt als auch gegen Entgelt erstanden werden. Aus verschiedenen Gründen sind die im Deep-Web liegenden Informationen für Suchmaschinen teilweise nicht zu erschließen (vgl. Sherman/Price 2001: 57). Zu den Gründen gehören technische Hürden, nicht verlinkte Webseiten, „Real-time“-Inhalte wie beispielsweise Nachrichten und dynamisch generierte Webseiten aus den oben erwähnten Datenbanken (vgl. ebd.).

Die im Internet liegenden Inhalte sind eine heterogene Masse, die in verschiedenen Dimensionen beobachtet werden kann. Als eine Dimension beschreibt man die Vielfältigkeit und Komplexität der Medienarten (Audio, Video, Bild, Text) sowie deren zahlreiche Dateiformate (vgl. Ferber 2002: 286ff.). Den größten Anteil an Webin-

---

[12] Vgl. http://de.wikipedia.org/wiki/Internet (Abruf: 12.12.2006).

halten nehmen noch bis heute Webseiten ein, die durch Integration von Links[13] den Zugriff auf andere Dateien ermöglichen (vgl. ebd). Webseiten bestehen aus den zwei Elementen HTML-Code und Text. Diese beiden Elemente bieten zur Zeit für Suchmaschinen den einzigen Ansatz zum erfolgreichen Auffinden von Daten im Web und können folglich als relevantestes Arbeitsmittel für Suchmaschinen bezeichnet werden.
Die aufgezeigten unterschiedlichen Dimensionen sind große Herausforderungen für die Konzeption einer Suchmaschine.

1.2 Größe und Dynamik des Internets

Die Größe des WWW - hierunter ist die Anzahl der vorhandenen und verfügbaren Webseiten einerseits und die Anzahl anderer im Internet verfügbarer Dateiformate andererseits zu verstehen - ist nicht eindeutig bestimmbar. Zwei wissenschaftliche Untersuchungen ergaben folgende Schätzungen: Die Universität Bielefeld schätzte im Januar 2005, dass weltweit 10-15 Milliarden frei zugängliche, statische Webseiten existieren. In dieser Zahl sind Dokumente und Informationen aus Datenbanken, aus geschlossenen und dynamischen Webseiten sowie mit Webseiten verknüpfte Dokumente nicht eingerechnet.[14] Eine Schätzung von Gulli und Signorini[15] im Mai 2005 ergab eine ähnliche Zahl von 11,5 Milliarden existierenden Webseiten. Es ist darauf hinzuweisen, dass eine **Website** mehrere **Webseiten** enthalten kann. So ist z.B. „http://www.ccc.de/“ eine Website, von der es mehrere Unterseiten wie „http://www.ccc.de/calendar/2006/23c3?language=de“ gibt.

Netcraft, ein Internet-Service-Unternehmen in England, analysiert das Internet seit 1995. Im November 2006 gab dieses Unternehmen eine Anzahl von 100 Millionen Webseiten bekannt. Diese Angaben basieren auf angemeldeten Webadressen aller Top-Level-Domains. Um die Anzahl hinzukommender Webseiten zu verdeutlichen, ist hier ein Zuwachs von ca. 3.5 Millionen Sites, allein im Oktober 2006, zu erwähnen. Für das Jahr 2006 wurden bisher insgesamt 27,4 Millionen neue Webseiten verzeichnet. Damit ist der Zuwachs von 17 Millionen aus dem Jahr 2005 schon übertroffen worden. Seit 2004 hat das Internet seine Größe bereits verdoppelt, wie auch Abb. 1 veranschaulicht.

---

[13] Link ist die Abkürzung von Hyperlink und steht z.B. für einen Verweis von einer Webseite auf eine andere. Durch einen Klick auf diesen Link wird automatisch die verwiesene Webseite aufgerufen.
[14] Vgl. http://www.ub.uni-bielefeld.de/biblio/search/ (Abruf: 23.11.2006).
[15] Vgl. http://www.cs.uiowa.edu/~asignori/web-size/ (Abruf: 12.12.2006).

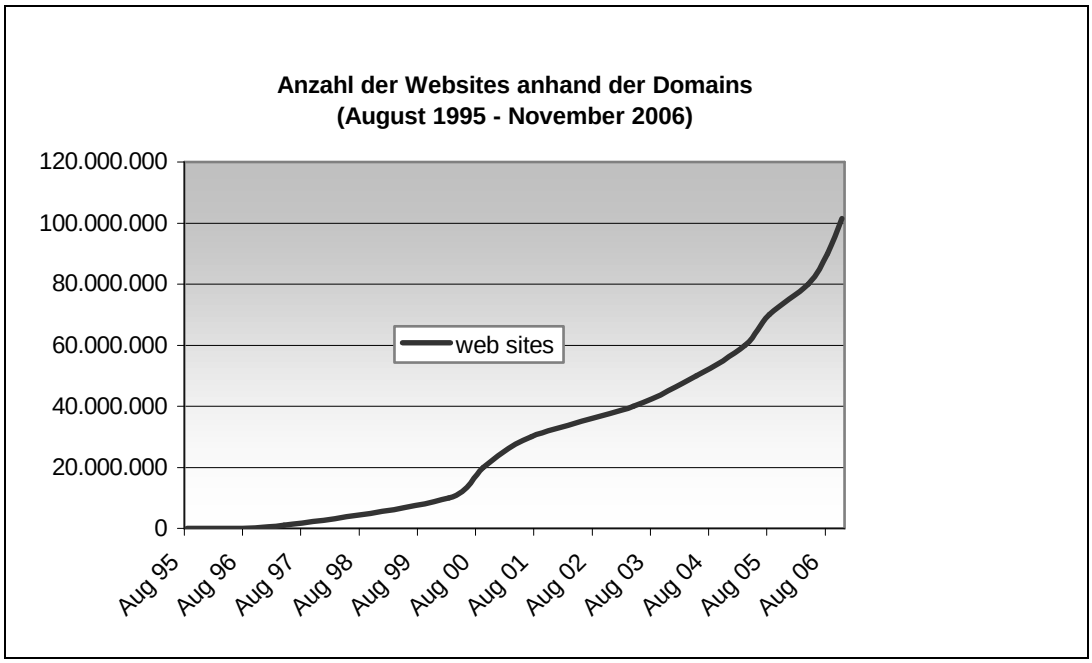


Abb. 1: Anzahl an Webseiten auf Webservern[16]

Allein die Schwankung in den oben dargestellten Ergebnissen verdeutlicht, dass es nahezu unmöglich ist, die Größe des WWW exakt zu eruieren. Über die Anzahl der Dokumente aus dem Deep-Web gibt es nur Vermutungen (vgl. Lewandowski 2005: 9).

Die konstant steigenden Rechnerkapazitäten, Bandbreiten, Teilnehmerzahlen und die sinkenden Zugangskosten zum Internet werden die Datenmenge weiter steigen lassen. Einen nicht unbedeutenden Beitrag zum Wachstum des Internets leistet die momentane Web 2.0-Bewegung[17] mit Anwendungen wie: Flickr[18], Youtube[19],

---

16 Vgl. http://news.netcraft.com/archives/2006/11/01/november_2006_web_server_survey.html (Abruf 02.11.2006).

17 Weiterführende Informationen unter: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (Abruf: 20.11.2006).

18 Flickr ist eine Fotoplattform, die es dem Nutzer erlaubt, Fotos bis zu einer bestimmten Datenmenge auf einem Server zuladen und in Alben zu verwalten. Zusätzlich kann der User den Fotos Schlagworte zuordnen. Weitere Informationen unter: http://www.flickr.com/about/ (Abruf: 20.11.2006).

19 Youtube ist eine kostenlose Videoplattform mit ähnlichen Funktionen wie Flickr. Weiterführende Informationen unter: http://www.youtube.com/ (Abruf: 20.11.2006).

Weblogs[20] und Wikis[21]. So werden dem Internet jeden Tag große Datenmengen neu hinzugefügt. Technorati[22] berichtet zum Beispiel von 75.000 neuen Blogs täglich.[23]

Gleichzeitig wird durch das exponentielle Wachstum des Internets seine Dynamik verdeutlicht. Unter Dynamik ist nicht nur die Entstehung neuer Inhalte zu verstehen, sondern auch das Wegfallen alter Inhalte und die Verschiebung von Dateien und Dokumenten. Untersuchungen von Ntoulas, Cho und Olsten (2004) ergaben, dass 20 Prozent der heute vorhanden Webseiten in einem Jahr nicht mehr existieren und 50 Prozent der heute vorhandenen Webseiten inhaltlich neu beziehungsweise verändert sein werden. Die höchste Veränderungsrate erreicht dabei die Verlinkungsstruktur mit 80 Prozent innerhalb eines Jahres (vgl. Ntoulas/Cho/Olsten 2004: 2).

Diese Untersuchungsergebnisse verdeutlichen, dass Suchmaschinen mit der Aktualisierungsfrequenz der Webseiteninhalte und der weitreichenden Veränderungen der Linkstrukturen einer großen Herausforderung gegenüberstehen.

1.3 Nutzung des Internets

Weltweit nutzen ca. eine Milliarde Menschen - von 6,5 Milliarden der Weltbevölkerung - das Internet.[24] In Tabelle 1 sind die im Internet am stärksten vertretenen zehn Länder nach Anzahl der Internetuser dargestellt, in der Deutschland zurzeit Platz fünf einnimmt.

---

20 Weblog ist ein Kunstwort, zusammengesetzt aus den engl. Worten Web und Log (für Logbuch), die Kurzform ist Blog. Es ist eine Webseite, die periodisch neue Einträge enthält. Neue Einträge stehen an oberster Stelle, ältere folgen in umgekehrt chronologischer Reihenfolge. Für Leser von Blogs besteht die Möglichkeit die Einträge zu kommentieren.

21 Wikis sind im WWW verfügbare Webseitensammlungen, die von den Benutzern nicht nur gelesen, sondern auch online geändert werden können. Weiterführende Informationen unter: http://de.wikipedia.org/wiki/Wiki (Abruf: 01.11.2006).

22 Technorati ist eine Blog-Suchmaschine. Weitere Informationen unter: http://de.wikipedia.org/wiki/Technorati (Abruf: 11.11.2006).

23 Vgl. http://www.heise.de/newsticker/meldung/69392 (Abruf: 11.11.2006) Wie viele davon letztendlich konstant mit neuen und interessanten Einträgen gepflegt werden, bleibt abzuwarten.

24 Vgl. http://www.internetworldstats.com/top20.htm (Abruf: 01.11.2006).

	Land oder Region	Internet Users	Population (2006 Est.)	Internet Penetration	% Users of World
1	United States	207.161.706	299.093.237	69,3 %	19,1 %
2	China	123.000.000	1.306.724.067	9,4 %	11,3 %
3	Japan	86.300.000	128.389.000	67,2 %	7,9 %
4	India	60.000.000	1.112.225.812	5,4 %	5,5 %
5	Germany	50.616.207	82.515.988	61,3 %	4,7 %
6	United Kingdom	37.600.000	60.139.274	62,5 %	3,5 %
7	Korea (South)	33.900.000	50.633.265	67,0 %	3,1 %
8	France	29.521.451	61.004.840	48,4 %	2,7 %
9	Italy	28.870.000	59.115.261	48,8 %	2,7 %
10	Brazil	25.900.000	184.284.898	14,1 %	2,4 %

Tabelle 1: Die Top-zehn-Länder der Internetnutzung[25]

Eine aktuelle Studie von AGOF Internet Facts 2006 belegt, dass 57 Prozent der deutschen Bevölkerung bereits ab 14 Jahren online sind. Die „Silver-Surfer", wie die Internetnutzer über 50 Jahre bezeichnet werden, sind bereits mit 23 Prozent im Internet vertreten und stellen in Deutschland zahlenmäßig die stärkste Altersgruppe (8,36 Mio.) dar. Ihnen folgen die 30-bis 39-Jährigen (8,05 Mio.) mit 22 Prozent, danach die 40- bis 49-Jährigen (7,78 Mio.) mit 21 Prozent und schließlich die 20-bis 29-Jährigen mit 19 Prozent (6,91 Mio.). Diese Zahlen belegen, dass sich die Strukturen der „Onliner"[26] insgesamt immer stärker an die Strukturen der Gesamtbevölkerung anpassen.

Die dargestellten Zahlen verdeutlichen, dass die Bedeutung des Internets in der Gesellschaft wächst. Das Internet ist vergleichbar mit einem Universum, das eine extrem große Menge an Informationen birgt. In dieser Menge an Informationen stehen Antworten auf viele Fragen der User bereit, die es bei Bedarf aufzufinden gilt. Suchmaschinen dienen dabei als „Wegweiser im Netz" (vgl. Machill/Welp 2003). Laut Fittkau und Maaß (2006) sind Suchmaschinen mit 85 Prozent (nach den E-Mail Providern) die im Internet am häufigsten besuchten Webseiten. Zahlen des Statistischen Bundesamtes (2005) verdeutlichen die hohe Bedeutung der Suchmaschinen für den User im Internet. So stiegen laut deren Studie 89 Prozent aller User mit einer Suchmaschine in die Informationen des Internets ein (vgl. Statistisches Bundesamt 2005: 19).

---

25 Vgl. http://www.internetworldstats.com/top20.htm (Abruf: 01.11.2006).

26 Der Begriff „Onliner" bezeichnet Personen, die mit einem Rechner der an das Internet angeschlossen ist, arbeiten (vgl. http://www.daserste.de/service/0206.pdf; Abruf: 20.01.2007).

Suchmaschinen zentralisieren den Zugriff auf Informationen innerhalb des dezentral aufgebauten Informationsnetzes Internet. Damit übernehmen sie eine bedeutende Schlüsselfunktion in der heutigen Informationslandschaft. So berichtet das Forschungsinstitut für Telekommunikation in seinem Newsletter ECIN, Suchmaschinen seien das wichtigste Einstiegsportal bei der Informationsbeschaffung im Internet. Ohne diese würden sich User in der heutigen Informationslandschaft des Internet nicht zurechtfinden (vgl. ECIN 2006). Abb. 2 zeigt die Nutzungshäufigkeit von Internet und Suchmaschinen im Vergleich. Deutlich wird hier, dass die befragten Personen, die das Internet an bis zu drei Tagen in der Woche nutzen, überwiegend Informationssuche durch Zuhilfenahme von Suchmaschinen betreiben. Die Personen, die das Internet hingegen häufiger - an vier bis sieben Tagen - nutzen, sehen dieses nicht nur als reines Informationsmedium.

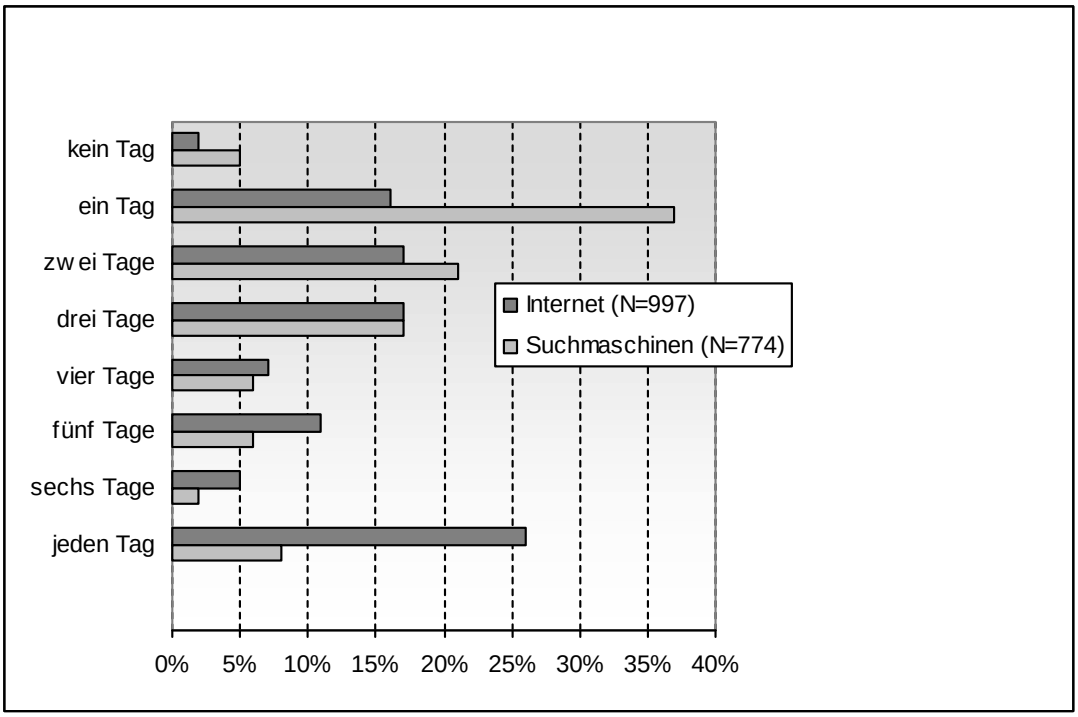


Abb. 2: Internet- und Suchmaschinennutzung im Vergleich[27]

Des Weiteren findet zurzeit ein Umbruch und Paradigmenwechsel in fast allen Bereichen der Internetlandschaft statt. Dieser Trend wird allgemein „Web 2.0“ genannt. Dieser Umbruch verändert die Art und Weise der Internetnutzung maßgeblich: Eine steigende Anzahl an Internetusern partizipiert freiwillig an der Erstellung und Verwaltung von frei zugänglichem „Content“[28], wie bspw. in der von Larry

[27] Vgl. Machill 2003: 144.

[28] „Content“ bedeutet im oben dargestellten Zusammenhang, dass Erstellen von Berichten, anlegen von Nutzerprofilen, abspeichern von Fotos, Videos und Lesezeichensammlungen.

Sanger geschaffenen Online-Enzyklopädie Wikipedia oder dem neuen experten-zentrierteren Citizendium[29]. Projekte wie das ODP - Open Directory Project[30] - und DOAJ - Directoy of Open Access Journals[31] - leben von freiwilligen Editoren, die durch „Peer-Review“[32] und das Beseitigen unseriöser Einträge für einen gewissen Qualitätsstandard sorgen. Die Verantwortung für Inhalte liegt dabei nicht mehr an einer zentralen Stelle, sondern wird den Usern des Internets zurückgegeben. Lessig betitelt dies als Transformation von der „read-only-culture“ zur einer „read-write culture“ (vgl. Lessig 2005). Es verkörpert ein Konzept, in dem das Internet nicht mehr als „one-to-many“ Medium, sondern wieder als „many-to-many“ Medium genutzt wird. SinnerSchrader formulieren dies wie folgt: „Das Internet findet wieder zu sich selbst: vom read-only zum writable Web“[33].

Im Rahmen dieses Paradigmenwechsels etabliert sich ein weiterer Zugang zu Informationen: sogenannte „Social-Bookmarking“ Plattformen wie „delicious“, „furl“, „digg“, „Blinklist“ und „Mr. Wong“ - um die Bekanntesten zu nennen - bieten den Usern eine Möglichkeit, Webadressen als Lesezeichen bei den genannten Dienste-Anbietern abzuspeichern. Zusätzlich können Schlagworte (sogenannte „Tags“) vergeben werden, die eine Suche nach Themen ermöglichen. Diese Vorgehensweise der Recherche ist äußerst effizient, da sie themen- oder interessenspezifisch gebündelte Webadressen liefert und dadurch eine Art von virtuellen Communities entstehen lässt. Im Folgenden werden „virtuelle Communities“ näher erläutert.

1.4 Virtuelle Communities im Internet

Mit dem Begriff „Community“ wird eine Gruppe von Personen bezeichnet, die ein gemeinsames Interessengebiet aufweisen. Die einzelnen Personen generieren einen Mehrwert aus dem Austausch von Wissen und Erfahrungen über das Thema ihres Interessensgebietes (vgl. Kim 2002).

Die Zunahme solcher Communities im Internet sowie der schnelle Anstieg der Mitgliederzahlen sind bemerkenswert. Tabelle 2 gibt einen Überblick über verschiedene Arten von Communities und deren aktuelle Mitgliederzahlen.

---

29 Weiterführende Informationen: http://www.citizendium.org/ (Abruf: 09.11.2006).
30 Weiterführende Informationen: http://www.dmoz.de/ (Abruf: 5.11.2006).
31 Weiterführende Informationen: http://www.doaj.org/ (Abruf: 15.11.2006).
32 „Peer-Review“ (engl. gleichrangige oder kollegiale Überprüfung) bezeichnet ein Verfahren zur Beurteilung von wissenschaftlichen Texten durch unabhängige Gutachter mit dem Ziel der Qualitätssicherung.
33 Vgl. http://www.nexttenyears.de/ (Abruf: 05.11.2006).

	Netzwerk - Community Name	Mitgliederzahl	Stand
Instant-Messaging und VOIP	MSN Messenger	155 Mio.[34]	April 05
	Jabber	4 Mio.[35]	Oktober 03
	ICQ	140 Mio. davon 6 Mio. aktive[36]	Juni 03
	Skype	6 Mio.[37]	März 06
Soziale- und Informations-Netzwerke	XING	1.5 Mio.[38]	Oktober 06
	MySpace	100 Mio.[39]	September 06
	Studivz	1. Mio. (nach eigenen Angaben)[40]	September 06
	dmoz	74.719[41]	November 06
	Wikipedia	230.000 angemeldete[42]	November 06
Peer-to-Peer-Technologien	Napster	270.000[43] zahlende User; in Spitzenzeiten vor der Legalisierung 40 Mio.[44]	Januar 04
	Seti@home	5 Mio.; davon 1 Mio. aktiv; im ∅ sind 600.000 gleichzeitig online[45]	November 03
	Folding@home	1,7 Mio. registriert davon durchschnittlich 200.000 aktiv[46]	Oktober 06
	Worldcommunitygrid	231.758[47]	November 06

Tabelle 2: Übersicht über Communities und Mitgliederzahlen[48]

Eine Motivation der User zur Partizipation ist die Generierung von persönlichem Mehrwert für den User selbst. Diesen erlangt er zum Beispiel durch den Austausch von Informationen oder Dateien zu den jeweiligen Interessensschwerpunkten (vgl. Bernecker 2002: 353). Bei den Instant-Messaging-Projekten wird ein Mehrwert durch die direkte Kommunikationsmöglichkeit geschaffen. XING[49] und Studivz generieren zumeist positive Netzwerkeffekte, die sich durch reale Treffen ergänzen oder sogar Geschäftsbeziehungen entstehen lassen (vgl. Thiedeke 2003: 23). Wie die Mitgliederzahlen der Communities zeigen, unterliegt das Internet zur Zeit der

34 vgl. http://www.convergedigest.com/bandwidth/newnetworksarticle.asp?ID=14365
35 vgl. http://news.zdnet.co.uk/software/applications/0,39020384,39117160,00.htm
36 vgl. http://de.wikipedia.org/wiki/Instant_Messaging
37 vgl. http://share.skype.com/sites/de/2006/03/post_1.html
38 vgl. http://www.tagesspiegel.de/computer-tipps/archiv/22.09.2006/2789971.asp
39 vgl. http://www.boinc.de/madrid_de.htm
40 vgl. http://www.studivz.net/blog/?p=78
41 vgl. http://dmoz.org/
42 vgl. http://upload.wikimedia.org/wikipedia/commons/0/06/Benutzer_angemeldet.png
43 vgl. http://www.zdnet.de/news/business/0,39023142,39129417,00.htm
44 vgl. http://www.politikerscreen.de/index.php/Lexikon/Detail/id/72842/name/Napster
45 vgl. http://www.boinc.de/madrid_de.htm
46 vgl. http://folding.stanford.edu/stats.html
47 vgl. http://www.worldcommunitygrid.org/stat/viewGlobal.do
48 Die Tabelle ist eine eigene Darstellung. Alle für diese Tabelle angegebenen Quellen wurden am 02.11.2006 zuletzt abgerufen.
49 Anmerkung: seit Ende 2006 in XING umbenannt vormals openBC.

„Kultur des Mitmachens“ oder wie es im Fachjargon der Web 2.0-Bewegung genannt wird, „Partizipation“ und „Kollaboration“ (vgl. Stieler 2006). Das Entstehen von Communities und die Bereitschaft der Mitglieder, auf freiwilliger Basis Aufgaben und Verantwortung zu übernehmen, ermöglichet erst die oben dargestellten Projekte. Des Weiteren zeigt es, dass die Internetuser sich mit dem Medium Internet sowie den damit einhergehenden Problemen und Herausforderungen auseinandersetzen. Die mehrjährige Interneterfahrung der User spiegelt sich in deren Verhalten wider (vgl. Fisch/Gscheidle 2006).

Kann dieser Trend der kollaborativen Zusammenarbeit[50] in Communities mit der Verfolgung gemeinsamer Ziele einen Lösungsansatz für die Problemfelder der Suchmaschinenbranche darstellen? Können positive Netzwerkeffekte auch in der Suchmaschinenbranche erreicht werden? Um Antworten auf diese Fragen zu finden, wird der Begriff „Suchmaschine“ im anschließenden Kapitel 2 genauer erläutert.

2 Suchmaschinen

„digitale Trüffelschweine” (Rossig/Prätsch 2005)
„Suchmaschine – Selektiermaschine” (Wolling 2005)
„Suchmaschinen als Gatekeeper”[51] (Machill/Beiler 2006)
„Wegweiser im Netz” (Machill/Welp 2003)

Für die meisten Menschen ist eine Suchmaschine eine Blackbox (vgl. Lehmann/Schetsche 2005: 53ff.), aus der man Informationen bzw. einen Verweis auf eine Information als Antwort auf eine Suchanfrage im Internet erhält. Weiter betrachtet fungiert sie für den User im Alltagsleben als Suchhilfe, die dazu dient, Webseiten mit bestimmten Informationen im Internet aufzufinden. Die genauere Arbeitsweise bleibt jedoch vorerst unbeachtet (vgl. Fuchs-Kittowski/Schewe 2002: 206ff.). In der Informationswirtschaft wird eine Suchmaschine detaillierter betrachtet. Sie sammelt Daten, selektiert diese und nimmt abschließend eine Beurteilung vor. Aufgrund dessen werden Suchmaschinen im Folgenden genauer untersucht.

---

50 Der Begriff „kollaborative Zusammenarbeit“ wird oft mit dem Begriff „Crowdsourcing“ beschrieben.

51 Gatekeeper (engl. Pförtner, Wächter). Im Kontext der Suchmaschinen kontrollieren und selektieren „Suchmaschinen als Gatekeeper“ die Informationen des Internets.

2.1 Eingrenzung und Definition des Begriffs „Suchmaschine“

Der Begriff „Suchmaschine“ ist erst seit 1999 mit folgendem Worteintrag im Duden vorhanden:

> *„Suchmaschine, die: auf ein bestimmten Namen lautendes Programm im Internet, das mithilfe umfangreicher, aus Internetadressen bestehender Datenbank die gezielte Suche nach Informationen im Internet ermöglicht.“*

Unter dieser sehr allgemein gehaltenen Definition können Suchdienste wie Webkataloge gefasst werden. Ein **Webkatalog** ist ein manuell erstelltes Verzeichnis von Internetadressen mit Webseiten und ist ein Synonym für den Begriff „Webverzeichnis“ (vgl. Strauch/Kuhlen/Laisiepen 2004a: 129). Alle Webseiten, die in das Verzeichnis aufgenommen werden, werden vor der Entscheidung über Aufnahme oder Ablehnung in das Verzeichnis von Mitarbeitern des Suchdienstes geprüft und redaktionell bewertet. Die aufgenommenen Webseiten und Links werden mit Schlagworten inhaltlich beschrieben und meist nach thematischen oder alphabetischen Kriterien sortiert (vgl. ebd.). Durch die manuelle, intellektuelle Erstellung des Verzeichnisses ist ein hoher Qualitätsstandard gegeben. Der personelle und finanzielle Aufwand der Verzeichniserstellung ist allerdings sehr hoch (vgl. Glöggler 2003: 2ff.). Des Weiteren enthält dieser Suchdienst nur eine kleine und selektierte Auswahl an im Internet verfügbaren Informationen. Zu den bekanntesten Webkatalogen gehören unter anderem das ODP (Open Directory Projekt) und das früher Webverzeichnis Yahoo.

Aus dieser Betrachtung lässt sich bereits die Metasuchmaschine ausgliedern. Eine **Metasuchmaschine** verfügt nicht über eine eigene Datenbank. Sie ist ein Recherchetool, das über eine eigene Webseite verfügt und bei Suchanfragen gleichzeitig auf mehrere Datenbanken anderer Suchmaschinen zugreift (vgl. Glöggler 2003: 8ff.). Das Ergebnis ist eine gemischte Trefferliste aus Ergebnissen der verschiedenen Suchmaschinen. Hierbei wird versucht, die Ergebnisse in der Trefferliste von Dubletten zu bereinigen. Diese Form der Suchmaschine ermöglicht es, mit einer einzigen Suchanfrage mehrere Suchdienstanbieter parallel abzufragen (vgl. ebd.). Als bekannte Metasuchmaschinen sind hier Metacrawler und MetaGer zunennen.

Ergänzend zu der oben stehenden sehr allgemein gehaltenen Definition des Dudens, beschreibt der Brockhaus (2005) eine Suchmaschine als:

> *„ein System, bestehend aus einem oder mehreren Computern sowie einer leistungsfähigen Software, das eine Sammlung von Dokumenten nach bestimmten Inhalten durchsucht (Dokumentensuchsystem). Der Begriff wird*

> *meist eingeschränkt auf das Internet verwendet, die zu durchsuchende Dokumentensammlung ist hierbei das komplette Internet.“*

Diese Definition unterscheidet sich von den vorherigen in der nicht zwingenden Verwendung auf das Internet. Somit kann auch die **Desktopsuche**[52] - Suche auf einer lokalen Festplatte - oder eine Suchmöglichkeit in einem firmeninternen Intranet unter diese Definition gefasst werden. Hier kann faktisch sowohl ein Webverzeichnis als auch die Metasuchmaschine eingeordnet werden.

Eine fachspezifischere Begriffserklärung geben Strauch, Kuhlen und Laisiepen. Die Autoren beschreiben eine Suchmaschine als diejenigen

> *„Webseiten oder Softwareprogramme, die in der Lage sind, eine große Anzahl anderer Webseiten in regelmäßigen Abständen nach bestimmten Regeln zu durchsuchen und deren Inhalte (meist Text aber auch Bilder und andere Formate) in Verzeichnissen abzubilden und abzulegen, um sie für eine inhaltsbezogene Suche zugänglich zu machen. Suchmaschinen machen einen Versuch, alle Informationen im Internet auf einem zentralen System abzubilden. Durch die Verwendung von komplexen Algorithmen bei der Verzeichniserstellung wird die Rangfolge in der Trefferliste beeinflusst.“ (vgl. Strauch/Kuhlen/Laisiepen 2004a: 116).*

Bei dieser Definition ist hervorzuheben, dass die Datensammlung auf einem **zentralen System** abgelegt werden soll. Demnach ist die Metasuchmaschine hier ebenfalls auszugrenzen. In diese Definition fließt jedoch ein weiterer Aspekt ein: Während sich die beiden Definitionen von Duden und Brockhaus auf das Finden und Abspeichern der Daten beschränken, wird hier zusätzlich auf die Ausgabe der Ergebnisse und deren Reihenfolge in der Trefferliste eingegangen.

Glöggler hingegen grenzt den Begriff der Suchmaschine durch vier Kernfunktionen einer Suchmaschine ein: welche die Datenbeschaffung, Datenanalyse, der Aufbau und die Verwaltung von Datenstrukturen sowie die Ausgabe von Suchanfragen mit der Berechnung der Relevanzwerte sind. Sie dürfen ausschließlich auf **automatisierten Verfahren** beruhen (vgl. Glöggler 2003: 5). Der Fokus dieser Definition liegt auf der Automatisierung der gesamten Datenverarbeitung in einer Suchmaschine. Unter dem Aspekt der automatisierten Verfahrensweise fällt der Webkatalog somit nicht unter den Begriff der Suchmaschine nach Glöggler, da hier das Sammeln sowie Analysieren der Webseiten manuell durchgeführt wird.

Eine **Payed-Listing-Suchmaschine** kann hinsichtlich der nicht automatisch generierten, sondern gekauften Rangposition ebenfalls nicht unter dieser Definition eingeordnet werden (vgl. Strauch/Kuhlen/Laisiepen 2004a: 94). Das Prinzip dieses

---

[52] Vgl. http://de.wikipedia.org/wiki/Desktopsuche (Abruf: 15.11.2006).

Suchmaschinentyps beruht auf dem Verkauf von Positionen bei anderen Suchdiensten gegen Höchstgebot (vgl. ebd.). In der Regel werden diese Ergebnisse in den Trefferlisten als „Sponsored Links“ gekennzeichnet. Zu den bekanntesten gehören Overture[53], Espotting und QualiGo.

An dieser Stelle wird eine eigene Begriffsklärung in Anlehnung an die oben aufgeführte Definition von Glöggler für die vorliegende Arbeit vorgenommen. Diese beschreibt die vier wesentlichen Kernfunktionen einer Suchmaschine bereits im Sinne dieser Arbeit. Der Aufbau und die Verwaltung von Datenstrukturen bedürfen jedoch einer detaillierteren Betrachtung. Eine Suchmaschine wird daher im weiteren Verlauf als ein komplexes System, bestehend aus vier Komponenten verstanden denen die vier Kernfunktionen zugeordnet werden. Eine Eingrenzung (der Anwendung einer Suchmaschine) auf das Internet ist dabei nicht zwingend erforderlich.

- Der ersten Komponente wird die Kernfunktion der **Datensammlung** zugeordnet und durch ein automatisiertes Computerprogramm umgesetzt welches im Folgenden als Crawler[54] bezeichnet wird. è Der Prozess ist das Sammeln.
- Die Aufgabe, die gesammelten Daten zu **analysieren**, wird durch Komponente zwei einem komplexen Computerprogramm, durchgeführt und unter der Bezeichnung Indexer weiter verwendet. è Der Prozess nennt sich automatisches Indexieren (vgl. Kapitel (II) 5.1.1).
- Die Abspeicherung erfolgt automatisch in ein zentrales oder dezentrales Datenbanksystem und stellt somit die dritte Komponente der Suchmaschine dar. Kernfunktion drei ist zuständig für die strukturierte **Abspeicherung** der vom Indexer analysierten Datenmengen.
- Die vierte Komponente ist eine **Benutzerschnittstelle** in Form einer Webseite. Diese dient der Interaktion mit dem User und der Ausgabe von Suchergebnissen. Die Berechnung der Relevanzwerte der einzelnen Suchergebnisse wird als Ranking bezeichnet.

Ausschlaggebend sind der **automatisierte Ablauf** der einzelnen Prozesse in jedem der vier Komponenten und die Erstellung eines **eigenen Datenbanksystems**. In

---

[53] Overture gehört seit 2003 zu Yahoo. Vgl. http://www.content.overture.com/d/ (Abruf: 15.12.2006).
[54] Der Crawler wird in der Fachliteratur auch als Spider, Robot oder Bot bezeichnet.

Abb. 3 sind die zuletzt genannten, signifikanten Aspekte übersichtlich dargestellt und eine Abgrenzung der einzelnen Suchmaschinenbegriffe visualisiert.

Suchdienste	automatisch generiertes Datenverzeichnis	automatisch generiertes Ranking	eigenes Datenbanksystem
(Web)-Suchmaschine	X	X	X
Desktopsuchmaschine	X	X	X
Webkatalog		X	X
Metasuchmaschine		X	
Payed-Listing			X

Abb. 3: Abgrenzung Suchmaschinen[55]

Zusammenfassend ist festzuhalten, dass mit einer Suchmaschine im allgemeinen der Service eines Internetsuchdienstes assoziiert wird. Schlussfolgernd darf der Term Suchmaschine nicht synonym mit dem Term Suchdienst verwendet werden. Ein Suchdienst ist allgemeiner zu betrachten und stellt einen übergeordneten Begriff der Suchmaschine dar.

Nach der vorgenommenen Begriffseingrenzung für die vorliegende Arbeit werden Webkataloge, Metasuchmaschinen und Payed-Listing-Suchmaschinen im weiteren Verlauf nicht näher betrachtet und die (Web)-Suchmaschine wird fokussiert. Zur Vollständigkeit ist an dieser Stelle noch eine weitere Unterteilung der (Web)-Suchmaschinen in Allgemeine- oder Universal- und themenspezifische Suchmaschinen zweckmäßig. **Universalsuchmaschinen** durchsuchen das gesamte Internet ohne Einschränkung auf bestimmte Informationen, Sprachen, Themen oder Länder. Zu den bekanntesten Universalsuchmaschinen zählen unter anderem Google, Yahoo[56] und MSN-Suche. Im Gegensatz dazu decken **Themenspezifische Suchmaschinen** nur einen ausgewählten Teilbereich des Netzes ab, der thematisch oder durch die Art der gefundenen Dokumente eingeschränkt sein kann. Beispielhaft können Jobsuchmaschinen wie Jobpilot und Monster, Suchmaschinen für Produkt- und Preisvergleiche wie Kelkoo, Expedia für Reisen und Scout24 für Immobilien genannt werden.[57]

In dieser Arbeit wird bei der Verwendung des Begriffs „Suchmaschine“ eine Universalsuchmaschine im Internet verstanden. Ist ein anderer Suchmaschinentyp

---

55 Quelle: eigene Darstellung.

56 Yahoo ist ursprünglich als Webverzeichnis/Katalog entstanden. Durch den Aufkauf der Firma Inktomi mitsamt Datenbestand und Crawler-Technologie ist Yahoo zu einer Kombination aus Webverzeichnis und Webkatalog geworden.

57 Vgl. http://www.suchfibel.de/4spez/index.htm (Abruf 15.11.2006).

gemeint, wird dies gesondert gekennzeichnet. Als Einstieg wird eine kurze Übersicht der Geschichte und den aktuellen Stand der Suchmaschinenentwicklung gegeben.

2.2 Historie und State-of-the-Art des Suchmaschinen-Marktes

Die Geschichte der Suchmaschinen ist noch recht jung, dennoch wird für die Einordnung der Suchmaschinen in das Untersuchungsgebiet ein kurzer Überblick der Suchmaschinen-Geschichte gegeben.

Entstanden sind Suchmaschinen aus Forschungsprojekten an Universitäten und deren Laboratorien. Als Pionier unter den Suchmaschinen gilt das von der kanadischen McGill-Universität entwickelte Suchtool „Archie“. Die Relevanz des Suchtools spiegelt sich darin wieder, dass es bereits 1992 zu den am meist genutzten Internetdiensten zählte.[58] Mit der kostenlosen Freigabe des Internet-Standards folgten diverse Suchdienstanbieter mit verschiedenen Suchtechnologien, von denen die meisten heute nicht mehr existieren. Einer der Ersten war das Such-Robotsystem “The Wanderer“, der das damals in seiner Größe übersichtliche Web durchsuchte und katalogisierte. 1994 veröffentlichten zwei Studenten eine Sammlung ihrer besten Webadressen und somit war Yahoo geboren. 1995 erfuhr der Suchmaschinenmarkt eine Trendwende, als die ersten von kommerziellen Firmen entwickelten Suchmaschinen auf den Markt kamen. Unter anderem waren aus diesen Entwicklungen Infoseek, Architext (später umbenannt in Excite) und AltaVista entstanden. Ein Jahr später wurde die Inktomi Corporation gegründet, deren Datensammlung von anderen Suchdiensten wie HotBot als Grundlage genutzt wurde. Ende 1998 wurde die nächste bedeutungstragende Entwicklung veröffentlicht: Page und Brin präsentierten ihre innovative Suchmaschinen-Ranking-Technologie Google auf dem Markt (vgl. Dielkmann 2006). Damit waren sie in der Lage, eine damals herausragende Qualität an Suchergebnissen innerhalb einer kurzen Antwortzeit zu liefern. In den Jahren 2003 und 2004 entstand ein starker Konkurrenzkampf unter den Suchdienstanbietern um die Anzahl der Webseiten im Datenbankverzeichnis und die Anzahl der User. Dieser führte zu gegenseitigen Übernahmen und Fusionen, wodurch eine komplexe Verflechtung der Suchdienste entstand und eine enorme Unübersichtlichkeit auf dem Suchmaschinenmarkt

[58] Vgl. http://www.bsi-fuer-buerger.de/suchmaschinen/geschichte.htm (Abruf: 15.11.2006).

resultierte. Hervorzuheben ist an dieser Stelle, dass sich Konkurrenten teilweise untereinander Suchergebnisse liefern.[59]

Der Suchmaschinenmarkt ist mittlerweile einer der wichtigsten Marketingbereiche im Wirtschaftsgeschehen (vgl. Rabe 2006). Im ersten Quartal 2005 benutzten 89 Prozent aller User eine Suchmaschine (vgl. Statistisches Bundesamt 2005)[60].

Im Wesentlichen gibt es heute drei große Anbieter die sich den Suchmaschinenmarkt untereinander aufteilen. Diese sind Google, Yahoo und die MSN-Suche von Microsoft. Eine Vielzahl kleinerer Suchdienste teilt sich den restlichen Markt. Die folgenden Abbildungen spiegeln die Marktanteile der Suchmaschinenbetreiber in Deutschland und den USA wider.

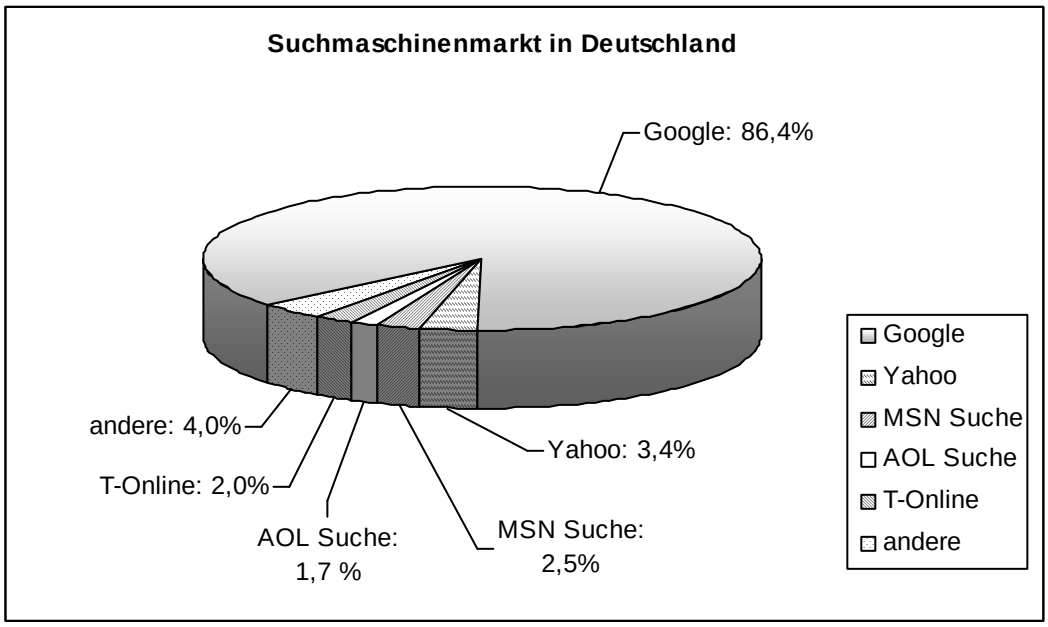


Abb. 4: Suchmaschinenmarkt in Deutschland[61]

[59] In einer Grafik von Karzauninkat 2004 wird das Beziehungsgeflecht der Suchmaschinen anschaulich darstellt. Siehe hierzu: http://www.suchfibel.de/5technik/suchmaschinen_beziehungen.htm (Abruf: 14.12.2006).

[60] Vgl. http://www.destatis.de/download/d/veroe/Tabellenanhang_Haushalte_IKT2005.pdf (Abruf: 15.11.2006).

[61] Quelle: http://www.webhits.de (Abruf: 18.01.2007).

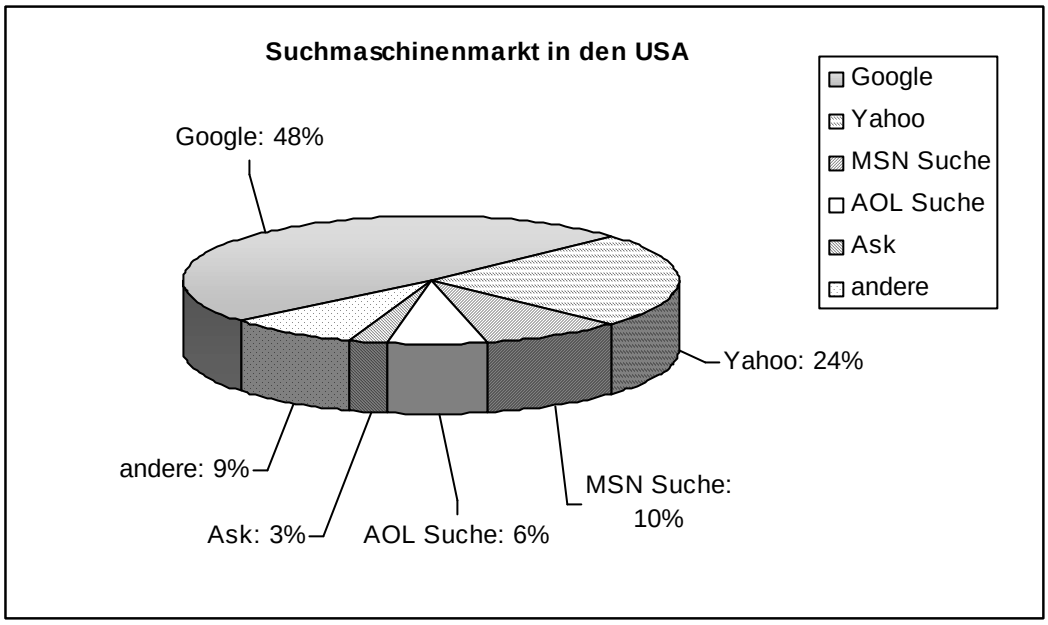


Abb. 5: Suchmaschinenmarkt in den USA[62]

Die Abbildungen zeigen, dass Google mit großem Abstand sowohl im deutschen Markt also auch im amerikanischen Markt den größten Marktanteil aufweist.

Neue Trends im Suchtechnologie-Sektor sind das Erkennen semantischer Zusammenhänge und linguistische Suchmaschinen, die sich allerdings noch im frühen Forschungsstadium befinden. Demnach sollen Suchmaschinen in Zukunft die Bedeutungen von Wörtern erfassen können. Wer beispielsweise nach „Büchern von Kindern“ sucht, wird Kinderbücher aufgelistet bekommen, aber nicht Bücher, die von minderjährigen Autoren stammen.[63] Aktuell erlangt die Personalisierung[64] von Suchergebnissen eine große Bedeutung. Der Fokus der technologischen Entwicklung wandelt sich jedoch weg von einer Fixierung auf optimale Rankingverfahren hin zu einer grundlegenderen Orientierung an den subjektiven Bedürfnissen des einzelnen Users (vgl. Griesbaum/Bekavac 2004: 2). Weitere Trends sind das sogenannte „Personal-Ranking“ Verfahren und Multimedia-Suchmaschinen. In Zukunft sollen Suchdienste auch in der Lage, sein Farb- und Formschemata zu identifizieren. Diese Art des Retrievals wird unter dem Begriff „Multimedia-Retrieval“ zusammengefasst.

---

62 Quelle: http://www.searchenginewatch.com (Abruf: 08.11.2006).

63 Vgl. http://www.netzeitung.de/internet/433461.html (Abruf: 19.11.2006).

64 Im Kontext von Suchmaschinen ist unter dem Begriff Personalisierung ist die Anpassung der Suchergebnisse auf das Informationsbedürfnis der User zu verstehen. vgl. http://www.muenchner-kreis.de/site/fileadmin/dateien/HTML/Vortraege.htm (Abruf: 23.11.2006).

2.3 Nutzung von Suchmaschinen

Die drei häufigsten Gründe, die zur Nutzung einer Suchmaschine führen, sind: die Recherche nach fachlichen Informationen, die Suche nach Informationen zu Produkten oder die Suche nach Informationen zu Personen (vgl. Schmidt-Mänz/Bomhardt 2005: 5-8). Um eine Aussage über das Suchverhalten der User machen zu können, werden diese in drei Usergruppen eingeteilt. Experten wie Informationswirte - die für Informationsrecherchen ausgebildet bzw. darauf spezialisiert sind, kennen und recherchieren in den verschiedenen Datenbanksysteme und mit deren Recherche-Tools und Techniken. Ein professioneller User, der Bedarf an wissenschaftlichen Fachinformationen hat, kennt und recherchiert überwiegend in den einschlägigen Fachdatenbanken. Dabei wird das Internet eventuell als Einstiegs-Informationsquelle verwendet. Die dritte Gruppe bilden die Laien, die sich ausschließlich mit dem Oberflächenweb und den dort angebotenen Suchmaschinen auseinandersetzen (vgl. Stock/Lewandowski 2005: 2). Für den weiteren Verlauf dieser Arbeit wird von einem Durchschnittsuser ausgegangen, in dessen Alltag das Medium Internet sowie die Nutzung von Suchmaschinen fest etabliert sind und der über grundlegende Anwendungskenntnisse verfügt.

Themenschwerpunkte von Suchanfragen

Es ist bekannt, dass User eine Vielzahl von Themen in einer Suchmaschine abfragen. Spink hat anhand von Logfile-Analysen eine Klassifizierung der Suchanfragen in elf Themenbereiche vorgenommen. Die aktuellsten Ergebnisse für deutsche Suchmaschinenuser gehen aus einer Untersuchung von Lewandowski (2005) hervor. Die Untersuchung basiert auf Daten der Suchmaschinen MetaGer, Fireball und Seekport[65]. Pro Suchmaschine wurden 500 Suchanfragen verteilt über den Tag erhoben, um möglichst realistische Ergebnisse zu erhalten. In der nachstehenden Tabelle sind die Themenschwerpunkte in der Reihenfolge ihrer häufigsten Nachfrage dargestellt. Platz eins nimmt der Themenbereich “Commerce, travel, employment or economy“ ein.

[65] Ein Zugang zu den Logfiles der marktführenden Suchmaschinen war nicht vorhanden. Laut Schmidt-Mänz dürfte jedoch kein Unterschied zu den Suchanfragemustern zwischen den kleinen und großen Suchdienstanbietern bestehen.

Durchschnittliche Häufigkeit aller 1500 Suchanfragen bei drei Suchmaschinenanbietern 2005	
29,0%	Commerce, travel, employment or economy
12,8%	People, places and things
7,7%	Entertainment or recreation
7,4%	Computers and Internet
7,3%	Health or sciences
4,5%	Sex or pornography
4,0%	Society, culture, ethics or religion
3,4%	Government
2,1%	Education or humanities
1,2%	Performing or fine arts
20,1%	Unknown or other

Tabelle 3: Suchanfragen nach Themenbereichen[66]

Bemerkenswert an den Ergebnissen ist, dass der Themenbereich „sex or pornography“ nicht Platz eins, sondern Platz sechs einnimmt. Allerdings wurde eine Verzerrung im Bereich „Sex or pornography“ verzeichnet, da die Suchmaschine Seekport Anfragen zu dem Themenbereich in der „Livesuche“[67] rausfiltert.

2.3.1 Suchverhalten der User

Eine Forschung zum Thema Suchverhalten der User mit dem Titel „Muster in Suchanfragen“ brachte Erkenntnisse zu der durchschnittlichen Anzahl der Suchtermeingabe von 1,6 bis 1,8 Wörtern pro Suchanfrage. Suchanfragen mit nur einem Suchterm sind am häufigsten vertreten (vgl. Schmidt-Mänz 2006: 1). Diese Ergebnisse werden bei einem Vergleich mit den zehn beliebtesten Anfragen auf Google Zeitgeist[68] bestätigt. Im Gegensatz dazu lässt sich in einer Analyse der versehentlich offen gelegten Logfiles der AOL Research Abteilung ein Trend zur Eingabe von Suchanfragen mit zwei bis drei Suchtermen erkennen.[69] Des Weiteren geht aus der Veröffentlichung von Machill/Welp (2003: 224) hervor, dass dabei die Userkenntnisse bei der Verwendung von Suchoperatoren eine bedeutende Rolle spielen. Die Möglichkeit der Phrasensuche wurde in dieser Untersuchung nur sehr selten genutzt, sie lag bei 1,7 bis 2,1 Prozent aller Suchanfragen (vgl. Schmidt-Mänz 2006).

---

66 Quelle: Lewandowski 2006: 6.

67 Die Livesuche bietet die Möglichkeit zu verfolgen, was gerade gesucht wird. Weitere Informationen unter: http://www.seekport.de/q?liveseek

68 Vgl. http://www.google.com/press/zeitgeist/archive.html (Abruf: 15.12.2006).

69 Vgl. http://www.sistrix.com/news/494-aol-daten-eine-kurze-auswertung.html (Abruf: 15.12.2006).

Weiterhin wurde eine Klassifikation der Suchbegriffe nach Auftrittshäufigkeit innerhalb eines bestimmten Zeitrahmens vorgenommen. Die Suchbegriffe wurden eingeteilt in Eintagsfliegen, Evergreens (Dauerbrenner), Impulse und Events (vgl. ebd). Zu den **Eintagsfliegen** gehören Suchbegriffe, die nur sporadisch auftauchen. **Evergreens** sind Suchbegriffe, die eine konstant hohe Nutzung aufweisen (z.B. Kameras, Reise, Flug) und in 90 Prozent aller Suchanfragen vertreten sind (vgl. ebd). **Impulse** sind nachrichtenbedingte Suchanfragen (z.B. zum Tod von Papst Johannes Paul II. oder nach der Tsunami-Katastrophe). **Events** sind periodisch wiederkehrende Suchanfragen wie Weihnachten oder das wöchentliche Kinoprogramm (vgl. Schmidt-Mänz 2006). Ergänzend wurde festgestellt, dass eine starke Konzentration auf nur wenige Suchworte besteht. Die häufigsten zehn Suchwörter machen bereits 15 Prozent der gesamten Sucheingaben aus, die häufigsten 60 bereits 25 Prozent und nur 700 Wörter decken 50 Prozent aller Suchanfragen ab.[70]

Die folgenden Abbildungen illustrieren den Verlauf unterschiedlicher Suchanfrage-Klassen.

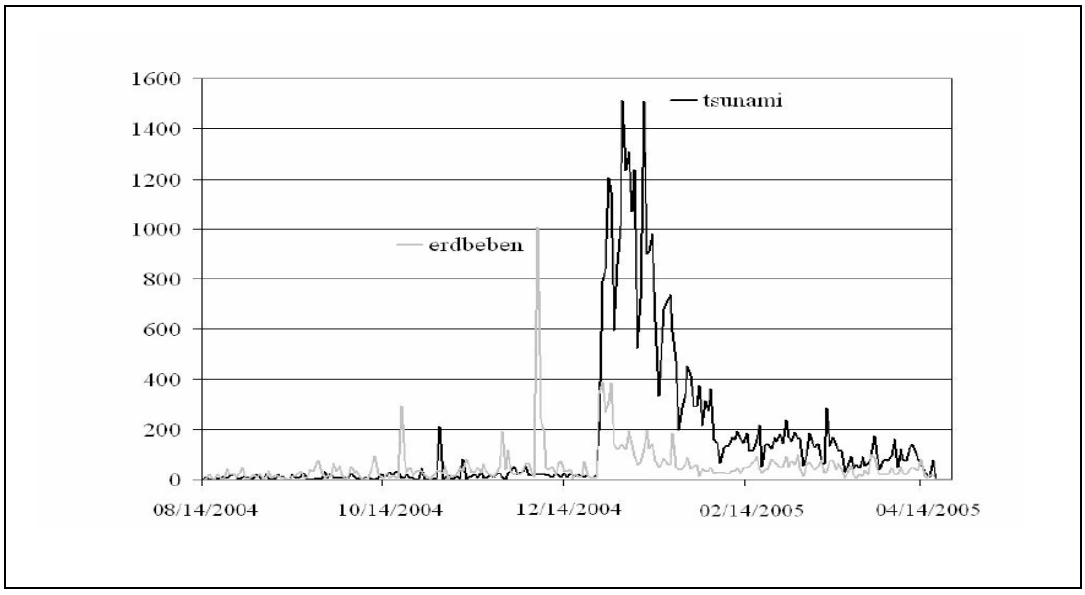


Abb. 6: Suchanfragenmuster von Impulsen [71]

In Abb. 6 wird der Verlauf von nachrichtenbedingten, plötzlich und zeitlich nicht bestimmbar auftretenden Impulsen visualisiert. Der Verlauf des Graphen visualisiert das plötzliche starke Aufkommen des Informationsbedürfnisses nach Erdbeben und Tsunami, das im weiteren Verlauf wieder abflacht. Ein Vergleich der plötzlich

[70] Vgl. http://www.suchfibel.de/aktuell/suchroboter_und_verzeichnisse.htm (Abruf: 20.11.2006).
[71] Vgl. Schmidt-Mänz 2006: Slide13.

aufkommenden sowie neuen, langsam ansteigenden Informationsbedürfnisse zeigt die folgende Abb. 7, welcher mit dem Google-Trends-Tool[72] konstruiert wurde.

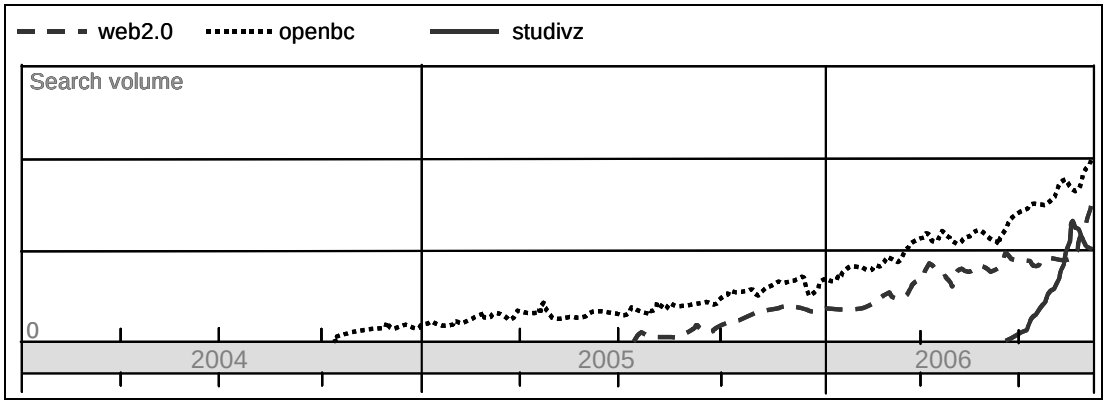


Abb. 7: Entdeckung neuer Trends[73]

Die folgenden zwei Abbildungen (Abb.8 und Abb. 9) sind Beispiele für die Kategorie „Suchanfragen zu periodisch wiederkehrenden Events“. Als Beispielbegriffe wurden „Sylvester“, „Ostern“ und „Weihnachten“ angeführt. Zur Überprüfung der Ergebnisse wurde durch Eingabe der Begriffe „Köln Marathon“, „Photokina“ und „IAA“ ein weiterer Graph erstellt. In beiden Abbildungen wird der leichte Anstieg vor dem Ereignis selbst und die abrupt aufhörende Nachfrage veranschaulicht.

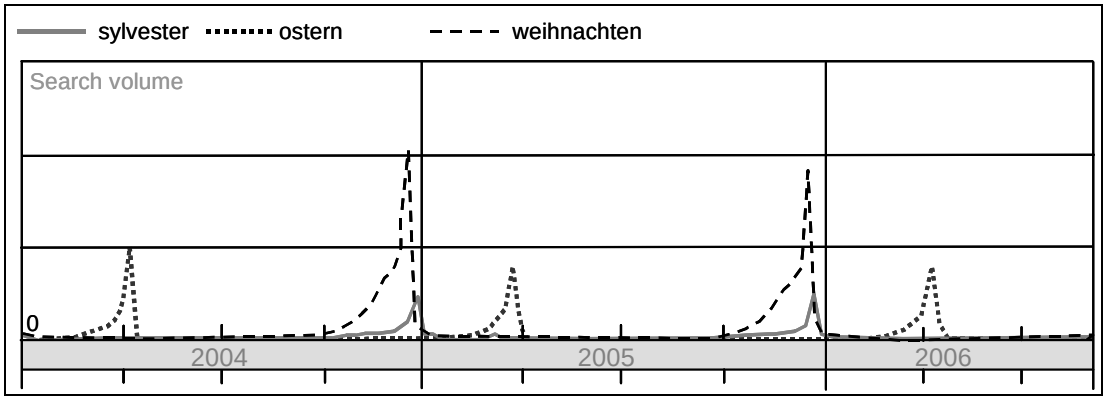


Abb. 8: Suchanfragenmuster von periodischen Events[74]

[72] Weiterführende Informationen: http://www.google.com/trends (Abruf: 15.12.2006).

[73] Erstellt am 28.10.2006 mit durch Eingabe der Suchbegriffe „web2.0“, „openbc“ und „studivz“ mit www.google.com/trends. (Abruf: 20.11.2006).

[74] Erstellt am 28.10.2006 durch Eingabe der Suchbegriffe „Sylvester“, „Ostern“ und „Weihnachten“ mit www.google.com/trends. (Abruf: 20.11.2006).

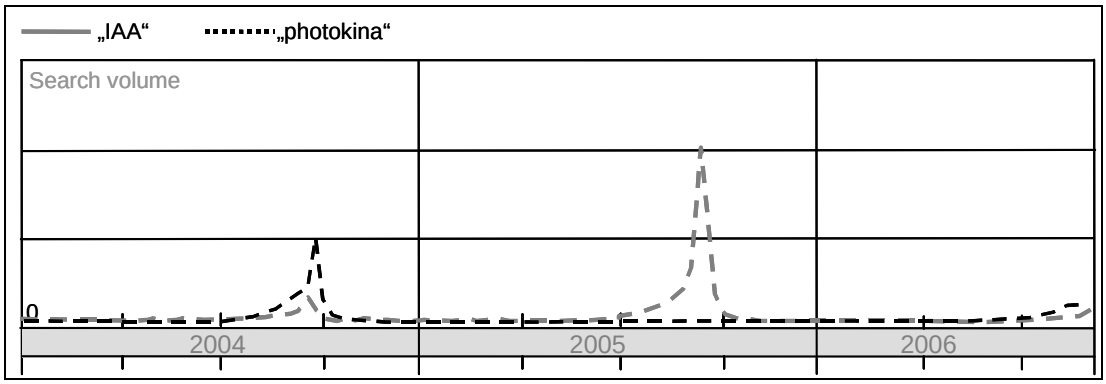


Abb. 9: Suchanfragenmuster von periodischen Events[75]

Die Abb. 10 visualisiert „Dauerbrenner“, die annähernd konstant über jede Zeitperiode hinweg nachgefragt werden. Bezüglich des Suchverhaltens der User (im Folgenden auch Informationsbedürfnis genannt) kann also eine starke Abhängigkeit von Jahreszeiten und anderen Ereignissen festgehalten werden.

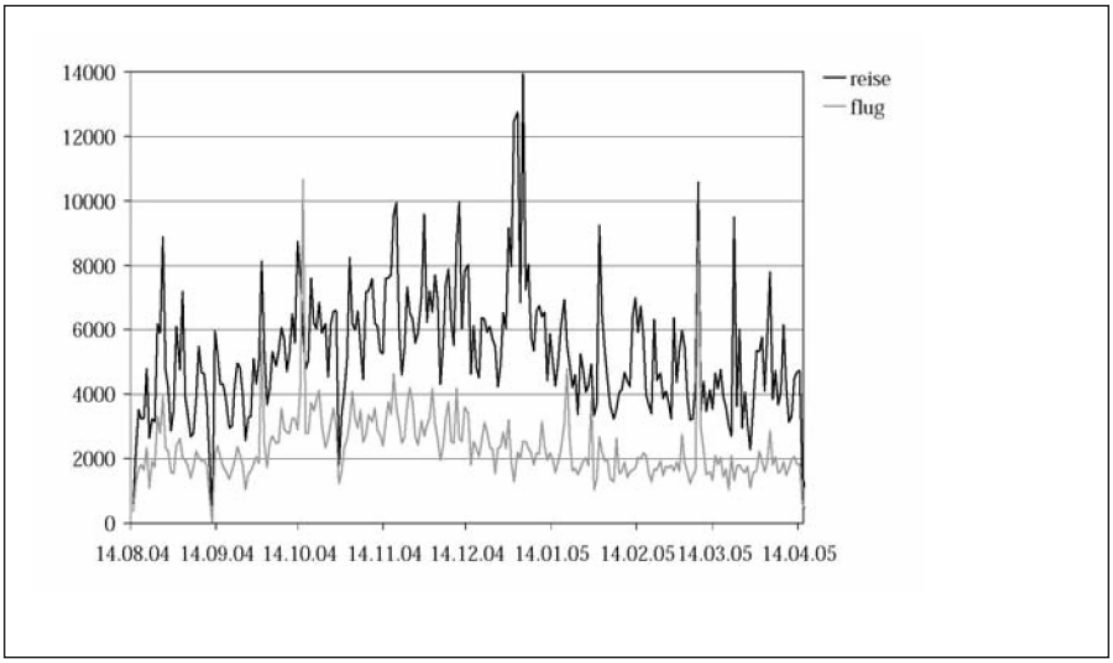


Abb. 10: Suchanfragenmuster von Dauerbrennern[76]

Diese von Schmidt-Mänz erläuterten Trends von Suchanfragenhäufigkeiten in entsprechenden Zeiträumen lassen sich durch das „Google-Trends“ Tool auch für die große Datenbank des Suchmaschinenbetreibers Google bestätigen.

---

75 Erstellt am 28.10.2006 durch Eingabe der Suchbegriffe „photokina“ und „IAA“ mit www.google.com/trends. (Abruf: 20.11.2006).

76 Vgl. Schmidt-Mänz 2006: Slide 11.

2.3.2 Anzahl und Arten von Suchanfragen

98 Prozent der Internetuser benutzen eine Suchmaschine. Die Frage, wie viele Suchanfragen pro Tag im Durchschnitt an eine Suchmaschine gestellt werden, lässt sich nicht leicht beantworten, denn die Suchmaschinenbetreiber halten sich mit Zahlenangaben generell zurück. Die Angaben in den folgenden Tabellen stammen aus einer Untersuchung von Schmidt-Mänz (2006). Diese Angaben konnten auf Basis der Livesuche einzelner Suchdienstanbieter erhoben und entsprechend ausgewertet werden. Die AOL-Daten basieren auf dem versehentlich im Web veröffentlichten Datenfile (siehe Kapitel 2).

Tabelle 5 enthält Angaben zu Brutto-Suchanfragen und Termen, welche die Gesamtzahl aller gestellten Anfragen/Terme darstellen. Die Netto-Angaben stellen die unterschiedlichen, eigenständigen Terme dar. Die Anzahl der User, die diese Suchanfragen gestellt haben, ist nur für die AOL-Daten vorhanden.

Suchmaschine	Brutto Suchanfragen	Netto Suchanfragen	Brutto Terme	Netto Terme	Zeitraum (Monat)
Fireball	132.833.007	17.992.069	241.833.877	6.296.833	13
Lycos	189.930.859	29.322.366	344.242.099	11.232.710	13
Metager (Top 4000 Terme)	4.407.566	678.655	7.333.343	430.338	10
Metaspinner	4.089.731	1.287.417	7.853.501	627.507	12

Tabelle 4: Anzahl Brutto- und Netto-Suchanfragen und Terme[77]

Suchmaschine	Brutto Suchanfragen	Netto Suchanfragen	Klick auf Anzahl unterschiedliche URLs	Anzahl unterschiedlicher User	Zeitraum (Monat)
AOL (sponsored by Google)	36.389.577	21.011.340	19.343.540	657.426	1

Tabelle 5: AOL-Daten zu Anzahl der Suchanfragen, Terme und User[78]

Arten von Suchanfragen

In einer Untersuchung zum Thema „Websuche und Userverhalten deutscher Suchmaschinenuser“ werden Suchanfragen in informationsorientierte, navigationsorientierte und transaktionsorientierte Suchanfragen eingeteilt (vgl. Broder 2002).

---

77 Vgl. Schmidt-Mänz 2006.
78 Vgl. http://www.sistrix.com/news/494-aol-daten-eine-kurze-auswertung.html (Abruf: 20.11.2006).

Informationsorientierte Suchanfragen werden von Usern gestellt, die sich über ein bestimmtes Thema informieren möchten (vgl. Lewandowski 2005: Kapitel 2.5). Der Suchende hat also keine bestimmte Seite als Ziel, sondern sammelt durch das Anschauen verschiedener Websites Informationen zu seinem Thema. Bei navigationsorientierten Suchanfragen ist es das Ziel des Suchenden, eine bestimmte Seite zu finden (vgl. ebd.). Diese kann ihm bereits bekannt sein oder er ahnt deren Existenz. Bei der Suche nach großen Firmen, deren Name auch gleichzeitig in der URL[79] enthalten ist, kann dies beispielsweise der Fall sein. Bei transaktionsorientierten Suchanfragen strebt der Suchende die Durchführung einer Transaktion an. Als Transaktion werden z.B. der Download einer Datei, ein Online-Einkauf oder eine Datenbankrecherche beschrieben.
Aufbauend auf diese Einteilung fanden 2004 weitere Untersuchungen von Rose und Levinson (2004) statt. 61 bis 63 Prozent der Suchanfragen werden dabei den informations-, 11 bis 15 Prozent den navigations- und 21 bis 27 Prozent den transaktionsorientierten Suchanfragen zugeordnet.
Ein weiterführender Versuch von Lewandowski im Oktober 2005 anhand von 400 Suchanfragen ergab nach Auswertung der Logfiles, einen Anteil von 39 bis 48 Prozent an informationsorientierter Anfragen. 20 bis 24,5 Prozent wurden den navigations-orientierten Anfragen zugeordnet und 22 bis 30 Prozent den transaktionsorientierten. Vergleicht man diese Ergebnisse, lassen sich Abweichungen feststellen. Dennoch wird deutlich, dass sich der Anteil an informationsorientierten Anfragen gegenüber den beiden anderen Bereichen abhebt. Diese Ergebnisse lassen sich durch die ARD/ZDF-Online-Studie 2004 belegen (vgl. Eimeren/Gerhard/Frees 2004: 355ff.). Diese beobachtete die gleichen Zugangswege zu den Webseiten im Web: Zum einen das Suchen mithilfe einer Websuchmaschine und zum anderen die Navigation durch direkte Eingabe der Internetadresse.

2.3.3 Selektionsverhalten der User in Suchergebnisseiten

Das Selektionsverhalten der User in den Suchergebnissen wird anhand Abb. 11 anschaulich illustriert. Diese Darstellung basiert ebenfalls auf dem bereits erwähnten AOL-Datenfile. Aufgrund der großen Datenmenge und der umfangreichen Userzahl, spiegelt diese Analyse das reale Verhalten der Suchmaschinenuser wider.

---

[79] Die URL (engl unified resource locator) ist der Speicherort einer Website und gleichbedeutend mit der Webadresse.

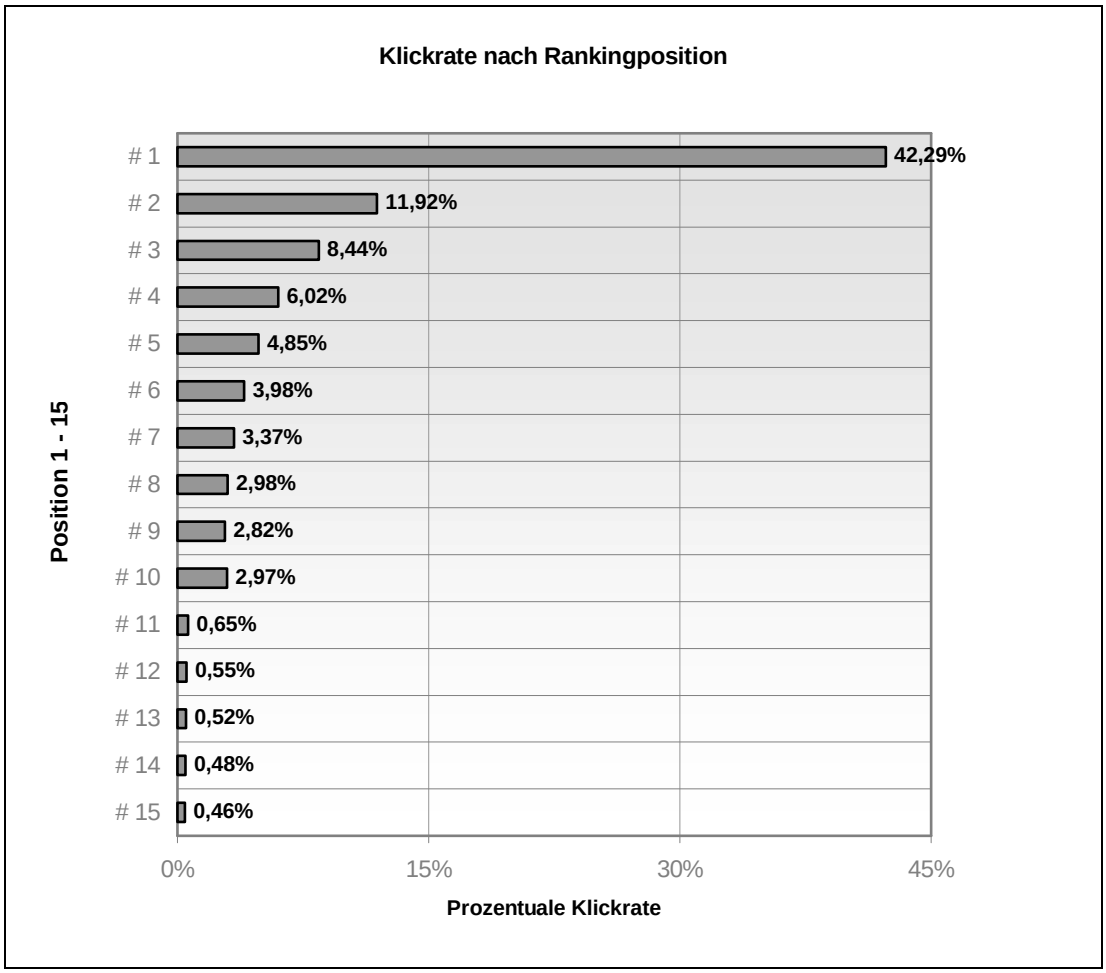


Abb. 11: Klickrate nach Rankingposition[80]

Abb. 11 zeigt, dass in über 80 Prozent aller Suchanfragen nur die erste Ergebnisseite betrachtet wird und die tatsächliche Auswahl auf die erste Rangposition aller angebotenen Links fällt. Im Rahmen einer Forschung mit dem Eye-Tracking-Verfahren, wurden diese Ergebnisse bestätigt und als „das Goldene Dreieck“ bezeichnet.[81]

Wie bereits anhand von Abb. 11 erkennbar, klicken maximal knapp ein Prozent der User auf die zweite Ergebnisseite. Der überwiegende Anteil der Suchmaschinenbetreiber gab in der Befragung von Machill einen Wert von 13 Prozent an. Aus diesen Ergebnissen kann der Schluss gezogen werden, dass zur Beurteilung der Qualität einer Suchmaschine die Gesamtzahl der Treffer nur eine untergeordnete Rolle spielt. Nur die ersten 20 Treffer, die in der Voreinstellung meist zwei Treffer-

---

80 Vgl. http://www.sistrix.com/news/494-aol-daten-eine-kurze-auswertung.html (Abruf: 20.11.2006).

81 Als „das Goldene Dreiecke“ wird der obere linke Bereich einer Suchmaschinentrefferliste bezeichnet, da dieser die meiste Aufmerksamkeit der User erhält. Vgl. http://www.prweb.com/releases/2005/3/prweb213516.htm und http://www.at-web.de/studien/goldenes-dreieck.htm (Abruf: 20.11.2006).

listenseiten entsprechen, sind für den User relevant. Eine vollständige Auswertung von mehreren Tausend oder sogar mehreren Millionen Treffern ist arbeitsökonomisch ausgeschlossen.

Über die Anzahl durchschnittlich angeklickter Treffer liegen in Machill (2003: 94ff.) stark variierende Zahlen vor. Die Angaben basieren auf einer Befragung der Suchmaschinenbetreiber selbst und liegen überwiegend bei einer durchschnittlichen Anzahl von zehn angeklickten Links pro Trefferliste. Machill beurteilt die Selbsteinschätzung der Suchmaschinenbetreiber als zu hoch und ermittelt einen Durchschnittswert von 4,5 angeklickten Links pro Ergebnisliste (vgl. ebd.).

Zusammenfassend kann festgehalten werden, dass über 90 Prozent aller User nur die erste Ausgabeseite nutzen und nur zehn Prozent auf die zweite Ergebnisseite blättern. Aus der Sicht eines durchschnittlichen Suchenden macht es letztlich keinen Unterschied, ob einige Hundert oder einige Tausend Treffer gefunden werden. Hölscher (2002) belegt in einer Studie, dass die Auswahl der Treffer stark von den Suchmaschinenkenntnissen der User abhängig ist. Im weiteren Verlauf dieser Arbeit wird von einem Durchschnittsuser ausgegangen.

3 Peer-to-Peer

> *„Wait until you see an online world as large as the web, unfettered by bottlenecks caused by primitive client/server technology. Everyone is gradually waking up to the fact that for an application to be of this global scale, it must be a distributed system. People are beginning to see that the future is peer-to-peer." (Crosbie Fitch 2000, Cyberspace in the 21st Century)*

In diesem Kapitel wird der Begriff „Peer-to-Peer" erläutert und die damit verbundenen Techniken skizziert. Der Stand der Technik soll eine Einschätzung über die Bedeutung der Technologie im Alltagsgebrauch geben. Um einen Gesamtüberblick zu erhalten, werden verschiedene Anwendungsbereiche kurz aufgeführt.

In den folgenden Kapiteln wird für den Ausdruck „Peer-to-Peer" die allgemein gängige Abkürzung „P2P" verwendet. Für den allein stehenden Begriff „Peer" werden im Folgenden mitunter die Begriffe „Rechner", „Client", „Node", „User" oder „Netzwerkteilnehmer" synonym verwendet.

3.1 Eingrenzung und Definition des Begriffs „Peer-to-Peer"

„Peer" steht im Englischen für Gleichgestellter und Ebenbürtiger. Allgemein betrachtet steht Peer-to-Peer in der Informatik für eine spezifische Netzwerk-

Architektur mit besonderen Eigenschaften.[82] In diesem Kontext kann der Begriff auch als gleichberechtigte Verbindung, Zusammenarbeit oder Kommunikation zwischen zwei oder mehreren Rechnern in einem Netzwerk bezeichnet werden. In der Literatur findet sich eine Vielzahl von variierenden Definitionen für Peer-to-Peer-Systeme. Im Folgenden werden verschiedene Definitionen in chronologischer Abfolge vorgestellt und diskutiert.

Shirky 2001

> *„P2P is a class of applications that takes advantage of resources - storage, cycles, content, human presence - available at the edges of the Internet. Because accessing these decentralized resources means operating in an environment of unstable connectivity and unpredictable IP addresses, Peer-to-Peer nodes must operate outside DNS (Domain Name System) and have significant or total autonomy of central servers.“*[83]

Mit dieser Definition kann ein P2P-System als eine Art “Internet auf Applikationsebene über das Internet“ verstanden werden (vgl. Dustdar/Gall/Hauswirth 2003: 162). Shirky schränkt die Anwendung eines P2P-Netzes in dieser Formulierung auf das Internet ein. Dennoch wird deutlich, dass eine Nutzung bislang ungenutzter Speicherressourcen durch das Zusammenspiel von Netzwerkkreisen, Usern und Content möglich wird. An dieser Stelle lässt sich bereits erkennen, dass es sich bei P2P-Technologien um Netzwerkgüter handelt, da das Internet als Basis für Peer-to-Peer fungiert. Besonders hervorzuhebende Aspekte sind zum einen die Dezentralität von Ressourcen und zum anderen die notwendigen Autonomie, welche zumindest teilweise oder sogar vollständig erlangt werden sollte.

Schoder, Fischbach, Teichmann 2002

> *„Peer-to-Peer-Systeme können als verteilte Systeme charakterisiert werden, in denen alle Knoten gleiche Rechte und Verantwortlichkeiten haben, alle Knoten somit gleichberechtigt sind. Ressourcen wie Informationen, CPU-Laufzeiten, Speicher und Bandbreite sollen wechselseitig zugänglich gemacht werden. Autonome Knoten, welche das System jederzeit betreten und verlassen können, tauschen ohne zentralen Server Daten untereinander aus. Jedes Teilsystem kann dabei sowohl Informationsanbieter als auch Informationskonsument sein.“*[84]

In dieser Definition findet keine Einschränkung auf das Internet statt. Jedoch wird das Hauptaugenmerk, wie in der vorherigen Definition, gleichermaßen auf die Eigenschaften Dezentralität und Autonomie gelegt. Hinzu kommt, dass die Ressourcen wechselseitig zur Verfügung zu stellen sind und eine Gleichberechtigung aller am Netz beteiligten Rechner erreicht werden soll. Daraus ergibt sich die

---

82 Vgl. http://www.itwissen.info (Abruf: 02.11.2006).
83 Quelle: Shirky (2001: 21).
84 Quelle: Schoder,Fischbach,Teichman (2002:3).

Sachlage, dass jeder Rechner sowohl Anbieter als auch Nachfragender von Daten oder Ressourcen seien kann.

Whatis.com 2004

> *„Peer-to-peer is a communications model in which each party has the same capabilities and either party can initiate a communication session. […] On the Internet, peer-to-peer […] is a type of transient Internet network that allows a group of computer users with the same networking program to connect with each other and directly access files from one another's hard drives."*[85]

Exakt wie in den beiden oben erwähnten Formulierungen werden hier ebenfalls Dezentralität und Gleichberechtigung fokussiert. verdeutlich wird hier erneut, dass Peer-to-Peer-Technologien zu Netzwerkgütern gehören und dass ein direkter Austausch von Daten zwischen den Netzwerkteilnehmern existiert.

Bigchampage 2005

> *„Peer-to-peer ("P2P") file sharing has been called a lot of things: a "piratical bazaar", a digital revolution, a consumer movement.... Certainly P2P is by now a genuine worldwide phenomenon and a lightning rod for controversy. Technically, P2P is an architecture for distributed computing. Socially, the best-known P2P networks are online swap meets, communities first popularized by Napster and now accessed using clients like eDonkey, Limewire, Bearshare, Kazaa and hundreds of others. P2P users congregate in a free exchange, chatting about, searching for and "sharing" digital music files as well as movies, images, software, games... virtually anything you can imagine parked in a shared folder on a desktop."*[86]

Aus dieser Erörterung des Peer-to-Peer Begriffs geht deutlich hervor, dass er nicht nur als eine technische Bezeichnung für eine Software-Architektur verwendet werden kann. Aus technischer Sicht spricht man zwar auch hier von dezentral verteilten Eigenschaften, zusätzlich wird jedoch ein sozialer Aspekt eingeworfen, der das Entstehen von Communities, in denen Tauschtreffen stattfinden, beschreibt. Die Tauschgüter in diesen Tauschtreffen sind digitale Güter wie Filme, Bilder, Texte, Software. Des Weiteren wird ebenso die direkte Kommunikation zwischen den Usern möglich.

3.1.1 Peer-to-Peer Eigenschaften

Wie sich anhand der oben aufgeführten Definition erkennen lässt, ist eine einheitliche klare Definition und Abgrenzung nicht zu formulieren. In einem Punkt stimmen jedoch alle Definitionen überein: sie versuchen, eine Abgrenzung zu dem im

---

85 Vgl. http://searchnetworking.techtarget.com/sDefinition/0,290660,sid7_gci212769,00.html (Abruf: 20.12.2006); Das Online Comupter Lexikon „Whatis.com" ist eine Empfehlung der Fachzeitschrift Wirtschaftsinformatik. Heft 3/2003: 335.

86 Vgl. http://www.bigchampagne.com/faqs.html?PHPSESSID=cf43ff430d64de4676ba4d9787 0449e8 (Abruf: 20.12.2006).

Gegensatz stehenden zentralistischen Client-Server-System zu finden. Neben dieser grundlegenden Abgrenzung kann Peer-to-Peer durch die folgenden drei wesentlichen Merkmale charakterisiert werden, die oftmals nur im Idealfall alle zusammen erfüllt werden. Diese sind: Dezentralität, Autonomie und gegenseitige Bereitstellung von Ressourcen (vgl. Schoder, Fischbach, Teichman 2002: 4).

- **Dezentralität** bedeutet zum einen, dass es keine zentrale Koordinationseinheit für die Organisation des Netzwerkes gibt, die für die Kommunikation innerhalb des Netzwerkes und die Ressourcennutzung zuständig ist. Somit ist eine zentrale Kontrolle der Peers nicht möglich. Die Kommunikation erfolgt ungefiltert und direkt zwischen den am Netz beteiligten Peers. Zum anderen ist unter Dezentralität die verteilte Nutzung der Ressourcen zu verstehen.

- Durch die **Autonomie** kann jeder Peer den Zeitpunkt sowie den Umfang der eigenen Ressourcen, welche er für andere Peers oder für das Netzwerk zur Verfügung stellen möchte, selbst bestimmen. Die teilnehmenden Peers sind voneinander unabhängig. Sie können beliebig in das Netz ein- oder austreten, ohne einen Verlust oder Zusammenbruch des gesamten Netzes zu verursachen (vgl. Shirky 2000: 22-23). Diese Autonomie gibt dem Netzwerk einen dynamischen Charakter.

- In einem Peer-to-Peer-Netz sind alle beteiligten Rechner gleichgestellt. Jeder Rechner kann sowohl als Client als auch als Server agieren. Dementsprechend können sie durch die **gegenseitige Bereitstellung von Ressourcen** (wie Informationen, Dateien, Speicherplatz und Rechnerleistung) sowohl Anbieter als auch Nachfrager sein.

Betrachtet man die verschiedenen Peer-to-Peer-Architekturmodelle und deren Anwendungen konkreter, wird deutlich, dass nicht zwingend jede der drei Eigenschaften erfüllt sein muss, um diese als Peer-to-Peer-Technologie bezeichnen zu können. Der wichtigste Aspekt ist die Dezentralität, die wie oben geschildert, zwei Ausprägungen annehmen kann. Im Folgenden werden die Architekturmodelle kurz beschrieben. Eine Überprüfung der oben genannten Eigenschaften trägt zur Verdeutlichung der Schwierigkeit einer klaren Abgrenzung bei. Des Weiteren befindet sich die Peer-to-Peer-Technologie konstant im Forschungsstadium und unterliegt somit einer ständigen Weiterentwicklung, wodurch sich neue Möglichkeiten und Eigenschaften entwickeln können (vgl. Fiutak 2005).

3.1.2 Peer-to-Peer Architekturmodelle

Ein Modell ist als ein vereinfachtes Abbild der Wirklichkeit zu betrachten. Es ist eine Abstraktion, die bewusst einzelne Merkmale zur Vereinfachung vernachlässigt, um grobe Modelleigenschaften hervorzuheben. Aus diesem Grund werden die P2P-Architekturen nach ihrer technischen Umsetzung und ihrer Topologie in vier Grundmodelle unterteilt: ein Zentrales Modell, ein Dezentrales Modell, ein Hybrides Modell und ein auf Distributed-Hash-Tabellen basierendes Modell (vgl. Schollmeier 2005: 6). In Steinmetz und Wehrle 2005 wurden die verschiedenen Modelle und deren Entwicklung anhand einer übersichtlichen Abbildung dargestellt (Abb. 12). In dieser Abbildung findet eine weitere Einteilung in eine Erste und eine Zweite Generation statt, anhand derer die technische Weiterentwicklung im Laufe der Zeit deutlich wird. Darüber hinaus wird die Abgrenzung zum Client-Server-Modell veranschaulicht.

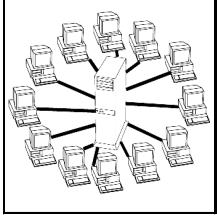
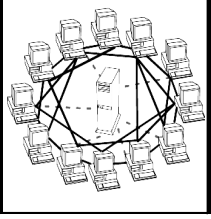
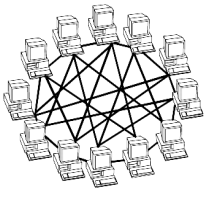
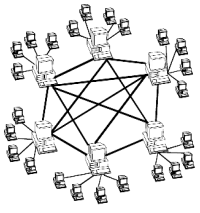
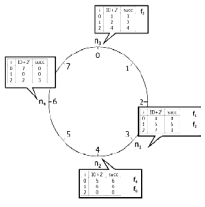
Client-Server	Peer-to-Peer			
	1. Dezentralität = keine zentrale Koordinationseinheit, direkte 2. Autonomie = freie Bestimmung von Zeitpunkt und Umfang der Beteiligung am 3. Gleichstellung = Peer kann sowohl Server als Auch Client sein (Servent) 4. Indexstruktur = dezentral / zentral			
	Unstrukturiertes P2P			Strukturiertes P2P
	1st Generation		2nd Generation	
	Zentrales P2P	Reines (pure) P2P	Hybride P2P	DHT-basierte P2P
1. Server als zentrale Einheit, einziger Anbieter von Inhalten und Services	1. Server als Vermittler der Peers	1. völlige Dezentralität – keine Hierarchie	1. Aufgabe des Servers teilen sich mehrere leistungsfähige Rechner (HKP)	1. völlige Dezentralität
2. Der Server ist in die höhere bietende Instanz.	2. Funktioniert nur wenn Server läuft	2. Keine Funktions-einbußen durch den Ausstieg eines Peers aus dem Netz	2. Keine Funktions-einbußen durch den Ausstieg eines Peers aus dem Netz	2. Keine Funktions-einbußen durch den Ausstieg eines Peers aus dem Netz
3. Die Clients die nachfragende, nehmende Instanz.	3. Direkter Kontakt und Austausch zwischen den Peers möglich	3. Alle Peers sind gleichgestellt, sowohl Server als auch Client	3. Dynamische (Hauptknotenpunkte= HKP)	3. Verbindungen im P2P Overlay Netz sind „fest“ definiert. Jedem Peer wird ein bestimmter Zuständigkeits-bereich zugewiesen
4. Index liegt auf dem Server	4. Index liegt auf dem Server	4. Index dezentral bei allen Peers die gerade online sind	4. Index liegt bei den online HKP	4. Index dezentral bei allen Peers die gerade online sind
Beispiel: WWW, Google	Beispiel: Napster, IM-Programme	Beispiel: Freenet, Gnutella 4.0	Beispiel: JXTA, Gnutella 6.0	Beispiel: Chord, Faroo
				

Abb. 12: Client-Server-Modell und Peer-to-Peer Modelle[87]

[87] Quelle: eigene Darstellung in Anlehnung an Steinmetz/Wehrle 2005: 36.

Jedes P2P-System impliziert, aufgrund der dezentralen Eigenschaft und der verteilten Abspeicherung und Nutzung von Ressourcen, immer eine Art von Suchmaschine. Bryn Loban subsummiert dies unter dem Begriff „Ressourcen Retrieval“ (vgl. Loban 2004), zumal nicht nur das Auffinden von „Content“, sondern auch das Auffinden ungenutzter Bandbreiten, Speicherkapazitäten oder CPU-Leistungen hierbei im Mittelpunkt steht. Daher wird in diesem Kapitel der Begriff der „Suche“ auf den Aufbau einer Kommunikationsebene zwischen den Peers sowie das Auffinden von Ressourcen bezogen. Die in Kapitel 2.1 erarbeitete Definition einer Suchmaschine kann demnach hier nicht angewendet werden.

Wie in Abb. 12 veranschaulicht, besteht die bislang als Standard geltende **Client-Server-Architektur** aus einer zentralen Servereinheit und angeschlossen Clients. Über diese zentral koordinierende und organisierende Servereinheit werden den Clients gespeicherte Daten und Informationen angeboten. Eine Vergabe von Rechten an die einzelnen Clients durch den Administrator ist möglich. Es können nur Daten und Services gefunden und genutzt werden, die der Verwalter der zentralen Einheit anbietet. Die Clients können in dieser Architektur keine eignen Dateien anbieten oder direkt untereinander austauschen. Die zentrale Einheit ist eine zusätzliche Hardware, für die finanzielle und personelle Ressourcen bereitgestellt werden müssen. Prinzipiell lässt sich festhalten, dass eine Abhängigkeit der Clients vom Server besteht. Ein Ausfall des Servers hat ein Ausfall für alle angebundenen Clients zur Folge.

Das **Zentrale-P2P-System** ist dem Client-Server-Modell sehr ähnlich, denn es besitzt ebenfalls eine zentrale Servereinheit. Jeder Peer ist mit dieser zentralen Einheit verbunden. Dieser Server dient als Lookup-Server, auf dem die Daten der am P2P-Netz angemeldeten Peers mit den zur Verfügung stehenden Ressourcen und Dateien in einem Verzeichnis eingetragen werden. Dieses zentral verwaltete Verzeichnis dient den Peers als Vermittlungszentrale und schafft eine Übersicht über die verfügbaren Ressourcen im dezentralen Netz. Anlässlich einer Suchanfrage wird von der zentralen Verwaltungseinheit eine Liste mit Peers, die die Anfrage beantworten können, erstellt. Aus dieser Liste wählt sich der anfragende Peer die gewünschte Datei aus und stellt eine direkte Verbindung zum anbietenden Peer her. Der Download erfolgt auf direktem Weg zwischen diesen beiden Peers (Steinmetz/Wehrle 2005: 37-38). In der Literatur wird diese Architektur aufgrund der zentralen Servereinheit als zentrales System bezeichnet. Da der Austausch der Daten jedoch in direkter Kommunikation der Peers erfolgt, kann es als ein P2P-Modell bezeichnet werden. Als Beispiele gelten das Filesharing-System Napster

sowie die Instant Messaging Progamme ICQ[88], MSN-Messenger[89] und Skype[90]. Ein weiteres Beispiel ist die Grid-Computing-Anwendung seti@home[91]. Diese Anwendung gilt hinsichtlich ihrer P2P-Zugehörigkeit je nach P2P-Definiton als umstritten. Hier werden Ressourcen von einer Vielzahl von Peers lediglich lokalisiert, zusammengebracht und genutzt. Eine direkte Kommunikation zwischen den einzelnen Peers findet jedoch nicht statt.

Ein weiteres Architekturmodell ist das **Dezentrale-P2P-Netz**, in der Literatur häufig auch „Pure P2P" genannt. Dieses gilt als idealtypisch und erfüllt am ehesten die drei oben genannten Eigenschaften. Es besteht aus der Vernetzung von einer Vielzahl gleichberechtigter Rechner, die alle in direkter Kommunikation miteinander stehen (Steinmetz/Wehrle 2005: 42-43). Im Gegensatz zu dem Zentralen System kommt das Pure P2P ohne jegliches zentral koordinierendes Element aus. In diesem System herrscht keine Hierarchie, jeder Teilnehmer ist Client und Server zugleich. Durch anspruchsvolle Algorithmen werden Verbindungsanfragen nicht von einem zentralen Server, sondern von den Peers selbst bearbeitet. Aufgrund dieser Eigenschaft leidet, ab einer bestimmten Netzgröße jedoch die Performance. Auf diese Art und Weise arbeiten Filesharing-Systeme wie zum Beispiel Freenet und Gnutella 0.4 (ebd.: 42-43).

Eine Weiterentwicklung sind die **Hybriden Netze.** Sie sind eine Mischung aus dem Zentralen- und dem Dezentralen System. Die Aufgabe des Servers teilen sich mehrere leistungsfähige Rechner, sogenannte „Hauptknotenpunkte". Diese Hauptknotenpunkte besitzen ebenfalls Autonomie und können dynamisch ihre Funktion an andere Hauptknotenpunkte abgeben (ebd: 49-51). Zur Verbesserung der Performance wurde das P2P-Protokoll von Gnutella 0.6 auf das Hybride System umgestellt.

Das letzte in Abb. 12 dargestellte Modell ist das strukturierte P2P-Netz auf Basis von sogenannten **Distributed-Hash-Tabellen** (DHT[92]). Es erfüllt ebenfalls alle drei vorab definierten Eigenschaften. Dieses P2P-Netz ist vollkommen dezentral, jeder Peer besitzt völlige Autonomie, die Ressourcen werden geteilt und die Kommunikation untereinander findet ohne jegliche Zwischeninstanz statt. Im Unterschied zu den oben beschriebenen P2P-Architekturmodellen, findet bei diesem

---

88 Weiterführende Informationen: http://www.icq.com/products/whatisicq-ger.html (Abruf: 10.11.2006)
89 Vgl. http://get.live.com/messenger/overview (Abruf: 10.11.2006).
90 Vgl. http://www.skype.com/intl/de/ (Abruf: 10.11.2006).
91 Durch die Installation der P2P-Software auf dem eigenen Rechner ist es möglich, diesem Projekt Rechnerressourcen für rechenintensive Aufgaben zu Verfügung zu stellen. Weiterführende Informationen: http://setiathome.ssl.berkeley.edu/ (Abruf: 10.11.2006).
92 engl. distributed Hash Tables

Modell aufgrund der Verwendung der DHTs eine Strukturierung und Zuteilung von Verantwortlichkeiten und Zuständigkeiten statt. Dies beeinflusst die Geschwindigkeit und Performance in großen Netzwerken positiv. Anwendung findet ein solches System in P2P-Netzen wie z.B. Chord.

3.1.3 Peer-to-Peer Zusammenfassung

Zusammenfassend wird Peer-to-Peer als Kommunikationsmodell zwischen zwei gleichberechtigten Partnern verstanden. Auf die Computertechnik übertagen bezeichnet man dies als „logisch dynamisches Netz im Internet“.[93] Dieses logisch dynamische Netz besteht aus gleichberechtigten, miteinander kommunizierenden Rechnern. Gleichberechtigt heißt hierbei, dass alle beteiligten Rechner in ihrer Verantwortung gleichgestellt sind. Kein Rechner hat eine gesonderte Verantwortung für Koordination oder Kontrolle von Prozessen innerhalb eines P2P-Netzes (vgl. Loban 2004). Alle an dem Netz teilnehmenden Peers haben direkten Zugriff auf Ressourcen anderer in diesem Netz befindlichen Rechner. Der Zugriff auf diese Ressourcen kann vom User individuell, mehr oder weniger restriktiv, bestimmt werden. Unter Ressourcen werden z.B. Speicherplatz, CPU-Laufzeiten und Dateien verstanden. Oft liegen diese Ressourcen brach, da sie vom Rechnerinhaber nicht genutzt werden. P2P-Technologien bieten die Möglichkeit, diese verteilten, dezentralen und ungenutzten Ressourcen zu lokalisieren, zu bündeln und zielgerichtet für eine Reihe von Anwendungen zu nutzen. Durch diesen Zusammenschluss können sogenannte „Super-Computer“ entstehen, die in der Lage sind, eine enorme Rechenleistung zu liefern (vgl. Frascaria 2002: 1ff.). Die Peer-to-Peer-Technologie ist ein verteiltes, Informationen bzw. Daten verarbeitendes System, das sich aus einer Vielzahl von eigenständigen und freiwillig beteiligten Rechnern zu einem Netzwerk zusammenschließt. Diese Rechner kooperieren - in einigen Fällen kommunizieren - direkt oder indirekt über ein Telekommunikationsnetzwerk mit anderen Rechnern, um ein gemeinsam angestrebtes Ziel zu erreichen (vgl. Loban 2004). Eine konkrete Zuordnung der P2P-Anwendungen in die einzelnen Architekturmodelle ist durch den modellhaften Charakter jedoch nicht immer möglich und auch nicht immer beabsichtigt. Oftmals finden Kombinationen dieser Modelle Anwendung.

---

[93] Vgl. whatis.com http://searchnetworking.techtarget.com/sDefinition/0,,sid7_gci212769,00.html, (Abruf: 26.11.06).

3.2 Historie und State-of-the-Art der Peer-to-Peer-Technologie

Die P2P-Technologie ist keine technologische Innovation. Bei dieser Einschätzung sind sich Experten wie Shirky, Loban, Schollmeier und Oram einig. Der Ursprung der P2P-Technologie liegt bereits in den Anfängen des Internets. Begonnen hat die Entwicklung mit dem ARPANET in den sechziger Jahren, als eine Vernetzung von Universitäten und Forschungseinrichtungen in den USA aufgebaut wurde. Ziel war es damals, zunächst die knappen Rechnerressourcen effizient zu nutzen und den Universitäten einen Austausch von Ressourcen zu ermöglichen (vgl. Minar/Hedlund 2001: 4-5). Das ARPANET sollte möglichst ausfallsicher sein. Verwirklicht wurde diese Anforderung durch ein dezentral konzipiertes Netzwerk in dem alle Rechner gleichrangig und gleichberechtigt über das Telekommunikationsnetzwerk miteinander kommunizieren konnten. Somit vernetzte das ARPANET die Rechner bereits mit einer Peer-to-Peer-Beziehung untereinander. In den 1980er Jahren entstand das Usenet[94], das mit seinen Diskussionsforen ebenfalls P2P-Strukturen aufwies und in der Fachliteratur auch als solches bezeichnet wird (vgl. Minar/Hedlund 2001: 6).

Ende der 1980er Jahre wurde das Internet ursprünglich zur unentgeltlichen Nutzung freigegeben. Bereits Anfang der 1990er Jahre folgte jedoch schnell die Kommerzialisierung des Internets. Der PC (Personal Computer) verbreitete sich massenhaft in Privathaushalten und die ersten Internet Service Provider (ISP) und Onlinedienste entstanden (vgl. Gscheidle/Fisch 2005: 570ff.). Das Internet und die Anzahl der an diesem Netz teilnehmenden Rechner unterlagen einem starken Wachstum.

Als die Teilnehmerzahl der am Internet angeschlossenen Rechner plötzlich stark anstieg, reichte die Anzahl der bis dato verfügbaren, auf einem statistischen Modell basierenden, IP-Adressen nicht mehr für steigenden Anzahl der Netzwerkteilnehmer aus. Bis Mitte der 1990er basierte das Internet ausschließlich auf diesem Modell der Vernetzung, welches die Rechner für eine kontinuierliche Verbindung zum Internet mit einer permanenten statischen IP-Adresse vorsah (vgl. Frascaria 2002: 1ff.).

Durch den Browser konnte ein weiteres innovativeres Modell eingeführt werden. Damit ein User auf Webseiten zugreifen konnte, musste sein Rechner vorher über ein Modem mit einer eigenen IP-Adresse am Netzwerk des Internet angemeldet werden. Die Netzteilnehmer konnten durch den Browser nun individuell das Internet betreten und verlassen, es war keine konstante Verbindung mehr vorgegeben und

---

[94] Das Usenet ist ein weltweites, elektronisches Netzwerk bestehend aus Newsgroups, in denen Diskussionen zu den verschiedensten Themen stattfinden. Die Funktionsweise ist dem Internet ähnlich, allerdings wird anstatt des HTTP-Protokolls ein NNTP-Protokoll zur Kommunikation zwischen den Rechnern der Newsgroups und der News-Server verwendet.

erforderte somit eine dynamische Zuteilung der IP-Adressen durch die ISPs. Durch die erforderliche Anmeldung an einem zentralen Server des ISPs wird dem Rechner bei jeder Anmeldung am Internet eine temporäre IP-Adresse zugewiesen. Die direkte Kommunikation der Rechner untereinander war somit nicht mehr möglich. Allerdings wurde auf diese Weise das Problem der knappen IP-Adressen entschärft (vgl. ebd). Die steigende Zahl der Internetteilnehmer führte also zu einer Veränderung der Verbindungsmodelle. Weitere Ursachen für diese Veränderung waren die Entstehung von Firewalls als Sicherheitskonzept und dem Adresssystem NAT (Network Address Translation), welches ebenfalls wie die Einführung der dynamischen IP-Adressen zum Zwecke einer skalierbaren und sicheren Internetarchitektur entstand. Diese Entwicklungen schwächten aber die Peer-to-Peer-Kommunikationsmodelle. Durch die zentralisierte Internetinfrastruktur wurden die User in ihren Möglichkeiten der Partizipation eingeschränkt. Sie wurden zu reinen Clients degradiert und waren nicht mehr in der Lage, eigene Datenquellen zur Verfügung zu stellen oder Netz-Services anzubieten (vgl. ebd.).

Erst 2001 mit dem Erscheinen der Online-Musiktauschbörse Napster[95] erhielt die P2P-Technologie auch in der Öffentlichkeit erneute Aufmerksamkeit. Ausschlaggebend waren die wiedergewonnene Autonomie der User und die Möglichkeit des gegenseitigen freien Austausches von Inhalten.

Inzwischen haben P2P-Entwickler die Problematik der Firewalls durch technischen Fortschritt gemeistert. Die neue Generation von P2P-Applikationen verfügt über innovative Eigenschaften, die die Infrastruktur des neuen Webs ausnutzen und eine Anwendung der P2P-Technologie trotz Firewalls heute wieder ermöglichen (vgl. Frascaria 2002: 2ff.). Mittlerweile steigt die Anzahl der P2P-Anwendungen und die Nachfrage der User rapide, wie aktuelle Untersuchungen bei den ISPs zeigen (vgl. Cachelogic 2005). Studien von Bigchampage und Cachelogic haben bei der Beobachtung des gesamten Internetdatenverkehrs einen Anteil des P2P-Datenverkehrs von 50 bis 70 Prozent tageszeitenabhängig festgestellt.

Das anhaltende Wachstum des Internets in Bezug auf die Anzahl der User, leistungsfähigere Rechner sowie immer höhere Bandbreiten, werden von steigenden Anforderungen an die Technologie begleitet. Das heutige Standard Client-Server-Modell bedingt einen hohen technischen, finanziellen und personellen Aufwand, um diesen Anforderungen gerecht zu werden (vgl. Steinmetz/Wehrle 2005: 9).

---

95 Weiterführende Informationen: http://www.napster.de (Abruf: 31.11.2006).

Die Nachteile des Client-Server-Modells begünstigen eine Anwendung und Verbreitung von P2P-Applikationen. Der hohe Anteil des Peer-to-Peer Datenverkehrs wird im Folgenden durch Anwendungsbeispiele aufgezeigt. Im Alltag finden bereits eine Vielzahl von P2P-Systemen Einsatz. Um einen Überblick vom State-of-the-Art zu bekommen, werden nachstehend die häufigsten Anwendungsbereiche beispielhaft erläutert.

Filesharing-Systeme wie Freenet und Gnutella sind vermutlich die bekanntesten P2P-Anwendungsbeispiele, gefolgt von den umstrittenen Musiktauschbörsen Napster[96] und eDonkey[97], die folgendermaßen funktionieren: nachdem sich die User eine Software installiert haben, ist es ihnen möglich, durch eine Anmeldung mit Benutzernamen und Passwort nach Songtiteln und Interpreten zu suchen. Aus einer Ergebnisliste kann die gewünschte Datei, sofern sie bei einem anderen im Netz registrierten User zur Verfügung steht, heruntergeladen werden. Sowohl die Software als auch der Download sind bzw. waren kostenlos.

Science-to-Science[98] (S2S) ist ein weiteres Filesharing Projekt vom Deutschen Forschungsnetz. Das Projekt dient der Verbesserung des kooperativen wissenschaftlichen Arbeitens. S2S ist ein Netzwerk von Peers, die die Suche in verteilten Dokument-Beständen des Deutschen Forschungsnetz anbieten.

Des Weiteren werden Filesharing-Systeme als Wissensmanagement-Lösungen in Unternehmen eingesetzt (vgl. Schmücker/Müller 2003: 307-311). Aktuell ist eine Erweiterung für den Firefox-Browser mit dem Namen „AllPeers“ entwickelt und zur Nutzung freigegeben worden.[99] Dieses Tool erlaubt dem User einen Zugriff auf seine persönlichen Kontakte anhand einer erzeugten Buddy-Liste und erkennt so, wer gerade online ist. Per Drag-and-Drop ist ein Austausch von Dateien möglich, auch wenn die Rechner offline sind.

Instant Messaging (IM) ermöglicht den direkten Nachrichtenaustausch in Echtzeit zwischen zwei Usern. Mittels einer vorab installierten P2P-Client-Software ist dem User ersichtlich, welche anderen ihm bekannten IM-User gerade online sind. Der Austausch von kurzen Textnachrichten wird über ein Netzwerk - meist dem Internet, aber auch Intranet - abgewickelt. Der Empfänger kann unmittelbar auf die erhaltene Mitteilung antworten. Jegliche Form von Dateien können gleichermaßen über diesen

---

96 Weiterführende Informationen: http://www.napster.de/ oder http://de.wikipedia.org/wiki/Napster (Abruf: 20.11.2006).

97 eDonkey ist ein dezentrales P2P-Netzwerk und aus diesem Grund nicht abschaltbar.

98 Vgl. http://www.neofonie.de/pm/PM040213.html (Abruf: 07.11.2006).

99 Vgl. http://www.macwelt.de/news/339648/ (Abruf: 06.11.2006).

Weg ausgetauscht werden. Einige Instant-Messaging-Programme bieten zusätzliche Funktionen für Telefon- oder Videokonferenzen an. ICQ (Akronym für „I seek you“), AIM (AOL Instant Messaging), Yahoo Messenger, Skype und Jabber sind die bekanntesten Anwendungen im IM-Bereich.

Grid-Computing[100] ist ein Zusammenschluss bzw. die Bündelung mehrerer Rechner, oder genauer deren Rechenleistung, zu einem Super-Computer mit enormer Rechenleistung. Seti@home (Search for extraterrestrial intelligence at home) - eines der ersten Projekte des Grid-Computing - nutzt ruhende Rechnerressourcen der am Netz beteiligten Rechner zur Suche nach außerirdischer Intelligenz. Auch in der Krebs-[101] und AIDS-Forschung[102] wird Grid-Computing erfolgreich eingesetzt, um Berechnungen die große Rechenkapazitäten benötigen kostengünstig zu ermöglichen (vgl. Mauthe/Heckman 2005: 193ff.).

Zusammenfassend kann gesagt werden, dass Peer-to-Peer-Technologien die ursprüngliche Idee und das ursprüngliche Konzept des Internets wieder aufgreifen (Loban 2004). Ein Konzept, in dem jeder in der Lage ist, sowohl Client als auch Server zu sein und als aktive Person durch interaktive Beteiligung mitwirken kann.

3.3 Peer-to-Peer-Netze unter netzwerk-ökonomischen Gesichtspunkten

Sehr häufig wird an Stelle des Begriffs „Peer-to-Peer-System“ der Begriff „Overlay-Netzwerk“ verwendet. Dies verdeutlicht, dass ein P2P-System ein auf einer Netzwerktopologie aufgebautes, übergeordnetes Netzwerk mit unabhängiger Topologie darstellt, welches das Basisnetz wie Telekommunikationsnetzwerke überlagert (Hauswirth/Dustar 2003: 2). P2P-Applikationen sind somit eindeutig Netzwerkgüter und unterliegen besonderen ökonomischen Gesichtspunkten, die für diese Arbeit von Bedeutung sind und aufgrund dessen im Folgenden kurz beschrieben werden.
Das Internet als Netzwerk und die in diesem Netzwerk angebotenen Netzwerkgüter unterliegen speziellen ökonomischen Verhaltensweisen, bei denen die beteiligten User eine zentrale Rolle einnehmen (vgl. Shapiro/Varian 1999). Unter dem Begriff „Netzökonomie“ (vgl. Zerdick 1999) werden Netze und Ihre Netzwerkeffekte zusammengefasst. Abstrakt gesehen ist ein Netzwerk ein Konstrukt, in dem Objekte auf einem realen oder virtuellen Weg miteinander verbunden sind. In der Literatur wird das Internet als reales Netz bezeichnet (vgl. Linde 2005: 51). Hierbei sind die

---

100 Grid ist die englische Bezeichnung für das Stromnetz. Das Prinzip beim Grid-Computing ist ähnlich. Dabei kommt die Rechenleistung - ähnlich wie der Strom - aus der ’Steckdose’. Weiterführende Informationen: http://gridcafe.web.cern.ch/gridcafe/ (Abruf: 05.10.2006).

101 Weiterführende Informationen: http://www.nfcr.org/default.aspx?tabid=274 (Abruf: 05.10.2006).

102 Weiterführende Informationen: http://fightaidsathome.scripps.edu (Abruf: 05.10.2006).

Objekte die User und die Verbindungen die Kabel und Server. Ein Beispiel für ein **virtuelles Netzwerk** stellen Communities dar, die sich durch die Verwendung gleicher Software oder das Spielen gleicher Spiele bilden. Sie werden als ein virtuelles Netzwerk bezeichnet, da keine direkte physische Verbindung zwischen den einzelnen Teilnehmern besteht. Demzufolge ist ein Netzwerk eine Zusammenfassung von Usern eines bestimmten Gutes. Als Netzwerkgüter werden Güter bezeichnet, deren Existenz oder Verwendung in geringer Anzahl keinen großen Nutzen für die Konsumenten haben. Am Beispiel des Telefons oder Faxgerätes lässt sich dieser Sachverhalt verdeutlichen: Besitzen nur zwei Personen ein Telefon, existiert nur ein sehr kleines Netzwerk und der Nutzen ist gering. Je mehr Telefone und Teilnehmer an das Netz angeschlossen werden, desto größer wird der Nutzen. Das Erreichen einer sogenannten „Kritischen Masse“ an Usern ist notwendig, um die Akzeptanz und die Bereitschaft zur Verwendung des Netzwerkgutes beim User zu erlangen und positive Netzwerkeffekte (Rückkopplungen) zu erzielen. Mit Netzwerkeffekten sind Handlungen gemeint, die Auswirkungen (im positiven oder negativen Sinne) auf das gesamte Netzwerk haben können. In Abb. 13 werden die beiden typischen Lebenszyklen von Netzwerkgütern und deren zeitlichen Verlauf im Verhältnis zur Userzahl dargestellt.

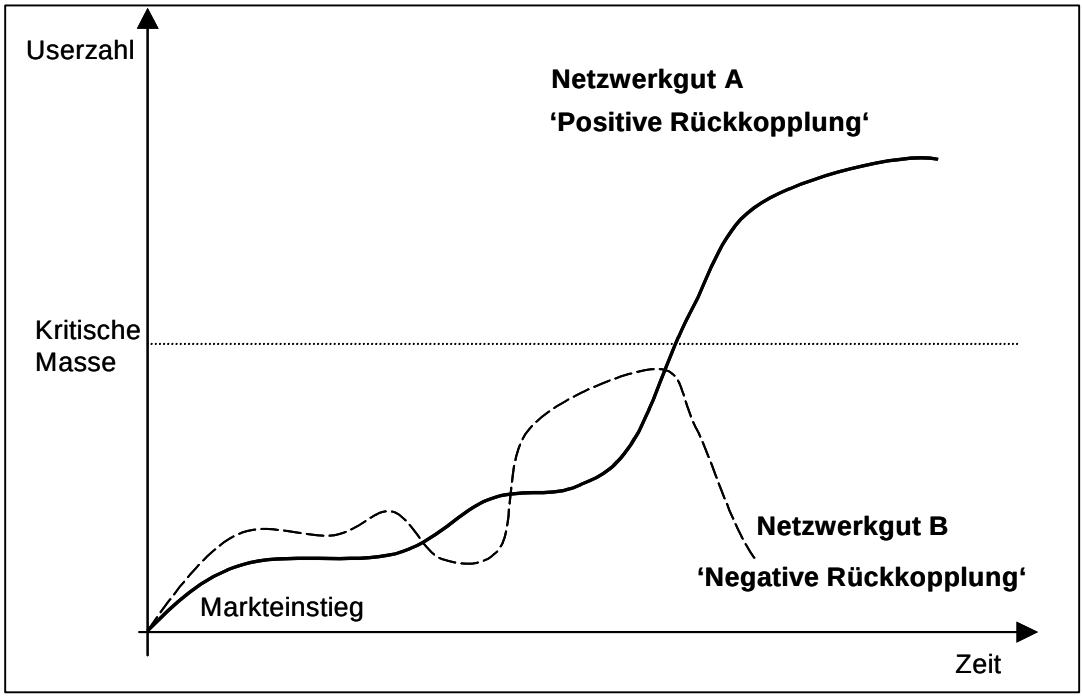


Abb. 13: Lebenszyklus von Netzwerkgütern[103]

[103] Quelle: eigene Darstellung in Anlehnung an Stock 2004 und Dietl/Royer 2000.

Im Falle einer positiven Rückkopplung wird die kritische Masse erreicht das Netzwerkgut A erzielt eine Userzahl, die groß genug ist damit sich das Gut auf dem Markt etabliert. Während Netzwerkgut B die kritische Masse nicht erreicht und es in diesem Fall zum Marktversagen für dieses Gut kommt.

4 Peer-to-Peer-Suchmaschine Faroo

In diesem Kapitel erfolgt eine Beschreibung des Aufbaus und der Funktionsweise von Faroo auf Basis der in den vorherigen Kapiteln erarbeiteten theoretischen Grundlagen. Die einzelnen Untersuchungsbereiche (Suchmaschine und Peer-to-Peer) werden hier zusammengeführt und dem Hauptbereich (Peer-to-Peer-Suchmaschine) gegenübergestellt. Speziell sollen Unterschiede in den einzelnen Arbeitsprozessen einer konventionellen Suchmaschine und einer Peer-to-Peer-Suchmaschine aufgezeigt werden. Als Gerüst und Leitfaden dienen die in Kapitel 2.1 erarbeiteten Komponenten einer Suchmaschine. Vorab eine kurze allgemeine Beschreibung Faroos:

Faroo ist ein Computerprogramm, das eine Universalsuchmaschine für das Web auf Basis der Peer-to-Peer-Technologie realisiert. Dazu muss es vorab von den Usern heruntergeladen und auf dem eigenen Rechner installiert werden. Die Verwendung der P2P-Technologie ermöglicht eine divergente Umsetzung der einzelnen Kernfunktionen einer Suchmaschine, wodurch diese neue Funktionen und Möglichkeiten erhält. Die Faroo-P2P-Web-Suche unterscheidet sich hautsächlich in drei der vier, in Kapitel 2.1 beschriebenen, Komponenten: der Datensammlung (Crawling), der Datenspeicherung im Datenbanksystem und dem Ranking der Suchergebnisse. Für diese Arbeit werden die Datensammlung und -Speicherung fokussiert, während das Ranking der Vollständigkeit halber nur peripher tangiert wird.

4.1 Verteiltes Crawling von Faroo

Die in Kapitel 2.1 beschriebenen Suchmaschinen betreiben die Datensammlung im Internet über einen Crawler, ein Computerprogramm, das selbstständig anhand der Verlinkung von Webseiten untereinander Informationen und Webseiten, die auf den Webservern liegen, auffindet und sammelt (vgl. Glöggler 2003: 25ff.). In diesem Punkt begründet sich der erste signifikante Unterschied zu Faroo: Bei Faroo werden die User an der Datensammlung beteiligt und zusätzlicher Datenverkehr (durch den Crawler) auf den Webservern vermieden. Jeder User ist Teil des Crawlers, der

durch das Aufrufen von Webseiten im Browser die Datensammlung und Aktualisierung des Datenbestandes im Index übernimmt.[104] Der Crawler erhält somit einen dezentralen, verteilten Charakter und ist nicht kontrollierbar, da jeder User ein Teil des Crawlers darstellt. Ausgenommen von der Indexierung sind HTTPS-Seiten[105], Webseiten mit Robots-Meta-Tag[106] direkt im Quellcode der jeweiligen Webseite sowie Webseiten, die Formularfelder zur Eingabe von Benutzernamen und Passwörtern enthalten. Des Weiteren werden alle Informationen bereits vor der Übertragung in den Index über ein Secret-Key-Verfahren[107] verschlüsselt. Somit sind die Informationen sowohl bei der Übertragung als auch im Index selbst zu jedem Zeitpunkt verschlüsselt. Zusätzlich wird die gesamte Kommunikation zwischen den Peers über ein Public-Key-Verfahren verschlüsselt. Die Verschlüsselung beugt Angriffen durch Dritte vor und bietet darüber hinaus folgende Sicherheitseigenschaften:

- Vertraulichkeit: Daten bleiben sowohl beim Indexieren als auch bei Eingabe der Suchanfragen vertraulich.
- Integrität: Daten können während der Übertragung nicht unbemerkt verändert werden. Durch die elektronische Unterschrift können nachträgliche Manipulationen des elektronischen Dokumentes nachgewiesen werden.
- Firewalls können keinen Filter mit Fingerprint[108] für das Faroo-Protokoll setzen.
- Sniffer[109] können die Nachrichten und das Faroo-Protokoll nicht analysieren.

Hier ist anzumerken, dass Links, die sich auf den aufgerufenen Seiten befinden, mit in den Index aufgenommen werden. Jedoch wird nur der Text, mit dem der Link betitelt ist, indexiert. Ausgenommen sind Links, die z.B. nur mit „Klicken sie hier“ gekennzeichnet sind. Im Index wird dann vermerkt, dass für diese Webseite noch keine vollständige Indexierung stattgefunden hat. Eine vollständige Indexierung oder Aktualisierung des Seiteninhaltes erfolgt erst nach Aufruf der Webadresse im

---

104 vgl. http://www.faroo.com (Abruf: 20.12.2006).

105 HTTPS steht für „HyperText Transfer Protocol Secure“ und dient zur Verschlüsselung und zur Authentifizierung der Kommunikation zwischen Web-Server und Browser. Dieses Protokoll findet unter anderem Anwendung beim Online-Banking.

106 Die Robots-Meta-Tags enthalten Anweisungen, die das Indexieren durch einen Crawler untersagen. Weitere Informationen unter: http://www.robotstxt.org/wc/exclusion.html (Abruf: 21.11.2006).

107 Secret-Key- und Public-Key-Verfahren dienen der Ver- und Entschlüsselung von digitalen Daten.

108 Fingerprint (engl. Fingerabdruck) ist ein Hashwert, der unter anderem zur Verschlüsselung von Daten eingesetzt wird.

109 Sniffer (engl. Schnüffler) sind Softwaretools, die Datenpakete innerhalb eines Netzwerkes aufspüren, registrieren und auswerten können.

Browser. Sogenannte „Dead-Links“ werden einmalig aufgerufen und anschließend nach Erkennen der Fehlermeldung 404 (welche für „Webseite nicht mehr unter der URL verfügbar“ verwendet wird) aus dem Index gelöscht.

4.2 Indexierung von Faroo

In der zweiten Komponente, der automatischen Indexierung, unterscheidet sich Faroo nicht wesentlich von den herkömmlichen Suchmaschinen. (Dieser Prozess wird im anschließenden Kapitel dargestellt) Jede im Browser aufgerufene Webseite wird im Cache[110] des Browsers zwischengelagert. Faroo greift auf diesen Cache zu und zerlegt die Webseiten in reine Texte, um die automatische Indexierung durchzuführen. In diesem Schritt werden Stoppworte, wie in der deutschen Sprache „und/oder“, Artikel (der, die, das) etc., eliminiert. Die Webseiten durchlaufen danach eine Volltextindexierung. Jedes Wort wird in den invertierten Index geschrieben und den Webadressen zugeordnet. Um bei der Ausgabe in den Trefferlisten ein Ranking erstellen zu können, werden an dieser Stelle ergänzende Wortstatistiken generiert. Zusätzlich zur Abspeicherung der URL wird das Datum der Aufnahme in den Index angespeichert.

Faroo verwendet zur Erstellung des Index sowohl ein linguistisches als auch ein statistisches Indexierungsverfahren. Das linguistische ist ein Stemmingverfahren (Lemmatisierung), das Worte auf ihre Grundform zurückführt. Dieser Vorgang vereinfacht dem User später das Retrieval. So werden zum Beispiel bei der Sucheeingabe des Begriffs „indexierten“ auch die Begriffe „indexieren“, „indexiertes“ und „indexierte“ gefunden. Auch unregelmäßige Verben werden bei diesem Verfahren auf ihre Stammform zurückgeführt. Stoppworte werden auf Basis einer vordefinierten Liste ausgefiltert. Das statistische Verfahren dient unterer anderem als Grundlage für das Ranking. Hier werden Inverse Dokumentenhäufigkeit (IDF) und Term Frequeny (TF) bestimmt. Es wird zu jedem Wort festgehalten, ob es in der URL, im Titel oder im Content vorkommt. Außerdem wird die Wortposition abgespeichert, wodurch eine Verwendung des Near-Operators möglich wird. Es wird kein weiterer semantischer Zusammenhang zwischen den Worten hergestellt und es findet keine Einordnung in eine Taxonomie statt. Das bedeutet, dass jeder Schreibfehler in einer HTML-Webseite als eigenständiges Wort mit indexiert wird. Aufgrund der Dynamik sowie der Weiterentwicklung in den vielfältigen Sprachen ist dies anders kaum umzusetzen, wenn man alle Sprachen indexieren will. Andernfalls

---

[110] Der Cache ist ein Pufferspeicher oder Zwischenspeicher, der Kopien von anderen Speichern enthält und somit den Zugriff auf die dort befindlichen Daten beschleunigt.

würden Wörter- und Rechtschreibbücher für jede einzelne Sprache benötigt. Außerdem entwickeln sich Sprachen stetig weiter so führen unter anderem Blogger[111] ständig neue Wortschöpfungen in eine Sprache ein. Hierzu sei auf Googlewhack[112] verwiesen.

Die N-Gram-Analyse[113], ist ein Verfahren, um mathematische Gesetze, die auf Vektoren angewandt werden, auch auf andere Objekte anzuwenden. Mit diesem Analyseverfahren können Zusammenhänge gesucht werden, beispielsweise die Wortgruppe „Indexierung von Faroo“ in einer großen Anzahl von E-Mails. Dabei ist die verwendete Sprache nicht von Bedeutung. Die N-Gram-Analyse funktioniert in jeder Sprache und in jedem Alphabet (vgl. Stock 2000: 150).

Faroo erzeugt „auf Wunsch“ einen zweiten Index - und zwar den sogenannten „Desktopindex“. Hier werden alle auf der Festplatte gespeicherten Dokumente einer Volltextindexierung unterzogen. Die indexierten Daten bleiben auf dem eigenen Rechner und werden nicht in den verteilten Index eingespeist. Der Desktopindex kann Vorteile bei der Erstellung des Rankings bringen. Verdeutlicht sei dies an einem Beispiel: Eine Person wohnt in Köln und sucht mit dem Begriff „Pizzeria“ in einer Web-Suchmaschine. Ist ein Desktopindex vorhanden, werden die Ergebnisse von Pizzerien aus Köln für diese Person in den Ergebnislisten höher gerankt. Basierend auf der Annahme, dass eine Person, die in Köln wohnt, mit hoher Wahrscheinlichkeit eine Vielzahl an Dokumenten die den Begriff „Köln“ enthalten - in diesem Fall gleich dem Wohnort - auf ihrem eigenen Rechner befinden. Dieses Verfahren wird als „Personal-Ranking“ bezeichnet.[114] Desktopindices können also eine Beeinflussung der Suchergebnisse von Websuchmaschinen haben. Sie ist jedoch separat abschaltbar. Die Desktopsuche wird in dieser Arbeit nicht weiter fokussiert.

4.3 Verteiltes Datenbanksystem von Faroo

Der nächste grundlegende und bedeutende Unterschied zwischen Faroo und den etablierten Suchmaschinen liegt in der dritten Komponente, der strukturierten Datenspeicherung in einem Datenbanksystem. Die Abspeicherung erfolgt ebenfalls

---

111 Ein Blogger ist der Schriftführer eines Weblogs. Ein Weblog ist ein Kunstwort, das aus den beiden englischen Begriffen Web und Log zusammen gesetzt wurde. Im Internet wird darunter eine Webseite, die regelmäßig neue Einträge erhält, bezeichnet. Die deutsche Übersetzung ist Online-Tagebuch.

112 Weiterführende Informationen: http://www.googlewhack.de (Abruf: 15.12.2006).

113 Das Verfahren wurde 1995 vom amerikanischen Geheimdienst NSA patentiert und seitdem konstant weiter entwickelt.

114 Vgl. http://www.faroo.com (Abruf: 20.01.2007)

automatisch, jedoch im Vergleich zu den herkömmlichen Suchmaschinen in einem **eigenen dezentralen Datenbanksystem**. Dieses wird durch die P2P-Technologie realisiert. Jeder User, der Faroo benutzten möchte, muss vorab eine kostenlose Software herunterladen und auf dem jeweiligen Rechner installieren. Diese Software ermöglicht eine Kommunikation aller Rechner, die diese Software ebenfalls installiert haben. Die Rechner der User bilden die Infrastrukturlieferanten des Faroo-Netzes. Die Rechnerressourcen dienen zur Abspeicherung von Teilen des Suchmaschinenindex (Zustimmung des jeweiligen Users vorausgesetzt). Jeder User kann selbst entscheiden, ob und wie viel Speicherplatz er dem Faroo P2P-Netz zur Verfügung stellt.

Das Grundprinzip des Faroo-Index basiert auf der in Kapitel 3.1.2 dargestellten strukturierten, DHT basierten P2P-Architektur. Strukturiert bedeutet, dass durch die Verwendung von DHT jedem aktiven Online-Peer ein bestimmter Zuständigkeitsbereich (des Alphabets, Zeichen, Zahlen) zugeteilt wird. Aus Abb. 14 lässt sich erkennen, dass Peer 1 in dem beispielhaft dargestellten P2P-Netz einen Zuständigkeitsbereich für die Begriffe von „Anker“ bis „Auto“ hat. Alle Webseiten, die diesen Begriff enthalten, werden in den invertierten Index dieses Peers geschrieben.

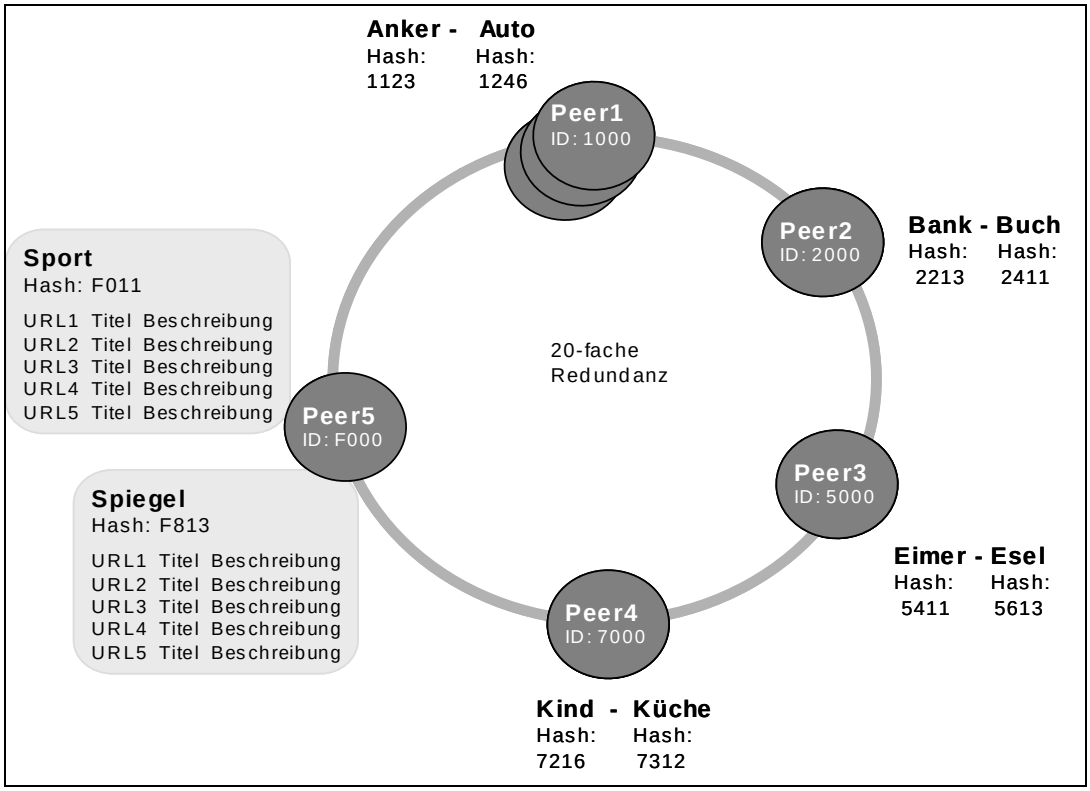


Abb. 14: Distributed Hash Table (DHT) und Distributed Inverted Index[115]

[115] Quelle: eigene Darstellung in Anlehnung an: vgl. Steinmetz/Wehrle 2005: 84

Diese Strukturierung des P2P-Netzes ermöglicht einen schnelleren und zielgerichteteren Zugriff auf den Index. Ein Vergleich der Hashwerte[116] bringt die Suchanfrage zu dem zuständigen Peer, ohne das gesamte P2P-Netz mit der Suchanfrage zu belasten. Folglich wird die Netzwerklast gering gehalten und die Antwortzeit verkürzt. Eine 20-fach erzeugte Redundanz des gesamten verteilten Index gewährleistet eine konstante und vollständige Verfügbarkeit des Index, erhöht die Performance bei einer Parallelsuche, balanciert die Belastung der einzelnen Peers aus und hält diese gering. Hash-Tabellen organisieren sich selbst. Peers werden mit einem Hashwert versehen (ge-'hash'-t), sobald sie das erste Mal in das Faroo-Netz eintreten. Die in den Index aufgenommenen Daten werden ebenfalls mit einem Hashwert versehen und sind somit verschlüsselt. Jeder Peer ist für einen bestimmten Bereich verantwortlich. Tritt ein neuer Peer in das Netzwerk ein, wird der Zuständigkeitsbereich der benachbarten Peers neu aufgeteilt. In dem Fall, dass ein Peer das Netz verlässt, fangen die Nachbarn den jeweiligen Zuständigkeitsbereich wieder auf. Mit den DHT wird die hohe Dynamik, die es in P2P-Netzen zu meistern gilt, erfolgreich bestritten. Jeder Peer besitzt vollkommene Autonomie.

Faroo erfüllt also alle drei der in Kapitel 3.1.1 dargestellten Eigenschaften. Die Suchmaschine ist vollkommen dezentral konzipiert. Zum einen, da sie keine zentrale Servereinheit zur Abspeicherung des Index besitzt und zum anderen, da jeder User Teil des verteilten Index, des verteilten Crawlers sowie Teil des verteilten Rankings ist. Eine zentrale Kontrolle des Index ist nicht gegeben, auf den Crawler kann kein Einfluss genommen werden.

4.4 Verteiltes Ranking von Faroo

Die Benutzerschnittstelle von Faroo wird, ähnlich wie bei den etablierten Suchmaschinen, in Form einer Webseite umgesetzt. Für das Ranking wurde ein neuartiges Verfahren entwickelt, welches erst auf Basis der P2P-Technologie ein verteiltes Ranking (das sogenannte „PeerRank") ermöglicht.[117] Die Faroo-Software zieht das Userverhalten (vgl. Garbe 2001: 515) beim Betrachten der Seite zur Bewertung heran. Dies erfolgt automatisch, ohne dass der User eine explizite Bewertung abgeben muss. Damit entscheidet der User, für den die Seite bestimmt ist, auch über deren Bewertung. Bei bisherigen Verfahren entscheiden nur die Webseitenbetreiber durch die Verlinkung zwischen den Seiten über deren Bewertung.

---

[116] Hashwerte sind Prüfsummen die mittels Hashfunktionen zur Verschlüsselung einer Nachricht angewendet werden. Die Verschlüsselungsfunktion soll so angelegt sein, dass jeder Hashwert einzigartig ist und wird somit häufig als digitaler Fingerprint bezeichnet.

[117] Vgl. http://www.faroo.com (Abruf: 20.11.2006).

Durch PeerRank wird das Ranking auf eine breitere Basis gestellt und damit demokratisiert.[118] Erstmalig entscheiden die User selbst darüber, welche Ergebnisse für sie am wichtigsten sind. Dieses Verfahren ist somit eine Form des Social-Search, vermeidet aber dessen bisherige Nachteile. So wird nicht nur der Bekanntenkreis in die Bewertung einbezogen, sondern alle Faroo-User. Es ist keine Anmeldung nötig und es werden keine Userprofile und keine Suchhistorien auf einem zentralen Server gespeichert. Bisherige Social-Search-Verfahren gehen von der Annahme aus, dass sich das Interessens- und Suchprofil von Bekannten ähnelt, bzw. dass man zu jeder Frage Experten im Bekanntenkreis hat.[119]

Zusätzlich zum Social-Ranking wird dem User auf Wunsch die Möglichkeit zu einem personalisiertem Ranking (siehe Kapitel 4.2) angeboten. Die personalisierte Seitenbewertung basiert auf den Interessensgebieten des suchenden Users selbst. Um den Interessensschwerpunkt zu ermitteln, werden neben den besuchten Webseiten auch die Inhalte lokaler Dokumente des Users ausgewertet. Wenn ein User z.B. nach dem Begriff „Auto“ sucht und ein PDF-Prospekt von Audi auf seinem Desktop liegt, werden die Ergebnisse für den Begriff „Auto“, in denen ebenfalls die Bezeichnung „Audi“ vorkommt, höher gerankt. Die Auswertung erfolgt nur auf dem Rechner des jeweiligen Users. Diese Informationen werden zu keinem Zeitpunkt nach außen gesendet. Der User kann die Personalisierung jederzeit deaktivieren. Gegenwärtig ist jedoch noch kein Ranking aktiviert[120]. Dieses neuartige Ranking-Verfahren wird in dieser Arbeit nicht weiter fokussiert, da die genauere Betrachtung den Umfang einer weiteren wissenschaftlichen Arbeit entsprechen bzw. sogar übersteigen würde.

4.5 Netzwerk Faroo

Wie in Kapitel 3.3 beschrieben, sind P2P-Technologien Netzwerkgüter und unterliegen ökonomischen Besonderheiten. Eine besondere Eigenschaft die diese Netzwerkgüter haben ist die Abhängigkeit der sicheren Existenz von der Anzahl der User. Eine sichere Existenz ist nur mit einer Vielzahl von Usern gegeben, weil sich erst dann der optimale Nutzen des Netzes entfaltet. Denn mit jedem hinzukommendem User steigt der Wert des Netzes exponentiell.

Faroo ist sowohl vom technischen als auch virtuellen Standpunkt ein Netzwerkgut. Aus technischer Sicht wird ein reales, physisches auf Leitungen basierendes Netz-

---

118 Vgl. http://www.faroo.com (Abruf: 20.12.2006).
119 Vgl. http://www.faroo.com (Abruf: 20.12.2006).
120 Vgl. http://www.faroo.de (Abruf: 20.12.2006).

werk aufgebaut. Die Rechner der Faroo-User bilden die Infrastruktur der Suchmaschine selbst und werden zu einem Teil einer weltweit verteilten Suchmaschine. Der User bestimmt durch sein Surfverhalten, welche Inhalte in den Index einfließen und somit als relevant gelten. Der virtuelle Aspekt ergibt sich aus der entstehenden Community der Faroo-User, die das Ziel verfolgen, qualitativ hochwertige und aktuelle Suchergebnisse zu generieren und zu erhalten. Das Netzwerk wird nicht von einer zentralen Einheit geleitet und verändert, sondern stellt einen freien Zugang zu ungefilterten, objektiven Informationen sicher. Durch die Dezentralität gehört die Suchmaschine, bildlich gesprochen, den Usern.

Aus dieser Eigenschaft lässt sich folgende Aussage ableiten: Das P2P-Prinzip ist gewissermaßen das "Prosumer" - Prinzip auf Rechner übertragen. Prosumer ist ein Kunstwort und setzt sich aus den Begriffen „Poducer“ und „Consumer“ zusammen. Es beschreibt den Zustand, in dem ein User sowohl Produzent als auch Konsument eines Produktes oder einer Dienstleistung ist. Folglich kann ohne Netzwerkteilnehmer Faroo als P2P-Netz nicht existieren.

5 Wachstum des Index einer Peer-to-Peer-Web-Suchmaschine

Die vorliegende Arbeit beschäftigt sich mit dem Wachstum des Index. Im Folgenden Kapitel wird der Begriff „Index“ mit besonderem Bezug auf eine Suchmaschine erläutert, um im Anschluss das Wachstum des Index zu beleuchten und Überlegungen zu dessen Verlauf zu erarbeiten.

5.1 Eingrenzung und Definition des Begriffs „Suchmaschinenindex“

Allgemein wird der Begriff „Index“ in der deutschen Sprache vielseitig verwendet und erlangt verschiedene Bedeutungen. Ursprünglich wurde eine Liste von Büchern, die nach päpstlichem Entscheid von den Gläubigen nicht gelesen werden durfte, unter einem Index verstanden (vgl. Fremdwörterbuch 2003). Heute wird zumeist eine Liste von nicht jugendfreien Filmen mit dem Begriff assoziiert. In der vorliegenden Arbeit geht es jedoch vielmehr um die Bedeutung des Begriffs in der Informatik und Informationswissenschaft.

Der Ursprung der digitalen Indices sind analoge Verzeichnisse[121], wie zum Beispiel Namens- und Sachverzeichnisse oder Stichwortlisten. In Bibliotheken werden Standorte von Büchern anhand von Katalogen mit Sach- und Autorenverzeichnissen aufgefunden. In einem Buch werden bestimmte thematische Kapitel anhand des

[121] Die Begriffe Index und Verzeichnis werden fortan synonym verwendet.

Inhaltsverzeichnisses oder des Sachverzeichnisses recherchiert. Ein Verzeichnis enthält demnach eine alphabetisch oder inhaltlich geordnete Zusammenstellung - in Form einer Liste, Tabelle oder Matrix - von Namen, Begriffen oder inhaltlichen Merkmalen eines Dokumentes, die auf eine bestimmte Seitenzahl oder einen Standort verweist und die Wiederauffindbarkeit bestimmter Themen und Inhalte erzielt.

Im untersuchten Fachgebiet wird der Index als eine Datenstruktur zum schnellen Auffinden von Datensätzen in Datenbanken verstanden. Im Kontext von Datenbanken wird die chronologische oder alphabetische Auflistung von Suchwörtern mit Angaben zur Worthäufigkeit, Position und Verweisen zu den Dokumenten[122] in denen die Suchwörter enthalten sind, als Index bezeichnet (vgl. Strauch/Kuhlen/ Laisiepen 2004a: 106).

Um die enorme Größe eines Suchmaschinenindex zu verdeutlichen, sei hier erwähnt, dass Suchmaschinen versuchen, das gesamte WWW in einem zentralen Index zusammenzufassen. Die Ausmaße der im Web verfügbaren Daten wurden in Kapitel 9 bereits erläutert. Diese Datenmengen erfordern Datenbanken mit beachtlicher Speicherkapazität. Somit bestehen diese Datenbanken aus mehreren hundert Servern (vgl. Gilder 2006), die die automatisch gesammelten Webadressen abspeichern. Um diese ihres Inhaltes entsprechend wiederauffindbar zu machen, wird der Webseitentext welcher zumeist aus HTML-Code besteht zerlegt und die bedeutungstragenden Wörter in den Suchmaschinenindex, zugehörig zur URL, abgespeichert. Über den Index wird also eine Verknüpfung zwischen Worten und den Dokumenten, in denen sie enthalten sind, hergestellt. Die gebräuchlichste und effizienteste Form ist ein „Invertierter Index“ (vgl. Baeza-Yates/Ribeiro-Neto 1999: 383). Dieser besteht aus einer mehrdimensionalen Tabelle, in der jedes Wort, das in mindestens einem Dokument vorkommt, eine eigene Zeile enthält. Dem Wort werden Angaben zur Position des Wortes im Dokument (in der URL oder im Titel) und Angaben über die Häufigkeit, mit der das Wort im Dokument auftritt im Index zugewiesen. Die URL dient als Verweis[123] und Weiterleitung auf das Dokument, in dem das entsprechende Wort enthalten ist. Abb. 15 zeigt ein einfaches Beispiel eines Invertierten Index:

---

[122] Dokumente sind Webseiten und dort eingebundene Texte, Bilder, Video- und Audiodateien.
[123] Dieser Verweis wird als Link in der Suchmaschinenergebnisseite dargestellt.

1	6	9	11	17	19	24	28	33	40	46	50	55	60
This	is	a	text.	A	text	has	many	words.	Words	are	made	from	letters.

Vokabular	Position	Häufigkeit	URL
letters	60	1	http://www.far...
made	50	1	
many	28	1	
text	11+19	2	
words	33+40	2	

Abb. 15: Beispiel eines invertierten Index[124]

Die Angaben im invertierten Index lassen eine Errechnung der Abstände der Worte untereinander zu und ermöglichen beim Suchprozess die Verwendung bestimmter Recherche-Operatoren, wie zum Beispiel dem NEAR-Operator. Wortposition und -Häufigkeit sind Einflussfaktoren, die später Bedeutung für die Berechnung der Rankings erlangen. Ebenso werden Angaben zur Begriffshäufigkeit innerhalb der gesamten Datenbank sowie die Anzahl der Dokumente in denen der Begriff mindestens einmal vorkommt, verzeichnet.

An dieser Stelle wird die Bedeutung des Index deutlich. Der Index ist das Herzstück der Suchmaschine. Um die Suchanfragen der User befriedigend beantworten zu können, wird ein Index mit einer umfangreichen Menge an indexierten Webseiten und Worten benötigt. Nur indexierte Daten sind wieder auffindbar. Ist der Index klein oder umfasst er nur den Wortschatz eines Spezialgebietes, können die Suchanfragen nicht ausreichend beantwortet werden, was zur **Unzufriedenheit der User** führen kann. Hier ist es wichtig zu erkennen, dass die Ordnung und das System, mit der die Datenabspeicherung im Index erfolgt, ausschlaggebend für die Qualität der Suchergebnisse beim Retrieval ist und nicht allein die Quantität der Webseiten. Die Informationswirtschaft unterscheidet diesbezüglich zwischen „Recall“[125] (Vollständigkeit) und „Precision“[126] (Genauigkeit) (vgl. Stock2000: 121ff.).

Ein Suchmaschinenindex, der das Internet vollständig indexieren könnte, würde eine komplette Kopie aller Informationen, des WWWs erzeugen und in der zentralisierten Datenbank abspeichern. Eine wichtige Frage aus Sicht der Suchmaschinenbetreiber ist daher, wie der Datenbestand möglichst effizient und kosten-

---

124 Quelle: Baeza-Yates/Ribeiro-Neto 1999: 383.

125 Der „Recall“ ist nicht messbar, da die Anzahl aller relevanten Dokumente in einer Datenbank nicht bekannt ist. Der folgende Satz verdeutlicht die Sachlage: „Woher weiß ich, was ich nicht gefunden habe?“ (vgl. Stock 2000:122).

126 Die „Precision“ ist messbar. Sie ist der Anteil der relevanten Dokumente relativ zur Gesamtheit der gefundenen Dokumente (vgl. Stock 2000: 128).

günstig erzeugt und aktuell gehalten werden kann. Die Anforderung nach Aktualität und möglichst umfangreicher Abdeckung der im WWW verfügbaren Informationen, ist nur durch ein automatisiertes Indexierungsverfahren zu realisieren.

5.1.1 Indexierungsprozess

Der Prozess zur Erstellung eines Index wird im deutschen Sprachgebrauch als „Indizierung“ bezeichnet. Jedoch wird unter Indizierung im Allgemeinen vielmehr die Aufnahme von nicht jugendfreien Filmen in eine entsprechende Liste verstanden. Im Zusammenhang mit der Erstellung eines Datenbankindexes wird meist der aus dem Englischen übertragene Begriff „Indexierung“ verwendet. Für den weiteren Verlauf dieser Arbeit wird unter dem Begriff „Indexierung“ das Erstellen eines Datenbank-index verstanden.

Der Prozess der Indexierung, speziell die Bearbeitung des Dokumentes (in Bezug auf Suchmaschinen sind Dokumente in erster Linie Webseiten), setzt erst nach dessen Auffinden durch das Crawlersystem ein und wird grob in fünf Hauptarbeits-schritte gegliedert (Baeza-Yates/Ribeiro-Neto 1999: 165ff.).

Im **ersten Schritt** wird das Dokument einer lexikalischen Analyse unterzogen. Ziel ist es, alle Formatierungen wie Kennziffern, Binde- und Trennstriche, Satzzeichen und spezielle HTML-Bezeichnungen und Formatierungen aus dem Dokument zu entfernen. Im **zweiten Schritt** wird eine Worterkennung durchgeführt. Es werden Stoppwörter, Nomen, Verben und anderen Worttypen in dem Dokument erkannt und unterschieden. Stoppwörter können auf Basis statistischer (automatisch generiert durch Worthäufigkeiten) oder linguistischer Verfahren (vorab definierte Liste mit Stoppwörtern) selektiert werden. Unter Stoppwörtern werden Wörter subsummiert, die in einer natürlichen Sprache nicht bedeutungstragend sind, wie zum Beispiel Artikel, Präpositionen und Konjugationen. Aufgrund dessen können diese in Schritt zwei des Indexierungsprozesses ohne Bedenken bei der Indexierung ignoriert werden. Bei der Betrachtung einer Liste von Stoppwörtern wird schnell ersichtlich, dass sie die am häufigsten verwendeten Worte einer Sprache sind (siehe Abb. 20). Somit bedarf fast die Hälfte der Worte eines Textes bei der Indexierung erst in Schritt fünf näherer Betrachtung, wodurch eine signifikante Reduzierung an Platz in einem Index möglich wird (vgl. Baeza-Yates/Ribeiro-Neto 1999: 146ff.). (Beim Retrieval werden sie bei einer Phrasensuche jedoch wieder notwendig.)

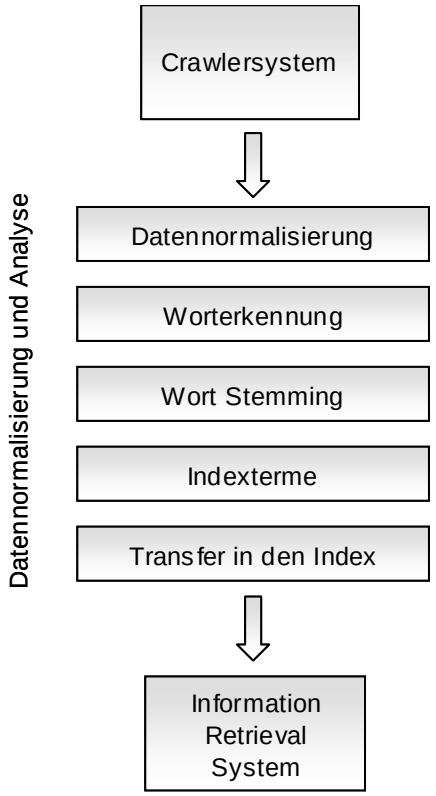


Abb. 16: Flussdiagramm eines Indexierungsprozesses[127]

Die Worte, die den gefilterten, bedeutungstragenden Rest bilden, werden im **dritten Schritt** einem Stemmingverfahren (Lemmatisierung) unterzogen. Dies bedeutet, dass Wörter auf ihre Grundform zurück gebracht werden, zum Beispiel: indexieren: indexiert, indexierte, indexierten. Darunter fällt das Entfernen von Flexionen und Affixen ebenso wie die Dekomposition von Komposita. Durch diesen Schritt wird das Retrieval-Ergebnis größer und das Formulieren der Suchanfragen einfacher. Wird nach dem Begriff „indexierte“ oder „indexiertem“ gesucht, wird das Dokument, welches „indexiert“ enthält, ebenfalls aufgefunden. Der **vierte Schritt** gilt der Auswahl von Wörtern oder Wortgruppen, die als Indexterme in Frage kommen. Gewöhnlich ist die syntaktische Eigenschaft eines Wortes von Bedeutung. Ein Nomen hat in den meisten Fällen einen beschreibungsfähigeren Charakter als ein Adjektiv, ein Verb oder ein Adverb. **Schritt fünf** ist für die Transformation der gerade extrahierten Indexterme in den jeweiligen Index zuständig. Hier werden zusätzliche Informationen wie Worthäufigkeit, Position des Wortes im Titel, in der URL oder im Text sowie das Datum der Indexierung erzeugt und dem Wort zugeordnet. Suchmaschinen verwirklichen die Generierung eines Index durch rein automatisierte Verfahren.

---

[127] Quelle: eigene Darstellung in Anlehnung an: Baeza-Yates/Ribeiro-Neto 1999: 150.

5.1.2 Größe des WWWs im Vergleich zur Größe der Suchmaschinenindices

Sowohl die Größe des WWWs als auch die Größe der Suchmaschinenindices beruhen auf Schätzungen und sind aufgrund der exponentiell wachsenden Datenmenge nicht eindeutig bestimmbar. Dennoch ist ein Vergleich zwischen der geschätzten Größe des WWWs und der Größe der Suchmaschinenindices interessant. Es zeigt sich deutlich, dass Suchmaschinen nicht in der Lage sind, das gesamte WWW abzubilden. Die Universität Bielefeld illustriert die Abdeckung der Webseiten durch die Suchmaschinen anschaulich mit der folgenden Abbildung:

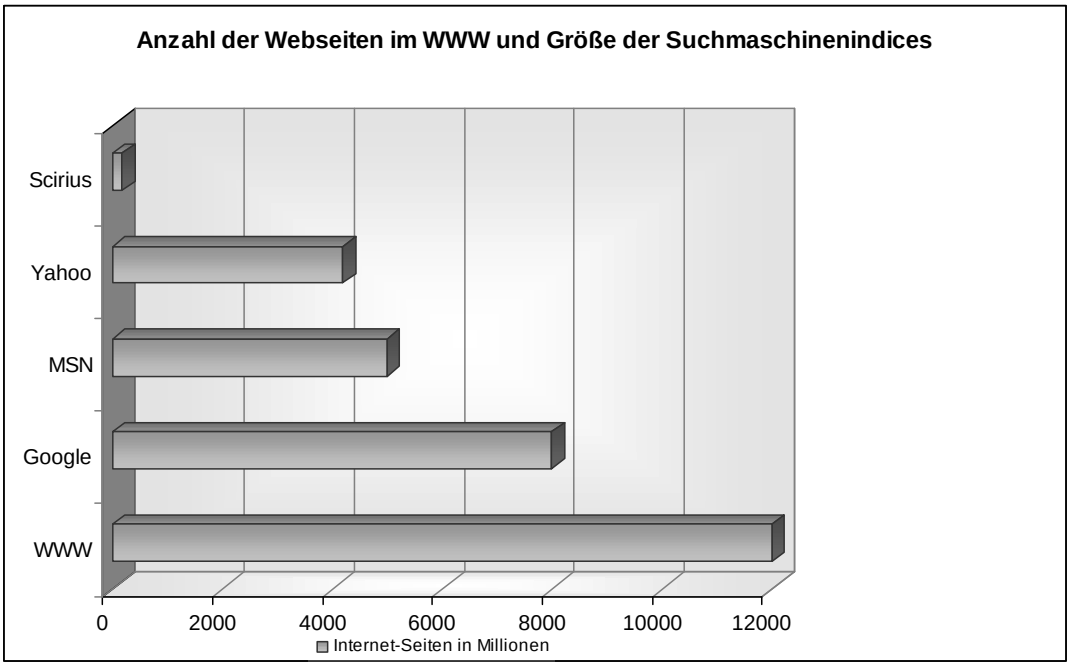


Abb. 17: Größe des WWWs im Vergleich zur Größe der Suchmaschinenindices[128]

Nach eigenen Angaben verfügt Google 2005 mit insgesamt ca. 8 Milliarden Seiten über den größten Index, gefolgt von der MSN-Suche mit ca. 5 Milliarden Seiten. Die Größe des Yahoo-Index wird auf 4,2 Milliarden Seiten geschätzt. Um einen Vergleich zu einer themenspezifischen Suchmaschinen-Datenbank zu bekommen, sei hier Scirus erwähnt. Scirus ist die größte Suchmaschine speziell für wissenschaftliche Informationen, die laut eigenen Angaben eine Indexgröße von 167 Mio. Seiten hat.[129] Die von der Universität Bielefeld gewonnen Daten werden durch eine Untersuchung von Gulli und Signorini[130] 2005 zur Größe des indexierbaren und bereits von den einzelnen Suchmaschinen indexierten Webs bestätigt. Die

[128] Vgl. http://www.ub.uni-bielefeld.de/biblio/search/services/ (Abruf: 23.10.2006).
[129] Vgl. http://www.ub.uni-bielefeld.de/biblio/search/services/ (Abruf: 23.10.2006).
[130] Vgl. http://www.cs.uiowa.edu/~asignori/web-size/ (Abruf: 23.10.2006).

Ergebnisse wurden von Sullivan[131] in der folgenden Tabelle 6 übersichtlich zusammengefasst:

Suchmaschine	Eigene Angaben Größe in Milliarden	Geschätzte Größe in Milliarden
Google	8.1	8.0
Yahoo	4.2	6.6
Ask	2.5	5.3
MSN-Suche (beta)	5.0	5.1
Gesamte Web	----	11.5

Tabelle 6:Suchmaschinenindexgrößen[132]

Allerdings entstehen auch Fragen im Zusammenhang mit der Indexgröße wie zum Beispiel welche Dateien (Formate) indexiert werden.

Im September 2005 entfernte Google die Größenangabe seines Index auf der Homepage mit dem Argument, dass die Indexgröße einer Suchmaschine nichts über die Relevanz der Ergebnisse aussagt. Seitdem gibt es nur noch Schätzungen über die tatsächliche Indexgröße, die auf wissenschaftlich durchgeführten Analysen (Retrievaltests) beruhen. In einer aktuellen Studie von Bar-Yossef und Gurevich (2006) wird das Verhältnis der oben dargestellten Indexgrößen umgekehrt. Die Indexgröße kann aufgrund der verwendeten Erhebungsmethode (Zufallsstichprobe) nur in relativen Zahlen angegeben werden. Demnach verfügt Yahoo über 28 Prozent mehr indexierte Webseiten als Google indexiert und MSN-Suche über 27 Prozent weniger als Google (vgl. Bar-Yossef/Gurevich 2006: 49).

Für diese Arbeit weitaus interessanter sind jedoch die Angaben zur Aktualität der indexierten Webseiten, die auf der gemessenen Anzahl von „Dead-Links“ basieren.

131 Vgl. http://searchenginewatch.com/showPage.html?page=2156481 (Abruf: 23.10.2006).
132 Vgl. http://blog.searchenginewatch.com/blog/050517-075657 (Abruf: 24.10.2006).

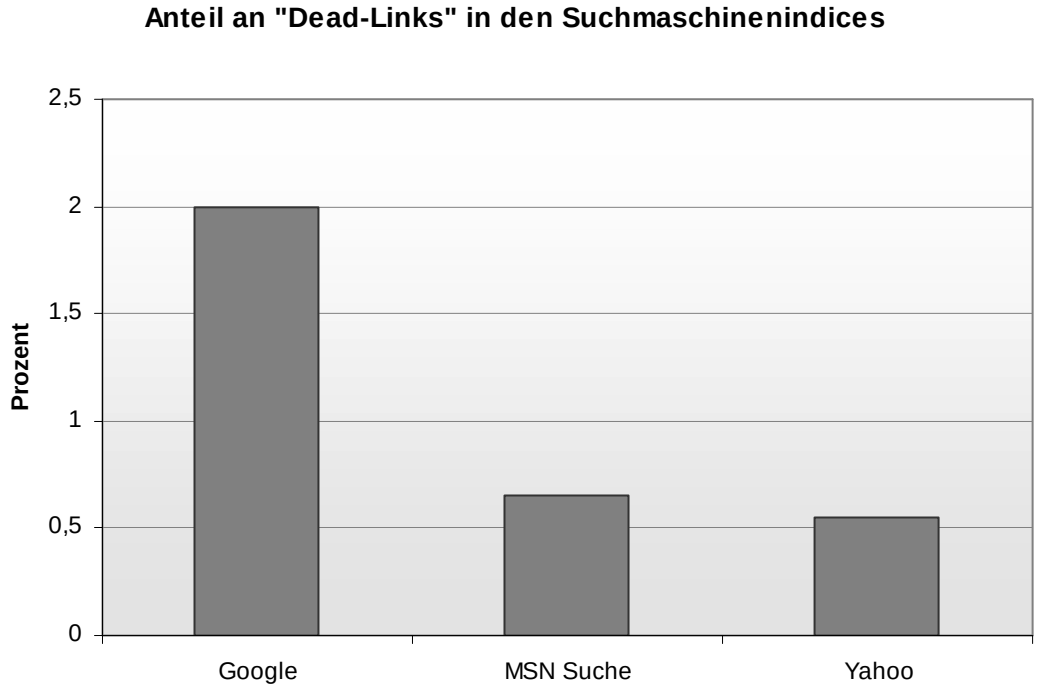


Abb. 18: Anteil an „Dead-Links“ in den Suchmaschinenindices[133]

Der Google-Index weist dabei den höchsten Anteil an Dead-Links (2,0 Prozent) auf, mit Abstand folgt die MSN-Suche (ca. 0,65 Prozent) und schließlich Yahoo (ca. 0,55 Prozent) mit dem geringsten Anteil an Dead-Links (vgl. Bar-Yossef/Gurevich 2006: 51). Dies liegt unter anderem an der teilweise manuellen Indexierung der Webseiten. Aus diesen Untersuchungen wird ersichtlich, dass die Aktualität der Suchmaschinenindices eine Herausforderung für die Betreiber darstellt.

5.2 Wachstum des Index

Wachstum ist der Anstieg einer Messgröße über einen bestimmten Zeitraum hinweg. Zu jedem Zeitpunkt kann der Messgröße ein bestimmter Wert zugeordnet oder abgelesen werden.[134] Da der Faktor Zeit eine stetige Eigenschaft besitzt, lässt sich das Wachstum in einem Diagramm als sogenannte „Wachstumskurve“ darstellen.

Die Anzahl der in einem invertierten Index enthaltenen Webadressen wird als Indexgröße bezeichnet (vgl. Lewandowski 2005). Die Messgröße des Wachstums eines Suchmaschinenindex ist somit der Anstieg einer Anzahl aufgenommener Webseiten (und Worte) während eines bestimmten Zeitraumes. Ob das Wachstum

---

133 Vgl. Bar-Yossef/Gurevich 2006: 51.
134 Vgl. http://de.wikipedia.org/wiki/Wachstum (Abruf: 20.12.2006).

endlich ist oder nicht, ist abhängig von den jeweiligen Eigenschaften der betroffenen Messgröße. Im Folgenden werden die Messgrößen des Indexwachstums auf ihre Wachstumseigenschaften hin überprüft.

Zuerst wird hierbei der Wortschatz betrachtet, da Wörter als Hinweis auf eine URL fungieren. Die Anzahl der Wörter im Index wird solange steigen, bis alle Wörter aus jeder im Web existierenden Sprache indexiert sind. Die Anzahl an Wörtern, allein in der deutschen Sprache, ist jedoch nicht eindeutig bestimmbar.[135] Sprachen und ihr Vokabular sind dynamisch und unterliegen einem ständigen Wandel.[136] Es kommen neue Wortbildungen hinzu, dafür fallen ausgediente weg bzw. werden in der heutigen Sprache nicht mehr verwendet. Auch durch das Aufkommen und die schnelle Verbreitung des Internets sowie die Schnelllebigkeit von Produkten und der Entwicklungen von Software, ist eine ständige Anpassung und Ergänzung des Vokabulars notwendig.[137] Am Institut für deutsche Sprache der Universität Tübingen beobachten Sprachwissenschaftler die Entwicklung sowie Wortneubildungen (Neologismen) in der deutschen Sprache. Beispielhaft können Worte wie „Weblog“, „Podcast“, „Mobbing“, „Surfen“ (Neubedeutung) und „Onlinebanking“ als Neologismen aufgelistet werden.

Als nächstes wird die eigentliche Messgröße - die Anzahl der Webseiten im Internet - betrachtet, welche sich, wie in Kapitel 1.2 dargestellt, exponentiell verhält. Durchschnittlich kommen am Tag mehrere hundert oder sogar mehrere tausend Webseiten hinzu (vgl. Kapitel 1.2). Einen großen Anteil der neu hinzukommenden Seiten bilden die sich immer mehr etablierenden Online-Zeitungen, durch die konstant tagesaktuellen Nachrichten und Berichte ins Internet gestellt werden. Erneut kann hier auch die Datenflut der Blogger angeführt werden. Daraus lässt sich schlussfolgern, dass das Wachstum eines Suchmaschinenindex auch langfristig nicht zum Stillstand kommt.[138] In diesem Zusammenhang sei auf die Archivierungsproblematik von digitalen Informationen hingewiesen.

---

135 Vgl. http://www.duden.de/index2.html?deutsche_sprache/zumthema/weg_eines_wortes.html (Abruf: 05.10.2006).

136 Vgl. http://www.gfds.de/ (Abruf: 10.10.2006).

137 Eine gelungene Kunstinstallation mit dem Namen „bit.fall“ von Julius Popp verdeutlicht die Informationsflut im Internet, die Schnelllebigkeit der gerade im Internet aktuellen und neu auftauchende Worte, Informationen und News sowie die Veränderung von Informationsbedürfnissen mit der Zeit. Die Installation ist ein Wasserfall der Buchstaben und Worte schreiben kann. Mithilfe eines Computerprogramms werden nach statistischen Regeln aktuelle Schlagworte von verschiedenen Nachrichtenwebseiten des Internets rausgefiltert. Diese liefern damit den Input für die Installation des Wasserfalls. online unter: http://www.youtube.com/watch?v=AICq53U3dl8 (Abruf:20.01.2007).

138 Vgl. http://www.wortwarte.de (Abruf: 20.01.2007).

Die Analyse des Crawlingverfahrens von Faroo macht unter diesen Gesichtspunkten deutlich, dass Webseiten nicht nur von einer Person besucht werden sondern dass der Besuch von Webseiten redundant erfolgt. Das Gesetz **„Heaps Law“**, das aus der Linguistik stammt, jedoch auch auf andere Bereiche anwendbar ist, eliminiert diese Redundanz. Es beschreibt das Wachstum des Vokabulars (Anzahl der einmal und eindeutigen, eigenständigen Wörter) als Funktion von der Textgröße (Anzahl aller vorkommenden Wörter eines Textes) (vgl. Baeza-Yates/Ribeiro-Neto 1999: 147,371ff.). Dieses empirische Gesetz besagt weiter, dass mit zunehmender Anzahl an Text (Textdokumenten) - aus einer bestimmten Sprache oder einem Wortschatz - die Entdeckung neuer eigenständiger Wörter dieses Wortschatzes abnimmt (vgl ebd). Das Verhältnis zwischen Textgröße (Anzahl aller in diesem Text stehenden Wörter) und der Anzahl der einmal vorkommenden Wörter in diesem Text wird mit folgender Funktion (siehe Abb. 19) beschrieben:

$$V_R(n) = Kn^{ß}$$

V_R ist die Anzahl der eigenständigen Worte eines Textes oder einer Textsammlung mit einer Gesamtanzahl von *n* Worten. *K* und *ß* repräsentieren freie Parameter, die auf Basis empirisch erhobener Daten bestimmt werden. Für englische Texte liegen die Werte der beiden Parameter normalerweise zwischen 10 und 100 für *K* und zwischen 0.4 und 0.6 für *ß*. Demnach beschreibt Heaps Law eine anfangs stark steigende Kurve, die mit zunehmender Textmenge abflacht, aber aufgrund von Wort- und Webseitendynamik nie aufhört zu steigen.[139]

[139] Vgl. http://planetmath.org/encyclopedia/HeapsLaw.html (Abruf: 20.10.2006).

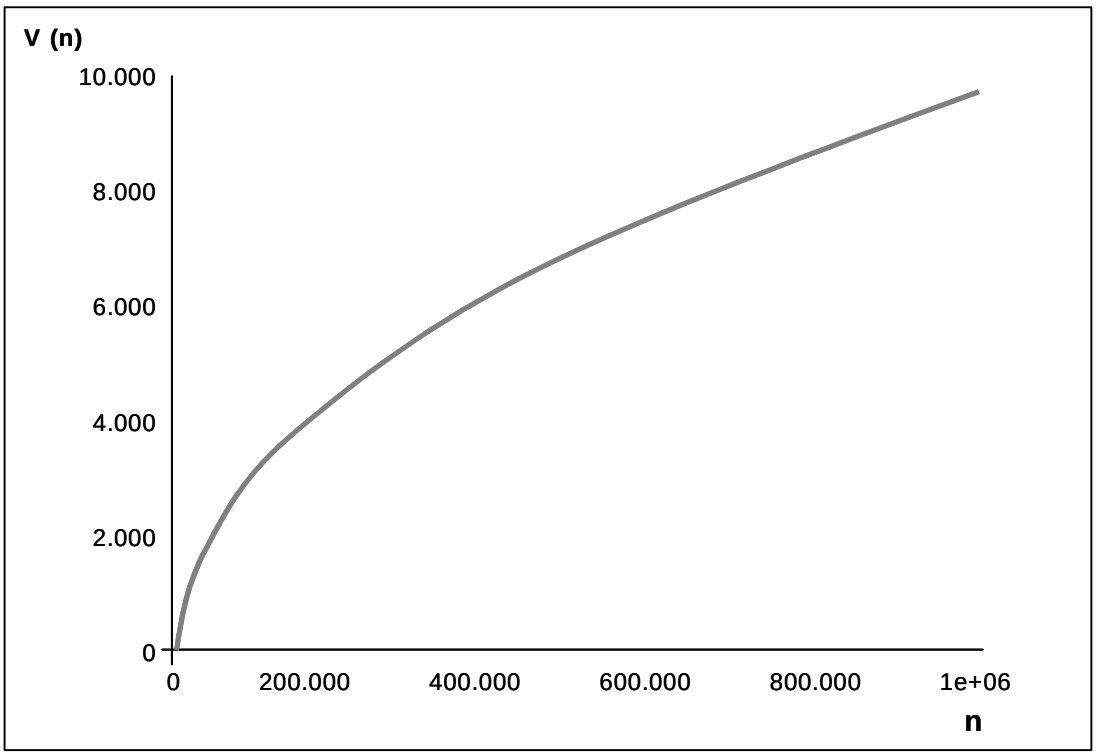


Abb. 19: Typischer Verlauf von Heaps Law

Bemerkenswert ist, dass Heaps Law in gleicher Weise in generellen Sachverhalten angewandt werden kann,[140] indem das Vokabular von einer Sammlung von Objekten ausgetauscht wird. Die Objekte müssten eigenständig und durch Attribute (Merkmalsausprägungen) erkennbar sein und Kategorien zugeordnet werden können. So könnten z.B. Personen als Objekte und das Herkunftsland der Personen als Kategorie gelten. Wählt man wahllos und absolut zufällig Personen aus der Gesamtmenge der Menschheit aus, dann besagt Heaps Law, dass sehr schnell sehr viele Herkunftsländer durch die zufällig ausgewählten Personen vertreten sind. Dagegen ist es nahezu unmöglich, mit der Zufallsmethode alle Herkunftsländer durch nur eine Person zu repräsentieren.[141]

Im folgenden Kapitel werden die deutsche Sprache und der tatsächlich verwendete Wortschatz näher beleuchtet, um Heaps Law zu verdeutlichen und weitere Überlegungen zum Indexwachstum anzustellen.

Sprache und Wortschatz

Die alltägliche (gesamte aktive) deutsche Sprache umfasst einen Wortschatz von ca. 75.000 Worten.[142] Die Gesamtgröße des deutschen Wortschatzes wird je nach Quelle und Zählweise auf 300.000 bis 500.000 Worte in ihrer Grundform (Lexem)

---

140 Vgl. http://planetmath.org/encyclopedia/HeapsLaw.html (Abruf: 20.11.2006).
141 Vgl. http://planetmath.org/encyclopedia/HeapsLaw.html (Abruf: 20.10.2006).
142 Vgl. http://de.wikipedia.org/wiki/Wortschatz (Abruf: 20.10.2006).

geschätzt.[143] Es wird zwischen aktivem und passivem Wortschatz unterschieden. Als aktiver Wortschatz werden die tatsächlich verwendeten Worte beim Sprechen bezeichnet. Ein durchschnittlicher Sprecher verfügt über einen aktiven Wortschatz von etwa 6.000 bis 10.000 Worten.[144] Als passiver Wortschatz wird die Menge eines Wortschatzes, bezeichnet den man beim Hören oder Lesen sinngemäß versteht, aber selbst nicht im Sprachgebrauch anwendet.

Für Alltagsgespräche einer Person reichen grundsätzlich 400 bis 800 Worte aus. Damit ein Leser einen anspruchsvolleren Text in Zeitschriften, Zeitungen oder klassischer Literatur verstehen kann, werden 4.000 bis 5.000 Wörter benötigt. In den neuen Kommunikationskanälen, wie der SMS beim Handy und dem Instant Messaging mittels P2P-Technologie und Internet, werden nur noch 100-200 Worte verwendet. Der Duden 2006 enthält ca. 130.000 Stichworte, davon ca. 5.300 Neubildungen (vgl. Duden 2006). Der Brockhaus multimedial 2007 wird 260.000 Artikel enthalten. Dies verdeutlicht die Existenz einer Vielzahl von Begriffen in der Deutschen Sprache. Die Häufigkeit der Verwendung einzelner Worte wird durch Untersuchungen anhand von Häufigkeitsklassen gemessen. Das Wort „der“ ist das in der deutschen Sprache am häufigsten verwendete Wort. Ausgehend davon gibt die Universität Leipzig z.B. für das Wort “Internet“ die Häufigkeitsklasse Acht an. Das bedeutet, dass das Wort „der“ acht mal häufiger verwendet wird als das Wort „Internet“.[145] Dieser Sachverhalt wird durch Zipfsches Gesetz[146] beschrieben, welches sich generell auch auf andere Bereiche anwenden lässt (vgl. Strauch/Kuhlen/Laisiepen 2004: 239).[147]

Durch Flexionen kann aus einer relativ geringen Anzahl an Worten in ihrer Grundform eine umfangreiche Menge an Wortformen entstehen. In der deutschen Sprache handelt es sich dabei um einen Faktor von ungefähr zehn. Im Vergleich dazu enthält die englische Sprache einen Flexionsfaktor von nur ungefähr vier. Grundlegend wird behauptet, dass je höher das Bildungsniveau eines Menschen ist, desto größer ist sein Wortschatz. Ein umfangreicherer Wortschatz ermöglicht einen differenzierteren Informationsaustausch. Die meisten deutschen Boulevardzeitungen nutzen einen Wortschatz von etwa 400 Worten, intellektuelle Tageszeitungen dagegen einen Wortschatz von etwa 5.000 Worten. Der Wortschatz einer

---

143 Vgl. http://www.duden.de/deutsche_sprache/newsletter/archiv.php?id=11 (Abruf: 11.11.2006).
144 Vgl. http://de.wikipedia.org/wiki/Wortschatz (Abruf: 20.10.2006).
145 Vgl. http://dict.uni-leipzig.de/cgi-bin/wort_www.exe?site=1&Wort=Internet&cs=1 (Abruf: 15.12.2006).
146 Vgl. http://de.wikipedia.org/wiki/Zipfsches_Gesetz (Abruf: 15.12.2006).
147 Weiterführende Informationen: http://en.wikipedia.org/wiki/Long_tail (Abruf: 15.12.2006).

Person ist auch abhängig von den Interessensgebieten dieser Person und der damit einhergehenden Fachtermini.

Nachstehend ist eine Liste mit den Worten abgebildet, die in der deutschen Sprache am häufigsten verwendet werden. Die Liste basiert auf von der Universität Leipzig ausgewerteten Quellen. Für andere Quellen ergeben sich andere Reihenfolgen, da die Anzahl der verwendeten Worte stark von Textsorte und Fachgebiet abhängt.[148]

1	der	26	aus	51	oder	76	unter
2	die	27	er	52	aber	77	wir
3	und	28	hat	53	vor	78	soll
4	in	29	daß	54	zur	79	ich
5	den	30	sie	55	bis	80	eines
6	von	31	nach	56	mehr	81	Es
7	zu	32	wird	57	durch	82	Jahr
8	das	33	bei	58	man	83	zwei
9	mit	34	einer	59	sein	84	Jahren
10	sich	35	Der	60	wurde	85	diese
11	des	36	um	61	sei	86	dieser
12	auf	37	am	62	In	87	wieder
13	für	38	sind	63	Prozent	88	keine
14	ist	39	noch	64	hatte	89	Uhr
15	im	40	wie	65	kann	90	seiner
16	dem	41	einem	66	gegen	91	worden
17	nicht	42	über	67	vom	92	Und
18	ein	43	einen	68	können	93	will
19	Die	44	Das	69	schon	94	zwischen
20	eine	45	so	70	wenn	95	Im
21	als	46	Sie	71	habe	96	immer
22	auch	47	zum	72	seine	97	Millionen
23	es	48	war	73	Mark	98	Ein
24	an	49	haben	74	ihre	99	was
25	werden	50	nur	75	dann	100	sagte

Abb. 20: Die 100 häufigsten Wörter der deutschen Sprache[149]

Aus dieser Liste wird bereits erkenntlich, dass (bestimmte und unbestimmte) Artikel, Pronomen, Präpositionen und Verben die in der deutschen Sprache am häufigsten gebrauchten Wörter sind und bedeutungstragende Wörter - Nomen – lediglich einen geringen Anteil ausmachen.

148 Vgl. http://wortschatz.uni-leipzig.de/html/wliste.html (Abruf: 15.12.2006).
149 Quelle: http://wortschatz.uni-leipzig.de/html/wliste.html (Abruf: 20.01.2007).

II Empirischer Teil und Versuch

1 Gegenstand und Ziel der Untersuchung

Im vorangegangenen Teil (II) wurden die für den Forschungsgegenstand dieser Arbeit relevanten theoretischen Grundlagen definiert und beschrieben. In diesem Kapitel (III) 1 findet zunächst eine genauere Eingrenzung des Untersuchungsgegenstandes statt. Davon ausgehend werden die Ziele der Untersuchung definiert und abschließend werden die Untersuchungshypothesen aufgestellt.

1.1 Gegenstand der Untersuchung

Die in Kapitel (II) 4 dargestellte P2P-Suchmaschine Faroo stellt den Kern des Untersuchungsgegenstandes dar. Mit Hilfe der angewandten Forschungsmethoden soll in der Praxis untersucht werden, welchen Einfluss die Variablen Userzahl und Zeit auf das Wachstum des Index haben. Für den Versuchsaufbau wurden 30 Probanden herangezogen, die sechs Wochen lang durch ihre alltägliche Internetnutzung und ihr Surfverhalten diesen Index generierten. Die Ergebnisse werden im Zuge der Untersuchung dargestellt und kritisch analysiert.

1.2 Ziel der Untersuchung

Ziel dieser Untersuchung ist die Überprüfung der neuartigen - dezentralen und userzentrierten - Vorgehensweise der Datensammlung auf ihre Eignung zum Aufbau eines Suchmaschinenindex. Die Überprüfung der Eignung wird anhand des Indexwachstums in Abhängigkeit von den Variablen Userzahl und Zeit durchgeführt.

Für diese Untersuchung werden nachstehend zwei Sachverhalte formuliert, die die Eignung definieren und eine Überprüfbarkeit herstellen.
Der **erste Sachverhalt** bezeichnet das Crawlingverfahren als „geeignet“, wenn das Wachstum des Index innerhalb eines Jahres eine gegenüber den drei marktbeherrschenden Suchmaschinen konkurrenzfähige Indexgröße erreicht. Als Vergleichswert für die Überprüfung werden demnach die Indexgrößen der drei Marktführer herangezogen.
Der **zweite Sachverhalt** wird anhand der theoretischen Grundlagen erarbeitet und bezeichnet das Crawlingverfahren als „geeignet“, wenn innerhalb eines Jahres ein Index generiert werden kann, der in der Lage ist, 90 Prozent **aller** Suchanfragen zu beantworten. Unter „beantworten“ soll hier verstanden werden, dass die Suchergebnislisten mindestens eine Antwort enthalten. Die „Precision“ soll hier nicht näher

untersucht werden. An dieser Stelle ist kritisch zu hinterfragen, ob eine Indexgröße wie die der drei Marktführer zwangsläufig für die Beantwortung der Suchanfragen notwendig ist. Diese zweite Variante nimmt Abstand von der Quantität der Webseiten im Index und rückt den User und dessen Informationsbedürfnis in den Mittelpunkt der Betrachtung.
Diese beiden Sachverhalte werden gegenübergestellt und in der abschließenden zusammenfassenden Diskussion (vgl. Kapitel (III) 6) bewertet. Dabei wird herausgearbeitet, welcher der Sachverhalte für die Überprüfung der Eignung der neuen Vorgehensweise zum Aufbau eines Suchmaschinenindex geeigneter ist.

Im anschließenden Kapitel werden Hypothesen formuliert, um eine Überprüfbarkeit der beiden Sachverhalte herzustellen.

1.3 Untersuchungshypothesen

Wissenschaftliche Hypothesen (griechisch: Vermutung, Unterstellung) sind wahrscheinlich richtige, aber noch zu prüfende und zu beweisende Aussagen, dargestellt in einem Satz. Um eine wissenschaftliche Hypothese von einer Alltagshypothese zu unterscheiden, sollten folgende Kriterien erfüllt sein: Die wissenschaftliche Hypothese muss sich auf reale, empirisch untersuchbare Sachverhalte beziehen und eine allgemeine Gültigkeit besitzen. Der Formulierung der Aussage sollte ein Konditionalsatz, welcher eine sinnvolle Wenn-Dann-Beziehung abbildet, zugrunde liegen. Diese Beziehung sollte anhand von empirisch erhobenen Ergebnissen verifizierbar oder falsifizierbar sein (vgl. Bortz/Döring 2002: 7ff.). An dieser Stelle ist ergänzend zu erwähnen, dass eine möglichst präzise Formulierung der Wenn-Dann-Beziehung den Verlauf der Untersuchung sowie deren Auswertung erheblich vereinfacht.
Der Wenn-Teil des Konditionalsatzes enthält die Bedingung, der Dann-Teil die Konsequenz der Aussage. Sowohl Wenn-Teil als auch Dann-Teil stellen Ausprägungen von Variablen dar. Die Variable im Wenn-Teil der Hypothese wird als unabhängige Variable bezeichnet und die im Dann-Teil als abhängige Variable. In einem Wenn-Dann-Satz werden also Beziehungen zwischen unabhängigen und abhängigen Variablen formuliert.
Mit dieser Arbeit liegt eine empirische Untersuchung vor, durch die folgende zentrale Frage eruiert werden soll:

Wie wirken sich Zeit und Userzahl auf das Wachstum eines P2P-Index aus?

Aus dieser Frage werden folgende drei Hypothesen abgeleitet, die durch die anschließende Ergebnisdarstellung und Analyse verifiziert oder falsifiziert werden sollen.

- Innerhalb des Versuchszeitraumes verläuft das Wachstum des Index nicht linear, sondern nach Heaps Law.
- Nach einem Jahr erreicht Faroo eine Indexgröße, deren Quantität mit den konventionellen Suchmaschinen konkurrieren kann (Bedingung: konstante Useranzahl von 500.000 oder 1.000.000).
- 500.000 User sind in der Lage, innerhalb eines Jahres eine Indexgröße zu generieren, mit der 90 Prozent aller gestellten Suchanfragen beantwortet werden können.

2 Forschungsmethode und Erhebungsinstrument

Die empirische Forschung dient der Überprüfung wissenschaftlicher Fragestellungen oder Theorien durch eine methodisch-systematisch angeordnete Untersuchung einer entsprechenden Situation oder eines Sachverhaltes. Hierbei liegt entweder eine bestehende Problematik zugrunde oder es gilt, eine genaue Aufklärung eines Sachverhaltes zu erarbeiten. Das hier betrachtete Problem stellt in Form einer Fragestellung (siehe Kapitel (III) 1.3) den zentralen Untersuchungsgegenstand dar. Für die konkrete Untersuchung und Beantwortung der jeweiligen Fragestellung werden existierende Theorien hinzugezogen und empirisch anerkannte Methoden verwendet (vgl. Bortz/Döring 2002: 34ff.).
In der empirischen Forschung wird zwischen qualitativen und quantitativen Forschungsmethoden unterschieden. Der Unterschied wird maßgeblich von der Form des erhobenen Datenmaterials bestimmt. Im Gegensatz zur qualitativen Forschung, die sich auf nichtnumerische Merkmale stützt, liegt der Schwerpunkt der quantitativen Forschung in der Ermittlung von Häufigkeiten, beziehungsweise quantitativ bezifferbaren Unterschieden. Die quantitative Forschung erhält ihre Erkenntnisse meist aus dem Vergleich der numerisch erhobenen Versuchsergebnisse (vgl. ebd: 295ff.). Weitere Unterschiede bestehen hinsichtlich der Forschungsmethode, des Untersuchungsgegenstands und des Wissenschaftsverständnisses (vgl. ebd).

Für diese Arbeit wurde eine quantitative Forschung mit einer entsprechenden Forschungsmethode (Feldexperiment) als Untersuchungsmethode gewählt, da der Untersuchungsgegenstand Merkmale mit numerischem Charakter aufweist. Im

Folgenden werden die in dieser Arbeit angewandte Forschungsmethode und das Verfahren zur Datenerhebung beschrieben.

2.1 Feldexperiment als Forschungsmethode

Das Experiment (lateinisch: experimentum = Versuch, Beweis, Prüfung, Probe) ist ein wissenschaftlich, methodisch angelegter Untersuchungsaufbau zur **zielgerichteten Untersuchung** einer unter **definierten Bedingungen** - möglichst reproduzierbar - hervorgerufenen Erscheinung. Das Experiment ist neben der genauen Beobachtung die wichtigste wissenschaftliche Forschungsmethode, um etwas über die Realität zu erfahren. Außerdem ist es die einzige Methode, die zuverlässige Aussagen über **Ursachen-Wirkungszusammenhänge** ermöglicht. Der Versuch ist die reale Umsetzung des Experimentes. Die durch den Versuch gewonnenen Ergebnisse dienen als Grundlage dazu, eine vorher aufgestellte Annahme oder Vermutung (**Hypothese**) an einem kontrollierten Modell beweisen oder widerlegen zu können. Eine wichtige Unterscheidung wird in der Literatur zwischen „echten Experimenten", „Ex-post-facto-" und „Quasi-experimentellem-" Forschungsdesign vorgenommen. Diese Forschungsdesigns unterscheiden sich des Weiteren durch die Art und Weise der Umsetzung in Laborexperiment und Feldexperiment (vgl. Schnell/Hill/Esser 2005: 220ff.).

Das **Laborexperiment** ist dadurch gekennzeichnet, dass die Durchführung des Versuchs in einem speziellen Labor stattfindet, so dass eine gute Kontrollierbarkeit von Störvariablen möglich ist. Somit führt ein Laborexperiment zu hoher interner Validität, was aber auch bedeutet, dass die Verallgemeinerbarkeit auf den Anwendungsbereich nur eingeschränkt möglich ist.

Im Gegensatz dazu wird das **Feldexperiment** dadurch gekennzeichnet, dass die Durchführung in der natürlichen Umgebung der Versuchsprobanden stattfindet. Somit ist eine Kontrollierbarkeit von Störvariablen nur eingeschränkt möglich. Die Probanden verhalten sich jedoch natürlich, wodurch eine Verallgemeinerbarkeit auf den Anwendungsbereich gegeben ist (vgl. Schnell/Hill/Esser 2005: 225ff.). Störvariablen können beispielsweise zwischenzeitliches Geschehen, Reifungsprozess der Probanden, Hathrown Effekt[150], Messeffekte, verzerrte Auswahl oder Ausfälle sein.

Um die quantitativen Daten erhebeben zu können, wurde für diese Arbeit ein Feldexperiment mit 30 Probanden durchgeführt. Die Methode des Feldexperimentes

---

[150] Dieser Effekt beschreibt eine Verhaltensänderung der Probanden, welche auf das Ziel ausgerichtet ist. Daraus folgt, dass das Ziel der Untersuchung den Probanden nie im Voraus genannt werden sollte, um diese Verhaltensänderung und die damit einhergehende Ergebnisverfälschung zu vermeiden. (vgl. Bortz/Döring 2002: 269).

wurde zum einen aufgrund der numerischen Eigenschaften der Variablen gewählt und zum anderen, da möglichst die unverfälschte Realität wiedergegeben werden soll. Die zu untersuchenden Variablen sind in der Realität nicht direkt beeinflussbar. Die Variable „User“ ist in ihrem Verhalten - der Teilnahme - nicht vorhersehbar. Oft sind User in ihrem Verhalten unzuverlässig - diese Situation muss eine P2P-Technologie berücksichtigen und dennoch funktionieren.

2.2 Computergestützte Beobachtung als Erhebungsinstrument

Die Erhebung der Daten fand anhand der implementierten dezentralen Statistik - ersichtlich auf der Faroo Startseite - statt. In regelmäßigen Abständen wurde die Anzahl der Worte sowie die Anzahl der Webseiten abgelesen und tabellarisch erfasst. Einmal wöchentlich ließen die Probanden der Versuchsleiterin Screenshots ihrer Faroo-Startseite (siehe Abb. 21) zukommen. Dies diente der Überprüfung der Funktionstüchtigkeit von Faroo und sollte einen fehlerfreien Verlauf des Versuchs garantieren. Durch die bei den Usern installierte P2P-Softwareapplikation wurden alle Surfaktivitäten der User im Internet automatisch von dem implementierten HTTP-Monitor aufgezeichnet und dem Index zugeschrieben. Dem User selbst ist dies zumeist unbewusst. Auf diese Weise entfallen Verzerrungen der Ergebnisse durch Verhaltensänderungen (vgl. Bortz/Döring 2002: 269).

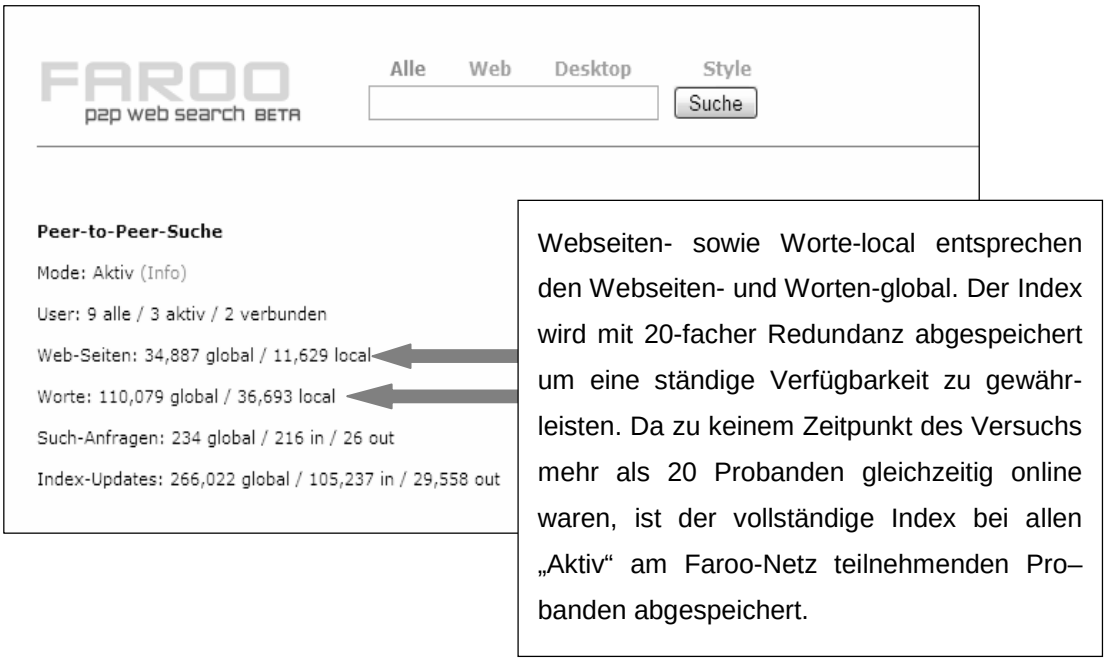


Abb. 21: Screenshot der Faroo Startseite

3 Versuchsplanung und Durchführung

In der Planungsphase wurde die Zielgruppe bestimmt, Rahmenbedingungen und mögliche Störvariablen definiert sowie die technische Umsetzung organisiert. Im Folgenden wird eine detaillierte Beschreibung des Versuchsaufbaus und der Durchführung vorgenommen.

3.1 Auswahl der Probanden

Für die Stichprobe wurden exemplarisch 30 Probanden - minimale Stichprobengröße (vgl. Kühl 2005: 100) - nach zukünftiger Zielgruppenorientierung ausgewählt. Die Auswahl der Zielgruppe fand anhand soziodemographischer Daten, Einstellungen und Lebensstilen statt. Für den Aufbau dieses Versuchs wurde der Fokus aus organisatorischen Gründen auf die deutschen Internetuser gelegt.

Beide Geschlechter waren gleich stark vertreten. Die Probanden sollten in der Altersgruppe der 20 bis 35-jährigen liegen und eine starke Affinität zum Internet besitzen. Diese Altersgruppe ist fast vollständig im Internet vertreten und weist aufgrund der bereits mehrjährigen Nutzung dieses Mediums fortgeschrittene Kenntnisse auf (vgl. Kapitel (II) 1.3). Das Medium Internet hat sich fest in ihrem Alltag etabliert, sie beteiligen sich aktiv und beschäftigen sich interessiert mit dem Internet und dessen Möglichkeiten. Die Muttersprache der Probanden ist deutsch. Da es sich indes überwiegend um Studenten und Hochschulabsolventen handelt, ist davon auszugehen, dass Englisch als Zweitsprache einen gewissen Anteil an Inhalten ausmacht. Die Probanden stammen aus diversen Fachrichtungen, wodurch eine einseitige themenspezifische Internetnutzung ausgeschlossen und eine große Interessenvielfalt gegeben ist.

Der erste Kontakt zu den Probanden wurde über ein Anschreiben via E-Mail mit einer Anfrage bezüglich ihres Interesses an einer Versuchsteilnahme hergestellt. Diese Herangehensweise impliziert eine freiwillige Teilnahme und schafft somit eine Grundlage für den Erfolg des Versuchs. Vor Versuchsbeginn wurde ein Flyer mit Informationen zu Kontext und Inhalt der Durchführung verteilt. Des Weiteren fand eine Informationsveranstaltung zum Ablauf und zur Klärung offener Fragen statt.

3.2 Erhebungszeitraum und Rahmenbedingungen

Der Erhebungszeitraum wurde vorab auf eine Laufzeit von sechs Wochen festgelegt. Um im Voraus einige Störvariablen wie zum Beispiel eine mögliche Inkompatibilität des Betriebssystems oder der Softwareapplikation Faroos zu eliminieren, wurden folgende Rahmenbedingungen festgelegt:

- Installiertes Betriebssystem: Windows XP[151]
- Verwendeter Browser: Internet Explorer[152]
- Internetzugang: Breitband/DSL mit Flatrate[153]

Die Aufgabe der Probanden bestand darin, das Internet nach der Installation der Faroo-Software wie gewöhnlich zu nutzen.

Durch vorzeitige Eindämmung von Störvariablen wird ein störungsfreier Untersuchungsverlauf garantiert. Ein Feldversuch ist jedoch auf Realitätsnähe ausgelegt und enthält somit Variablen, auf die kein Einfluss genommen werden kann. Dies sind zum Beispiel verschiedene Alltagssituationen wie Urlaub, Umzug oder Krankheit, aber auch technische Probleme mit der Software, der Kompatibilität oder der Installation (vgl. Bortz/Döring 2002: 528ff.).

3.3 Technische Umsetzung

Für die vorliegende Arbeit wurde ein verdecktes Peer-to-Peer-Netzwerk aufgebaut. Der Aufbau fand durch die Installation der Faroo-Software auf den Rechnern der Probanden statt. Über das in der Software implementierte IP-Discovery-Protokoll waren die Rechner in der Lage, sich zu erkennen, miteinander zu kommunizieren und ein Overlay-Netzwerk (siehe Kapitel (II) 3 bis 3.3) aufzubauen. Nach der Installation der Software wurde diese automatisch mit dem Booten des Rechners gestartet. Sobald ein User eine Webseite im Browser aufrief, wurde der Webseiteninhalt wie in Kapitel (II) 5.1.1 beschrieben zerlegt und dem Index zugeschrieben.

3.4 Ablauf des Versuchs

Nach Abschluss der Planungsphase wurde den Probanden die Software zur Installation ausgehändigt und der Versuch somit gestartet. Die Installation und die Funktionstüchtigkeit wurden anhand der zugesandten Screenshots überprüft. Während der Durchführung des Versuchs wurden die Probanden begleitet und bei Bedarf angeleitet. Parallel zum Versuch wurde ein Faroo-Forum zum Austausch über Probleme, Fragen und Anregungen eingerichtet. Nach Ablauf des festgesetzten Versuchszeitraumes wurden die Daten sortiert und komplettiert und werden im nachfolgenden Kapitel dargestellt.

---

[151] Die Kompatibilität mit anderen Betriebssystemen war zum Zeitpunkt des Versuchs noch nicht gegeben.
[152] Die Kompatibilität mit anderen Browsern war zum Zeitpunkt des Versuchs noch nicht gegeben.
[153] Diese Bedingung impliziert bereits eine regelmäßige Internetnutzung.

4 Darstellung der Ergebnisse

In diesem Kapitel werden die durch den Versuch erhobenen Daten und Informationen graphisch und textuell aufbereitet. Durch Darstellung der Versuchsergebnisse wird zunächst die Fragestellung erörtert: **Welche Indexgröße der Faroo-Index während des Versuchszeitraumes erreicht hat.** Des Weiteren ermöglichen sie eine Hochrechnung und Prognose der Indexgröße nach einem Jahr.

Zunächst soll festgehalten werden, dass sich die Teilnehmerzahl aufgrund technischer Probleme[154] von 30 auf 25 Probanden verringert hat. Die gewonnenen Daten sind in Tabelle 7 dargestellt. Die Daten wurden erstmals nach einer Woche abgelesen. Es ist erkennbar, dass zu diesem Zeitpunkt bereits 5.968 Webseiten und 19.141 eigenständige Worte indexiert waren. Stoppworte finden in den Angaben „Worte local“ keine Beachtung, sie wurden automatisch eliminiert.

Laufzeit in Wochen	Webseiten local	Worte local
1.	5.968	19.141
2.	9.048	26.120
3.	10.424	35.010
4.	13.494	43.620
5.	16.330	51.025
6.	19.726	58.284

Tabelle 7: Anzahl der indexierten Worte und Webseiten.

Als Endergebnis des Versuchs kann festgehalten werden, dass 25 Probanden durch ihr Surfverhalten in einem Zeitraum von sechs Wochen eine Indexgröße von 19.726 Webseiten mit 58.284 unterschiedlichen Worten indexiert haben. Die Bildung der Differenzen der einzelnen Werte in Tabelle 8 verdeutlicht die Steigerung des Wachstums von Webseiten und Worten in Abhängigkeit von der Zeit.

[154] Trotz vorher festgelegten Rahmenbedingungen traten Inkompatibilitäten zwischen der Faroo-Software und den Rechnern von fünf Probanden auf.

Laufzeit in Wochen	Webseiten local	Differenz Webseiten	Worte local	Differenz Worte
1.	5.968		19.141	
2.	9.048	3.080	26.120	6.979
3.	10.424	1.376	35.010	8.880
4.	13.494	3.070	43.620	8.620
5.	16.330	2.836	51.025	7.405
6.	19.726	3.396	58.284	7.259

Tabelle 8: Differenz von Worten und Webseiten pro Woche

Die Probanden besuchten demnach im gesamten Erhebungszeitraum ungefähr 3000 Webseiten pro Woche. (Mit Ausnahme von Woche drei, was auf das Surfverhalten[155] der User zurückgeführt werden kann).

Die Anzahl der wöchentlich neu indexierten **Webseiten** ist unabhängig von der Laufzeit der Untersuchung. In Abb. 22 ist der Verlauf des Wachstums von Webseiten im Index innerhalb des Versuchszeitraumes visualisiert. Die schwarzen Linien sowohl in Abb. 22 als auch in Abb. 23 stellen den realen Verlauf des Wachstums von Worten und Webseiten dar und geben die absoluten Werte wieder. Die leichten Schwankungen der absoluten Werte verdeutlichen den variierenden Zuwachs an Worten und Webseiten (im Index) in dem jeweiligen Zeitintervall und sind auf die unterschiedlich starke Internetnutzung der Probanden als auch auf den Indexierungsprozess von Faroo zurückzuführen.

[155] Surfverhalten beschreibt die Art und

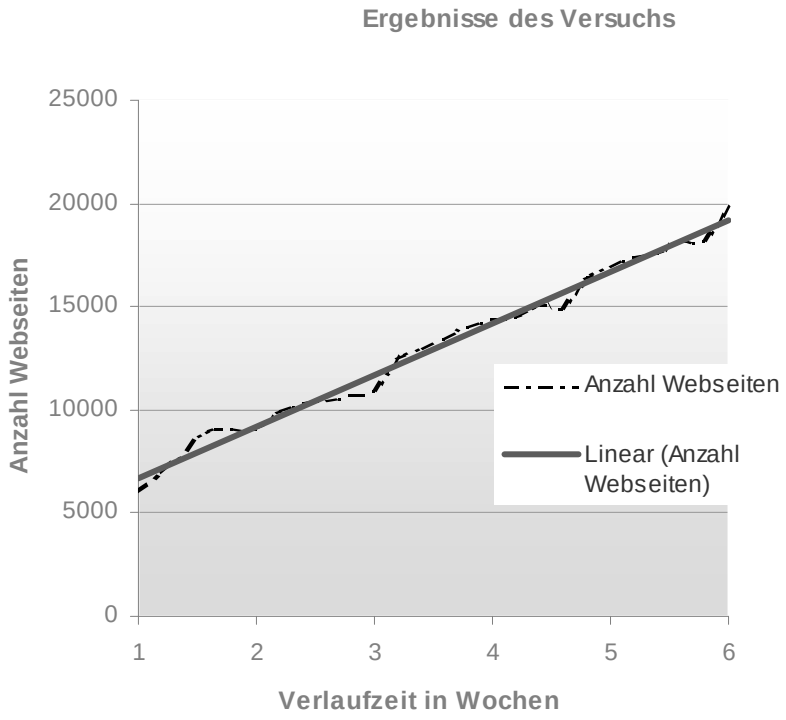


Abb. 22: Darstellung der erhobenen Webseitenanzahl

Der Verlauf der Neuentdeckung von **Worten** (Differenz der neu hinzugekommenen Worte pro Woche) veranschaulicht, dass mit zunehmender Laufzeit und neuen Webseiten (Text) weniger unterschiedliche Worte entdeckt werden (mit Ausnahme des ersten Wertes). Dieser Verlauf wurde mit Heaps Law vorausgesagt und ist in Abb. 23 anhand der eingefügten logarithmischen Trendlinie angedeutet. Die logarithmische Trendlinie wurde gewählt, da sie einem Kurvenverlauf von Heaps Law entspricht.

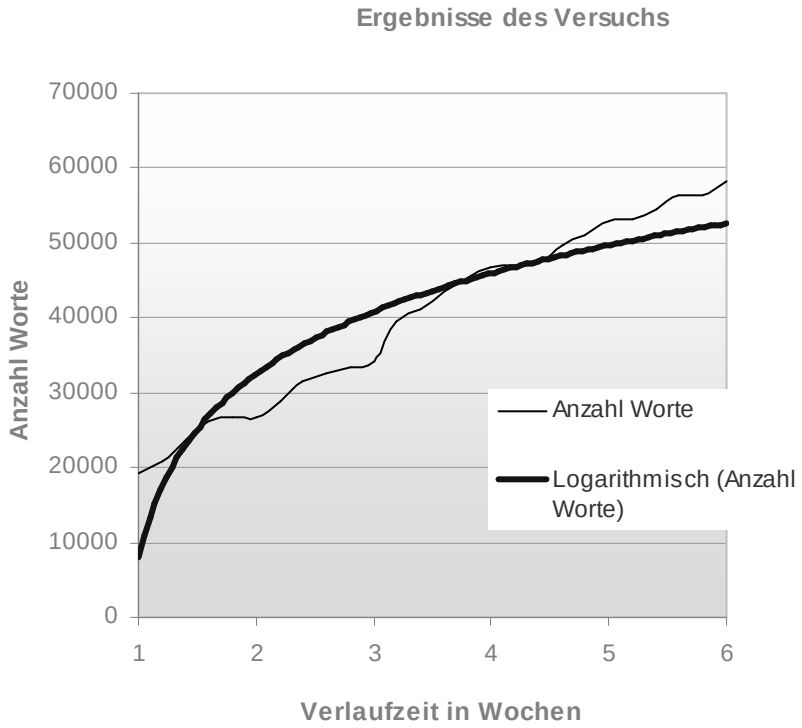


Abb. 23: Darstellung der erhobenen Anzahl an Worten

Als weiteres Ergebnis kann aus den erhobenen Daten die Anzahl der besuchten Webseiten pro User und Tag ermittelt werden: So hat jeder User pro Tag im Durchschnitt 18,8 Webseiten aufgerufen.

Anzahl indexierter Webseiten	Anzahl partizipierender User	Laufzeit in Tagen	Anzahl indexierter Webseiten/User/Tag
19.726	25	42	18,8

Tabelle 9: Ergebnisse des Versuchs

Auf Grundlage der oben dargestellten Ergebnisse des Versuchs werden im nächsten Kapitel weitere Überlegungen und Interpretationen durchgeführt und detaillierter beschrieben.

5 Analyse und Interpretation der Ergebnisse

In diesem Kapitel werden die vorab dargestellten Daten und Ergebnisse mit den im Theoretischen Teil vorgestellten Grundlagen miteinander verknüpft. Die Auswertung der Ergebnisse des Experimentes werden unter der Berücksichtigung der graphischen Darstellungen sowie Einbeziehung der wissenschaftlichen Fachliteratur zu einem Abschlussbericht zusammengefasst.

Zur Auswertung wird Heaps Law zur Hilfe genommen. Da die Anzahl aller aufgerufenen Webseiten proportional[156] zur Variablen Zeit verläuft, ist eine Anwendung von Heaps Law auf das Wachstum des Index möglich. Die folgende Darstellung soll ein Verständnis für diese Übertragung schaffen: So wie Heaps Law das Vorkommen von eigenständig Worten (V) in einem Text (mit einer Anzahl von n Worten) beschreibt, so verhält sich das Aufrufen unterschiedlicher Webseiten (V) zu allen aufgerufenen Webseiten (n).

Die zehn am häufigsten verwendeten Wörter der deutschen Sprache (wie z.B. „Prozent“) sind mehrfach in einem Text enthalten (siehe Kapitel (III) 5.2), während eigenständige Wörter wie „Internet“ und „Suchmaschine“ mit einer wesentlich geringeren Häufigkeit vorkommen. Dieser Sachverhalt ist in allen natürlichen Sprachen zu beobachten. Genauso verhält sich die Anzahl eigenständiger Webseiten zu allen aufgerufenen Webseiten. Webseiten mit tagesaktuellen Nachrichten wird von weitaus mehr als einer Person aufgerufen, während private Webseiten mit einer wesentlich geringeren Häufigkeit von Usern besucht werden. Da sich die Anzahl aller aufgerufenen Webseiten sowohl proportional zur Zeit als auch zur Anzahl der User verhält, kann Heaps Law auch für die Analyse angewendet werden.

Im Folgenden werden die beteiligten Variablen „Useranzahl“ und „Zeit“ auf deren Einflussmöglichkeit analysiert.

5.1 Analyse der Variablen „Zeit“ und „Userzahl“ anhand der Versuchsergebnisse

In diesem Kapitel werden die zwei unbekannten Parameter *K* und *ß* aus Heaps Law für die empirisch erhobenen Daten bestimmt. Die Parameter werden sowohl für die Anzahl der Webseiten als auch für die Anzahl der Worte errechnet. Dies ist erforderlich, um auf Basis von Heaps Law das Wachstum des Index (Anzahl der Webseiten) sowie das Wachstum der Anzahl der unterschiedlichen Worte auf ein Jahr zu prognostizieren.

Berechnung: Es werden die Koordinaten zweier Punkte bestimmt. Die erste Koordinate wird nach einer Woche aus Tabelle 7 abgelesen. Der zweite Punkt wird nach der sechsten Woche abgelesen. Somit erhält man für die Berechnung folgende Punkte:

---

[156] Steigt die Userzahl, steigt entsprechend auch die Anzahl aller aufgerufenen Webseiten. Verlängert sich die Zeit, steigt im Verhältnis stehend auch die Anzahl aller aufgerufenen Webseiten.

Parameter Webseiten:	Parameter Worte:
Punkt-a: (1/5.968)	(1/19.141)
Punkt-b: (6/19.726)	(6/58.284)

Im Folgenden werden die beiden Koordinatenpaare in die Formel von Heaps Law eingesetzt, um zwei Gleichungen zu erhalten. Im Fall der Webseiten steht die Variable V für Anzahl der eigenständigen, einzelnen Webseiten im Index und entspricht somit der Indexgröße. Im Fall der Worte steht die Variable V für die Anzahl der eigenständigen Worte im Index und entspricht somit der Anzahl beantwortbarer, unterschiedlicher Suchanfragen. Parameter n steht für die Anzahl der Wochen, die sich zur Anzahl aller aufgerufenen Webseiten bzw. Worte proportional verhalten (vgl. Kapitel (III) 5).

Heaps Law:	$V_R(n) = K\, n^{ß}$
Webseiten:	Worte:
Formel eins: $5.968 = K * 1^{ß}$	Formel eins: $19.141 = K * 1^{ß}$
$K = 5.968 / 1^{ß}$	$K = 19.141 / 1^{ß}$
Formel zwei: $19.726 = K * 6^{ß}$	Formel zwei: $58.284 = K * 6^{ß}$

Die erste Formel wird nach *K* aufgelöst und der Wert für *K* in die zweite Formel eingesetzt. Dadurch wird eine Gleichung mit nur einer unbekannten Variable (ß) erhalten.

$19.726 = [5.968 / 1^{ß}] * 6^{ß}$	$58.284 = [19.141 / 1^{ß}] * 6^{ß}$

Die Formel aufgelöst nach *ß* ergibt:

$19.726 / 5.968 = (6/1)^{ß}$	$58.284 / 19.141 = (6/1)^{ß}$
$_{6/1}log\ 19.726 / 5.968 = ß$	$_{6/1}log\ 58.284 / 19.141 = ß$
$ß \sim 0{,}65$	$ß \sim 0{,}69$

Somit ist der Parameter *ß* sowohl für Webseiten als auch für Worte empirisch bestimmt worden. Parameter *K* lässt sich jetzt ebenfalls durch Einsetzen in die Ursprungsformel von Heaps Law errechnen.

$19.726 = K * 6^{0{,}65}$	$58.284 = K * 6^{0{,}69}$
$K \sim 6.155$	$K \sim 16.628$

Für den Parameter *K* ist somit für die Anzahl der Webseiten ein Wert von *K*~6.155 und für die Anzahl der Worte ein Wert von *K*~16.628 empirisch bestimmt worden. Die beiden Parameter *K* und *ß* geben den Verlauf beziehungsweise den Krümmungsgrad von der durch Heaps Law definierten Funktion an.

Mit diesen Parametern kann im weiteren Verlauf berechnet werden, wie viele Webseiten und Worte innerhalb eines Jahres (n=52 Wochen) von 25 Usern (Anzahl der Versuchsteilnehmer) indexiert werden können.

$V_R(n) = Kn^{ß}$	$V_R(n) = Kn^{ß}$
$V = 6.155 * 52^{0,65}$	$V = 16.628 * 52^{0,69}$
V ~ 80.284	V ~ 264.268

Anhand dieser Vorgehensweise kommt die Untersuchung zu folgendem Ergebnis: Nach Heaps Law haben 25 User innerhalb eines Jahres die Kapazität, 80.284 unterschiedliche Webseiten und 264.268 Worte zu indexieren. Dieses Ergebnis gilt allerdings nur unter der Annahme, dass die User konstant das gleiche Internetverhalten und Informationsbedürfnis aufrecht erhalten, wie während des Versuchszeitraumes.

Aus dem hochgerechneten Jahresergebnis von 80.284 Webseiten lässt sich rechnerisch eine durchschnittliche Webseitenzahl von 8,8 Webseiten[157] pro Tag und User ermitteln. Aus dem Ergebnis lässt sich ableiten, dass in der Berechnung mit Heaps Law das Aufrufen redundanter Webseiten tatsächlich berücksichtigt wird und Redundanzen eliminiert werden. Die durchschnittlich aufgerufene Anzahl an Webseiten von 18,8 während des sechswöchigen Versuchszeitraums hat sich bei einer Laufzeit von einem Jahr auf 8,8 Webseiten reduziert.

5.2 Analyse der Variablen „Zeit“ und „Userzahl“ anhand der theoretischen Grundlagen

Die im vorigen Kapitel ermittelte durchschnittliche Webseitenzahl von 8,8 für zwölf Monate dient als Grundlage für weitere Überlegungen. Basierend darauf erfolgt in Tabelle 10 anhand möglicher Userzahlen eine Hochrechnung auf mögliche Indexgrößen innerhalb eines Jahres. Es lässt sich vermuten, dass eine divergierende Gruppe von Probanden andere Ergebnisse liefern würde, da diese mit hoher Wahr-

[157] Das ermittelte Ergebnis von 8,8 Webseiten liegt knapp über dem von Welp ermittelten Wert von 4,5 Webseiten. Der Wert von Welp wurde in einem Faroo-unabhängigen Versuch ermittelt (vgl. Welp 2003: 94 und Eimeren/Gerhard/Frees 2004: 355).

scheinlichkeit ein abweichendes Internetverhalten und abweichende Interessen aufweisen würden. Dennoch werden die für die vorliegende Arbeit ermittelten Ergebnisse als feste Größe zur weiteren Berechnung angesehen. Tabelle 10 zeigt anhand der in Kapitel (II) 1.1 und (II) 1.2 erarbeiteten Userzahlen im Internet selbst und den verschiedenen dort existierenden Communities auf, wie viele Webseiten mit dem userzentrierten Crawlingverfahren indexiert werden könnten. Dabei wird von durch den Versuch ermittelten 8,8 indexierbaren Webseiten pro User und Tag ausgegangen. Diese Überlegung veranschaulicht, dass **500.000 User** auf Basis der Versuchsergebnisse **1,6 Milliarden Webseiten** indexieren könnten. Wenn jeder deutsche „Onliner" am Faroo-Netz partizipieren würde, wären theoretisch die Anzahl aller im Web vorhandenen Webseiten innerhalb eines Jahres indexierbar.

Anzahl unterschiedlicher Webseiten pro User und Tag	**Userdefinition und -anzahl**		**Anzahl an Seiten innerhalb eines Jahres**
8,8	beim Start von Community Netzwerken	500.000	1.605.680.000
8,8	fortgeschrittene Anzahl an Usern junger Community Netzwerke nach meist einem Jahr	1.000.000	3.211.360.000
8,8	Anzahl von Skype Usern	6.000.000	19.268.160.000
8,8	Anzahl Internetuser Deutschland	50.616.207	162.546.862.512
8,8	Anzahl Internetuser USA	207.161.706	665.270.816.180
8,8	Anzahl Internetuser weltweit	1.000.000.000	3.211.360.000.000

Tabelle 10: Anzahl indexierter Webseiten anhand verschiedener Userzahlen[158]

Die vorab angestellten Überlegungen zum Crawlingverfahren (welche dieses Verfahren als „geeignet" bezeichnen, wenn eine ähnliche Indexgröße wie die der marktführenden Suchmaschinen innerhalb eines Jahres erreicht wird) werden als Basis zu weiteren Berechnungen verwendet. Im Folgenden werden die Indexgrößen der „Global Player" (siehe Kapitel (II) 5.1.2) als zu erreichendes Ziel dargestellt. Ausgehend von diesem Ziel wurde berechnet, wie viele unterschiedliche Seiten ein User pro Tag besuchen müsste, um die vorgegebene Indexgröße zu erreichen. Tabelle 11 stellt diese Berechnung übersichtlich dar. Für eine Indexgröße von acht Milliarden Webseiten (Google) und einer Useranzahl von 500.000 müsste ein User

[158] Quelle: eigene Darstellung.

jeden Tag 44 unterschiedliche Webseiten aufrufen. Bei einer Userzahl von einer Million müsste jeder User 22 unterschiedliche Webseiten innerhalb eines Jahres pro Tag aufrufen, um an die Indexgröße des Marktführers heranzukommen. Weiter zeigt Tabelle 11, dass bei einer Millionen User zwölf unterschiedliche Webseiten pro User und Tag aufgerufen werden müssten, um an eine mit MSN-Suche vergleichbare Indexgröße heranzukommen.

Suchmaschine		**Google**	**Yahoo**	**MSN-Suche**
Indexgröße		8.000.000.000	5.000.000.000	4.200.000.000
Anzahl nötiger Webseitenaufrufe bei einer Anzahl von ... Usern:	500.000	44	27	23
	1.000.000	22	14	12

Tabelle 11: Anzahl unterschiedlicher Webseitenaufrufe pro User und Tag für ein Jahr [159]

Ein Vergleich der in Tabelle 11 dargestellten und den empirischen ermittelten Werten zeigt deutlich, dass die Ergebnisse divergieren. Den erforderlichen 22 aufzurufenden Webseiten stehen bei einer Userzahl von einer Million laut Versuchsergebnis nur 8,8 Webseiten pro User und Tag gegenüber. Die Indexgrößen der etablierten Suchmaschinen basieren allerdings, im Gegensatz zur experimentellen Hochrechnung für ein Jahr, auf mehrjährigem Bestehen.

5.3 Anmerkungen zu den Analysen

Die Analysen und Hochrechnungen basieren auf einer sehr geringen Zahl von Usern. Es ist zu beachten, dass alle Probanden aus einer Altersklasse stammen, annähernd das gleiche Bildungsniveau besitzen und somit annähernd das gleiche Vokabular verwenden. Des Weiteren muss bedacht werden, dass weder Senioren („Silver-Surfer“) noch Kinder beteiligt waren. Die entsprechenden Informationsbedürfnisse dieser Gruppen finden demnach in den Ergebnissen keine Beachtung. Die sogenannten „Silver-Surfer“ verwenden z.B. ein älteres Vokabular als Suchbegriffe. So würde eine ältere Frau beispielsweise eher Informationen zu einem Hackbratenrezept suchen, während die jüngere Zielgruppe nach Rezepten zu Fingerfood oder Cocktails recherchieren würde. Diese Interessensunterschiede spiegeln sich in weiteren Bereichen, wie beispielsweise der Urlaubsplanung, wider (vgl. Faller 2005).

[159] Quelle:eigene Darstellung.

Des Weiteren finden über ein ganzes Jahr gesehen themen- oder veranstaltungsspezifische sowie jahreszeitenabhängige Suchanfragen statt, die den Index punktuell weiter wachsen lassen.

Anzumerken bleibt, dass sich die Ergebnisse ausschließlich auf Webseiten beziehen. Zum Versuchszeitpunkt war Faroo noch nicht in der Lage, PDF- oder andere Textverarbeitungsdateien zu indexieren. Welche Dokumentensammlungen (Datenbanken, FTP-Server, Newsgroups aus dem Usenet, das gesamte Web) und welche Dokumentenarten (HTML-basierte Webseiten, mit einem Text- oder Tabellenverarbeitungsprogramm erstellte Dokumente, Multimedia-Formate) indexiert werden können, wird zu einem ausschlaggebenden Einflussfaktor für die Indexgröße.

Bei der Betrachtung der Analyseergebnisse ist zusätzlich zu bedenken, dass die für die Generierung eines konkurrenzfähigen Index erforderliche Userzahl erst nach und nach gewonnen werden kann und nicht ad hoc vorhanden sein wird. Die präsentierten Ergebnisse beziehen sich jedoch bereits auf den Zeitpunkt, an dem die angenommene, erforderliche Anzahl der User erreicht ist. Dieser Zeitpunkt ist im Rahmen dieser Arbeit jedoch nicht bestimmbar.

6 Hypothesenprüfung und Diskussion

In diesem Kapitel werden die Erkenntnisse des Theoretischen Teils mit den Ergebnissen der empirischen Erhebung verglichen. Die in der Ergebnisanalyse gewonnenen Daten zum Wachstum des Index werden zusammengefasst und diskutiert. Im Folgenden werden die drei aufgestellten Hypothesen wiederholt, um sie anschließend einer Verifikation oder Falsifikation zu unterziehen.

Hypothese 1:
Innerhalb des Versuchszeitraumes verläuft das Wachstum des Index nicht linear, sondern nach Heaps Law.

Die Ergebnisse des Versuchs zeigen deutlich einen linearen Verlauf des Indexwachstums (Anzahl der Webseiten). Somit ist die aufgestellte Hypothese widerlegt. Dennoch schließen die gewonnenen Erkenntnisse einen Verlauf des Indexwachstums nach Heaps Law, auf einen längeren Zeitraum betrachtet, nicht aus. Eher deutet der näherungsweise lineare Anstieg der Webseitenanzahl mit der Zeit darauf hin, dass hierbei der Anfangsbereich der Heaps-Law-Kurve (ebenfalls annähernd linear) erfasst wurde. Die Steigung der Kurve nimmt erst zu einem späteren Zeitpunkt ab. Dies ist naheliegend, da mehrere Milliarden Webseiten im Netz existieren

und davon erst ein sehr kleiner Teil durch die geringe Anzahl an Probanden und einen relativ kurzen Zeitraum erschlossen wurde.

Demzufolge kann vermutet werden, dass sich der Verlauf der Versuchsergebnisse zur Zeit noch im Anfangsstadium des Wachstums befindet, sich jedoch bei einer längeren Versuchslaufzeit dem Verlauf von Heaps Law annähern würde. Das Ergebnis besagt zudem, dass wesentlich mehr User als nur 25 und ein wesentlich längerer Zeitraum als sechs Wochen nötig sind, um einen Verlauf des Indexwachstums nach Heaps Law (Abnahme der Anzahl neu indexierter Webseiten pro Woche) erkennen zu können. Insbesondere die „Dauerbrenner" sowie die impulse- und eventabhängigen Informationsbedürfnisse führen im ersten Jahr zu einem stetigen Zufluss neuer Informationsbereiche in den Index. Somit ist zu vermuten, dass das Wachstum des Index erst nach ungefähr einem Jahr (in dem die Dauerbrenner sowie die periodisch wiederkehrenden Events bereits abgedeckt wurden) den abflachenden Bereich der Wachstumskurve erreicht.

Im Rahmen dieser Arbeit gilt die Hypothese daher aufgrund der oben stehenden Erkenntnisse als falsifiziert und ist zu modifizieren. Die modifizierte Hypothese lautet: Nach der Laufzeit von einem Jahr verläuft das Wachstum des Index nach Heaps Law. Diese Hypothese ist durch weitere Versuche zu erforschen.

Hypothese 2:
Nach einem Jahr erreicht Faroo eine Indexgröße, deren Quantität mit den konventionellen Suchmaschinen konkurrieren kann (Bedingung: eine konstante Useranzahl von 500.000 oder 1.000.000).

Diese Hypothese stellt deutlich die Quantität (Anzahl der Webseiten im Index) in den Vordergrund der Betrachtung. Wie die Hochrechnungen in Tabelle 10 zeigen, ist eine mit den konventionellen Suchmaschinen konkurrierende Indexgröße **nicht** mit 500.000 Usern und auch **nicht** mit einer Million Usern innerhalb eines Jahres zu erreichen. Hier ist anzumerken, dass die Berechnung auf dem Surfverhalten einer speziellen Usergruppe basiert. Würde durch eine spezielle Marketingstrategie und Marketingkampagne eine Zielgruppe mit sehr hoher Internetnutzung oder sogar beruflich bedingter, intensiver Onlinerecherche[160] gewonnen werden, wäre eine quantitativ konkurrenzfähige Indexgröße möglicherweise generierbar.

[160] Zu der Gruppe mit beruflich bedingten Onlinerecherchen könnten z.B. professionelle Blogger, mit den verschiedenen themenspezifischen Interessensbereichen gehören.

Beim Vergleich mit einer Analyse der versehentlich veröffentlichten AOL-Userdaten werden die in dieser Arbeit angestellten Überlegung bestätigt und zeigen ebenfalls, dass eine quantitativ konkurrenzfähige Indexgröße nicht mit Durchschnittsusern erreicht werden kann. Bei einer Useranzahl von 657.426 wurden über einen Zeitraum von vier Wochen 19.343.540 unterschiedliche URLs angeklickt.[161] Demnach ergibt sich eine durchschnittliche Anzahl von 30 Webseiten pro User und Monat und knapp eine Webseite pro User und Tag. Davon ausgehend müsste der User weiteren 43 Verlinkungen (vgl. Kapitel (III) 5.2; Tabelle 11) folgen, um einen Index mit der Größe des Google-Index zu generieren.

Diese Hypothese gilt als falsifiziert und ist zu modifizieren. Die modifizierte Hypothese lautet: Nach Einsatz einer speziellen an Internetuser mit hohem Internet-Rechercheverhalten ausgerichteten Marketingkampagne, erreicht Faroo innerhalb eines Jahres eine Indexgröße, deren Quantität mit den konventionellen Suchmaschinen konkurrieren kann. Auch diese Hypothese bleibt durch weitere Versuche zu verifizieren.

Hypothese 3:
500.000 User sind in der Lage, innerhalb eines Jahres eine Indexgröße zu generieren, mit der 90 Prozent aller gestellten Suchanfragen beantwortet werden können.

Die Formulierung dieser Hypothese stellt nicht die Quantität der indexierten Webseiten in den Fokus. Vielmehr steht hier die „beantwortete Suchanfrage“ im Vordergrund. Eine Überprüfung der Hypothese erfolgt anhand der Versuchsergebnisse und wird durch weitere Überlegungen auf Basis der theoretisch erarbeiteten Grundlagen gestützt.

Suchanfragen sind mit Worten gleichzusetzen. Ist ein gesuchter Term im Index gespeichert, ist ebenfalls mindestens eine Webseite zu dieser Suchanfrage enthalten und die Suchanfrage kann damit theoretisch als beantwortet gelten. Eine Aussage über die Precision der als Antwort angebotenen Webseite bleibt an dieser Stelle außen vor.

Wie sich aus den Ergebnissen des Versuchs erkennen lässt, können bereits 25 User innerhalb eines Jahres 264.268 verschiedene Worte indexieren. Vergleicht man dieses Ergebnis z.B. mit der Anzahl der aktiven deutschen Worte (75.000) oder

---

[161] Vgl. http://www.sistrix.com/news/494-aol-daten-eine-kurze-auswertung.html (Abruf: 15.12.2006).

der Anzahl an Worten im Duden (130.000) bzw. im Brockhaus (260.000) wird erkennbar, dass eine Abdeckung zumindest des deutschen Wortschatzes in dieser Größenordung mit dem Faroo-Crawlingverfahren bereits durch 25 User innerhalb eines Jahres möglich ist. Im Fall der indexierten Worte wird der abflachende Bereich der Heaps-Law-Kurve also bereits im Versuchszeitraum erreicht. Das kann auf die homogene Zusammensetzung (=annähernd homogener Wortschatz) der Versuchsprobanden zurückgeführt werden. Im Vergleich zur vorigen Feststellung (Anzahl der indexierten Webseiten) liegt diese Feststellung (Anzahl indexierter Worte) ebenfalls nahe, da die gesamte Anzahl an Worten im Netz wesentlich niedriger ist als die gesamte Webseitenanzahl. Da weiterhin numerisch mehr Wörter als Seiten im Versuchszeitraum indexiert wurden, ist das Wachstum der Anzahl an Worten bereits weiter fortgeschritten als das Wachstum der Anzahl an Webseiten.

Um jedoch auch Homonyme und den möglichen richtigen Kontext abzudecken, gilt im Rahmen dieser Arbeit eine Suchanfrage als beantwortet, wenn zehn[162] verschiedene Suchergebnisse (=unterschiedliche Webseiten) angezeigt werden können. Dafür wird hier theoretisch die Anzahl an deutschen Worten im Brockhaus von 260.000 Einträgen verzehnfacht. Dementsprechend wäre ein Index von 2.600.000 (=unterschiedlichen Webseiten) in der Lage, alle Suchanfragen zu Wörtern der deutschen Sprache zu beantworten.

Stellt man die auf Basis des Versuchs ermittelte Indexgröße von 1,6 Milliarden Webseiten durch 500.000 Usern (vgl. Tabelle 10) gegenüber, wird deutlich, dass der Sachverhalt der aufgestellten Hypothese erfüllt wird und vermutlich sogar mehr als 90 Prozent aller Suchanfragen beantwortet werden können.

Gestärkt wird die Verifizierung der Hypothese durch weitere Überlegungen:

Ein Blick auf die Untersuchungen in Kapitel (II) 2.3.2 zeigt, dass innerhalb eines Jahres bei der Suchmaschine Lycos 11,2 Millionen und bei Fireball 6,2 Millionen „unterschiedliche“ Suchterme von den Usern verwendet wurden. In diesen Termmengen finden nicht nur Worte aus dem deutschen Wortschatz, sondern auch aus anderen Sprachen Beachtung. Angenommen, es sollten zehn unterschiedliche Treffer (Webseiten) pro Suchanfrage angezeigt werden, wäre dementsprechend eine Indexgröße von 112 Millionen beziehungsweise 62 Millionen unterschiedlichen

---

[162] Der Faktor zehn wurde aus zwei Gründen gewählt. Der erste Grund ist der Flexionsfaktor der deutschen Sprache von zehn (vgl. Kapitel (II) 5.2). Als zweiter Grund wurde das Selektionsverhalten der User in Ergebnisseiten (vgl. Kapitel (II) 2.3.3) als Grundlage verwendet.

Webseiten erforderlich. Wie Tabelle 10 zeigt, wird mit dem userzentrierten Crawlingverfahren bei 500.000 Usern eine Indexgröße von 1.6 Milliarden Webseiten erreicht. Hieraus lässt sich schlussfolgern, dass Faroo mit 500.000 Usern nach einem Jahr Laufzeit in der Lage wäre, mehr als 90 Prozent aller Suchanfragen zu beantworten. Dies gilt zumindest unter der Bedingung, dass der Durchschnittsuser nur die ersten zehn Treffer der Ergebnislisten als relevant einstuft.

Weitere Überlegungen machen deutlich, dass nicht zwingend mit dem Faktor zehn multipliziert werden muss. Ein Eintrag im Index entspricht einer Webseite. Diese besteht nicht nur aus einem Wort, sondern enthält Text, der sich immer aus mehreren Worten zusammensetzt. So enthält beispielsweise eine Webseite zum Suchbegriff „Cocktails“ ebenso Begriffe wie „Rezepte“, „Gläser“, „Zutaten“, „Longdrink“, „Früchte“, „Schnaps“, „Lounge“ und viele mehr. Dies verdeutlicht, dass bei der Suche nach einem dieser unterschiedlichen Begriffe rein theoretisch eine und die selbe Antwort geliefert werden kann. Es ist also eine Überschneidung der Informationen in den Webseiten vorhanden, so dass sich die erforderliche Indexgröße zur Beantwortung aller Suchanfragen verkleinert.

Die Hypothese gilt somit als verifiziert.

7 Weiterführende Überlegungen

Dieses Kapitel beleuchtet ergänzend, wie sich das Faroo-Crawlingverfahren von den herkömmlichen Verfahren unterscheidet. Diese Unterschiede der technischen Umsetzung haben ebenfalls Auswirkungen auf das Wachstum des Index, jedoch ist deren Umfang nur schwer messbar.

Bei Web-Katalogen erfüllen die Redakteure eine „Gatekeeperfunktion“, indem sie sich für oder gegen die Aufnahme einer Webseite in den Index entscheiden. Bei der Arbeitsweise von Robots ist kein Mensch unmittelbar an der Auswahl der Webseiten zur Indexierung beteiligt (vgl. Wolling 2005: 3ff.). Hingegen ist der Crawler - bzw. dessen Programmierer - für die Auswahl der Dokumente als Datenbasis zur Indexierung verantwortlich. Durch spezifische Programmierung der Crawler findet bereits bei der Datensammlung eine Selektion anhand vorgegebener Kriterien statt (vgl. Wolling 2005: 529ff.). Faroo ermöglicht hier eine Indexierung, sofern die Inhalte nicht passwortgeschützt, kostenpflichtig oder mit der robots.txt-Datei im Quellcode der Webseite versehen sind. Des Weiteren übernimmt bei Faroo der User das Crawlen. Somit ist es möglich, dass ein User Webseiten entdeckt, die für Crawler aus technischen Gründen nicht auffindbar sind. Die Informationen im Faroo-Index

unterscheiden sich in ihrer Art und Weise von denen, die von Webrobots gesammelt werden. Bei Crawlern von konventionellen Suchmaschinen wurde eine Selektion nachgewiesen, welche sich speziell im Bereich der Abdeckung bezüglich landesspezifischer Webinhalten niederschlägt. So werden US-Server von vielen Suchmaschinen-Crawlern bevorzugt aufgesucht und indexiert (vgl. Gartz 2000: 48). Darüber hinaus werden Tendenzen zum systematischen Ausschluss ganzer Domains von der Indexierung beobachtet (vgl. ebd.). Als Gründe sind hier nicht nur gewaltverherrlichende oder pornographische Inhalte zu nennen, sondern auch die Tatsache, dass nicht-kommerziellen Angeboten im privaten Bereich eine niedrigere Relevanz zugesprochen wird (vgl. ebd).

In den Indices der etablierten Suchmaschinen finden sich immer wieder Dead-Links, die möglichst schnell erkannt und eliminiert werden sollten, um die Qualität der Suchergebnisse nicht negativ zu beeinflussen. Mit einem konventionellen Robot-Verfahren ist es nicht einfach, die großen Datenbestände des Internets im Index auf dem aktuellsten Stand zu halten. Wie in Kapitel (II) 1.2 erläutert, unterliegen die Dokumente im WWW einer großen Dynamik. Studien zeigen, dass Robots einige Webseiten nur alle drei Monate besuchen und auf Änderungen überprüfen. In der Zwischenzeit können indes verschiedene Aktualisierungen oder gar Löschungen kompletter Sites stattgefunden haben. Bei der dezentralen Crawling-Technologie wird der Index jedoch direkt nach dem Aufrufen aktualisiert. Erhält Faroo die Fehlermeldung 404 wird der entsprechende Eintrag sofort aus dem Index gelöscht (vgl. Kapitel (II) 4.2) und bei Veränderung aktualisiert.

An dieser Stelle ist zu erwähnen, dass in dem P2P-Crawlingverfahren Potenzial für die Generierung eines Index mit hoher Aktualitätsrate liegt.[163] Anfänglich sind die User grundlegend für den Aufbau des Index verantwortlich, im weiteren Verlauf ebenso für dessen Aktualisierung und Erweiterung. Durch das redundante Aufrufen von Webseiten stellt nicht jede aufgerufene Webseite einen neuen Eintrag im Index dar und lässt den Index somit nicht weiter wachsen, hält ihn jedoch stets aktuell.

Eine weitere Möglichkeit, auf bestimmte Webseiten aufmerksam zu werden, ist eine Empfehlung von Freunden, Bekannten oder Arbeitskollegen. Hier wird oft die Webadresse ausgetauscht und in den eigenen Bookmarks abgelegt. Hinzu kommt, dass Webseiten auch über die direkte Eingabe von Webadressen im Browser, durch abgespeicherte Bookmarks oder Social-Bookmarking-Funktionen aufgerufen

---

[163] Vgl. http://www.faroo.com/home/deutsch/technology/description_architecture.html (Abruf: 18.12.2006).

werden können. Der Umfang dieser Seitenaufrufe konnte im Rahmen dieser Arbeit nicht ermittelt werden. Hierdurch kann der Index jedoch ebenfalls wachsen.
Es wird deutlich, dass der Index beim Crawlingverfahren von Faroo die Webseiten enthält, die von den Usern besucht werden. Die Verantwortung zum Aufbau, zur Erweiterung und Aktualisierung des Index wird beim dezentralen Crawlingverfahren nicht mehr dem Crawler in Form einer Softwareroutine überlassen. Stattdessen wird den Usern selbst ermöglicht, diese Verantwortung für die Indexierung ihrer Webseiten zu übernehmen.

III Fazit und Ausblick

Mit dieser Arbeit liegt eine **Untersuchung zum Wachstum eines verteilten Index einer Peer-to-Peer-Web-Suchmaschine** vor. Ziel dieser Arbeit war es, ein neuartiges Crawlingverfahren auf dessen Eignung zum Aufbau eines P2P-Suchmaschinenindex zu überprüfen.

Zusammenfassend zeigen der Versuch und die abschließende Diskussion, dass das userzentrierte Crawlingverfahren als „geeignet“ eingestuft wird; obwohl es nicht möglich ist, unter der Voraussetzung von einem einjährigen Marktbestehen und 500.000 bzw. 1.000.000 Usern einen zu den etablierten Suchmaschinenindices quantitativ konkurrenzfähigen Index zu generieren. Auch wenn die Quantität nicht erreicht wird, so wird jedoch die „Möglichkeit der Beantwortung von 90 Prozent aller Suchanfragen“ insofern erfüllt, als dass mindestens zehn Treffer pro Ergebnisliste geliefert werden können.

Ein Statement von Yahoo unterstreicht die Aussage, dass nicht die Quantität, sondern die Beantwortung der Suchanfragen der User, und somit die Befriedigung des Informationsbedürfnisses, das ausschlaggebende Kriterium zur Konkurrenzfähigkeit darstellt.

> *"We congratulate Google on removing the index size number from its homepage and recognizing that it is a meaningless number. As we've said in the past, what matters is that consumers find what they are looking for […]"*[164]

Bei näherer Betrachtung dieses Statements wird deutlich, dass der User im ersten Moment hinter einer hohen Indexgröße eine umfassendere Informationsmenge vermutet und damit eine höhere Wahrscheinlichkeit der Beantwortung seines Informationsbedürfnisses assoziiert. Die Indexgröße kann jedoch nicht als alleiniges Qualitätskriterium angesehen werden. Zur qualitativen Beurteilung müssen noch weitere Kriterien herangezogen werden: Hier sind vor allem Vollständigkeit (Abdeckung der im Internet vorhandenen Webseiten) und Aktualität eines Index zu nennen (vgl. Lewandowski/Wahlig/Meyer-Bautor 2006). Eine umfangreiche Informationsmenge stellt jedoch auch eine anspruchsvollere Herausforderung an den Selektierungs- und Rankingprozess zur präzisen Suchanfragenbeantwortung dar. Gekoppelt mit der Personalisierung durch die Desktop-Suche und den dadurch generierten personalisierten Index können Suchanfragen der User, auch aus einer

---

[164] Vgl. Battelle, John (2005) online unter: http://battellemedia.com/archives/2005/09/google_announce.php_test.php (Abruf: 20.01.2007).

umfangreicheren Informationsmenge eine höhere Genauigkeit erreichen. Eine höhere Aktualität des Index und eine umfassendere Abdeckung der im WWW verfügbaren Dokumente kann ebenfalls erzielt werden.

Ein Zusammenhang zwischen Wachstum des verteilen Index der P2P-Suchmaschine Faroo und der Zeit sowie der Useranzahl wurde nachgewiesen. Je länger der Zeitraum ist und je mehr User partizipieren, desto mehr Webseiten werden im Index gespeichert und abrufbar sein. Vorausgesetzt, dass 500.000 User zur Teilnahme an Faroo gewonnen werden, würde das Wachstum des Index seinen größten Schub am Anfang haben und die meistnachgefragten Webseiten relativ schnell im Index gespeichert sein. Wie Heaps Law (vgl. Kapitel (II) 5.2) den Kurvenverlauf bei der Neuentdeckung von Worten bei der Zunahme der Textmenge beschreibt, wird die Abdeckung des Long-Tail (der im Internet weniger stark nachgefragten Webseiten) erst zu einem späteren Zeitpunkt stattfinden. Der Verlauf des Wachstums sollte den angestellten Überlegungen nach, über einen längeren Zeitraum betrachtet, gemäß der Heaps-Law-Kurve verlaufen.

Die pauschale Annahme: „mehr User generieren mehr Inhalt“, stimmt nur bis zu einer „gewissen“ im Rahmen dieser Arbeit nicht bestimmbaren Anzahl an Usern. Denn mit steigender Userzahl steigt auch die Redundanz sowohl bei der Auswahl der Webseiten als auch bei der Eingabe von Suchbegriffen (Worten). Hieraus lässt sich ableiten, dass eine hohe Anzahl an Usern nicht zwingend benötigt wird, um einen Suchmaschinenindex aufzubauen, der mehr als 90 Prozent aller gestellten Suchanfragen beantworten kann. Eine hohe Anzahl wäre jedoch erforderlich, um 100 Prozent aller Suchanfragen beantworten zu können.

Aus diesen Ergebnissen sollen Handlungsempfehlungen abgeleitet werden: Eine wesentliche Bedeutung hat der Einsatz eines ansprechenden und zielgruppenorientierten Marketings. Des Weiteren gilt es, das Vertrauen der User in die P2P-Technologie der Suchmaschine zu gewinnen.

Besondere Bedeutung erlangt die Gewinnung der User mit hoher und vielfältiger Internetnutzung. Ebenfalls ist es wichtig, auch die Personen zu akquirieren, die „Nischen-Interessen“[165] aufweisen. Bei Markteintritt ist darauf zuachten, diese

[165] Nischen-Interessen sind spezielle Themengebiete, die nur von einer geringen Useranzahl nachgefragt werden.

speziellen Zielgruppen durch entsprechende zielgruppenorientierte Marketing-aktionen zu erreichen.[166]

Im Folgenden werden weitere Vermutungen und Hypothesen aufgestellt, die weitere Forschungsprozesse anstoßen können.

- Es ist zu vermuten, dass Motivationsanreize[167] zu einem schnelleren Wachstum des Index beitragen.

- Das innovative Crawlingverfahren ermöglicht (ab Erreichen der kritischen Masse) einen Suchmaschinenindex mit einer annähernden „just-in-time“[168] Aktualitätsrate.

Erreicht Faroo die kritische Masse und erlangt darüber hinaus eine weltweite Verbreitung, wird mit dem Crawlingverfahren eine annähernd vollständige Abdeckung des Internets erreicht. Der Verlauf des tatsächlichen Index-Wachstums bleibt abzuwarten und wird weitere Erkenntnisse liefern.

---

166 Weiterführende Informationen: Gladwell, Malcom (2002): Tipping Point, 3. Auflage; München: Wilhelm Goldmann Verlag.

167 Motivationsanreizen können hierbei emotional, monetär oder ideologisch sein.

168 Just-in-time bedeutet im oben stehenden Kontext: Das genau in dem Zeitpunkt in dem eine Änderung auf einer Webseite erfolgt, fast synchron eine Aktualisierung im Faroo-Index möglich ist.

Literaturverzeichnis

AGOF e. V. (2006): Internet Facts. Frankfurt. Online unter: http://www.agof.de/die_internet_facts.353.html, letzter Abruf: 20.01.2007.

Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier (2006): Modern Information Retrieval. Harlow: Pearson, Addison-Wesley.

Bar-Yossef, Ziv; Gurevich, Maxim (2006): Random Sampling from a Search Engine's Corpus. Online unter: http://www.ee.technion.ac.il/people/zivby/papers/se/se.techreport.pdf, letzter Abruf: 20.01.2007.

Beiler, Markus; Zenker, Martin: Die wachsende Macht von Suchmaschinen im Internet: Auswirkungen auf User, Medienpolitik und Medienbusiness. Wissenschaftlicher Workshop und Konferenz, Juni 2006, Berlin. Online unter: http://www.uni-leipzig.de/~journ/suma/pdf/Zusammenfassung_Suma-Tagung.pdf, letzter Abruf: 20.01.2007.

Bernecker, Michael (2002): Kundenbindung im Internet. In: Conrady, Roland; Jaspersen, Thomas; Pepels, Werner (Hg.): Online-Marketing-Instrumente. Angebot, Kommunikation, Distribution, Praxisbeispiele. Neuwied: Luchterhand, S. 342-357.

Bernhardt, Ute (2003): Filtern, Sperren, Zensieren? Vom Umgang mit unliebsamen Inhalten im Internet. In: Schulzki-Haddouti, Christiane (Hg.). Bürgerrechte im Netz. Bonn: Bundeszentrale für politische Bildung, S.319-335.

Bortz, Jürgen; Döring, Nicola (2002): Forschungsmethoden und Evaluation. Für Human- und Sozialwissenschaftler. 3. überarbeitete Aufl.. Berlin, Heidelberg, New York (USA), Heidelberg: Springer.

Brockhaus (2003): Fachlexikon: Computer und Informationstechnologie. Mannheim: Brockhaus.

Broder, Andrei (2002): A taxonomy of web search. Online unter: http://www.acm.org/sigir/forum/F2002/broder.pdf, letzter Abruf: 20.01.2007.

Neymanns, Harald (2005): Suchmaschinen: Das Tor zum Netz. Berlin: Bündnis 90; Online unter: http://www.g-bettin.de/cms/files/dokbin/63/63516.reader_1564_suchmaschinendas_tor_zum_net.pdf, letzter Abruf: 20.01.2007.

Cachelogic (2005): Peer-to-Peer in 2005; Online unter: http://www.cachelogic.com/home/pages/research/p2p2005.php, zuletzt geprüft am 20.01.2007.

Cloer, Thomas (2006): Google muss belgische Zeitungsberichte aus News entfernen. Online unter: http://www.computerwoche.de/nachrichten/581496/#, letzter Abruf: 20.01.2007.

Dielkmann, Peter (2006): Vergangenheit und Zukunft der Suchmaschine. Von Veronica zur "sozialen Suche". Online unter: http://www.tagesschau.de/aktuell/meldungen/0,1185,OID5242340_NAV_BAB,00.html, letzter Abruf: 20.01.2007.

Dietl, Helmut; Royer, Susanne (2000): Management virtueller Netzwerke in der Informationsökonomie. In: zfo, Jg. 69, Nr.6, S324-331

Duden (2006): Duden Deutsches Universalwörterbuch. Mannheim: Dudenverlag.

Dustdar, Schahram; Gall, Harald; Hauswirth, Manfred (Hg.) (2003): Software-Architekturen für verteilte Systeme. Prinzipien, Bausteine und Standardarchitekturen für moderne Software; Berlin: Springer.

ECIN (2006): Suchmaschinen sind wichtiges Nadelöhr. Forschungsinstitut für Telekommunikation. Online unter: http://www.ecin.de/news/2006/09/22/09943/index.html, letzter Abruf: 20.01.2007.

Eimeren, Birgit; Gerhard, Heinz; Frees, Beate (2004): Internetverbreitung in Deutschland: Potenzial vorerst ausgeschöpft? In: Media Perspektiven Nr. 8, S. 350-370. Online unter: http://www.daserste.de/service/ardonl04.pdf letzter Abruf: 20.01.2007.

Eimeren, Birgit; Frees, Beate (2005): Nach dem Boom: Größter Zuwachs in internetfernen Gruppen. In: Media Perspektiven Nr. 8, S. 362-379. Online unter: http://www.daserste.de/service/ardonl05.pdf, letzter Abruf: 20.01.2007

Faller, Heike (2005): David gegen Google. Wie ein Mathematiker aus New Jersey die größte Suchmaschine der Welt übertrumpfen will: Eine Reise ins Recht der Algorithmen. In: Die Zeit, Jg. 2005, Ausgabe 41, 2005.

Ferber, Reginald (2003): Information Retrieval. Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web. Heidelberg: dpunkt-Verlag

Fisch, Martin; Gscheidle, Christoph (2006): Onliner 2006: Zwischen Breitband und Web 2.0. Ausstattung und Nutzungsinnovation. In: Media Perspektiven, Nr. 8, S.431-440; Online unter: http://www.daserste.de/service/studie.asp, letzter Abruf: 20.01.2007.

Fiutak, Martin (2005): Peer-to-Peer wird Alternative zu Client-Server-Systemen. Forscher sprechen von "Ferrari der Online-Kommunikation". Online unter: http://www.zdnet.de/news/tkomm/0,39023151,39134180,00.htm, letzter Abruf: 20.01.2007.

Frascaria, Kareen (2002): Peer-to-Peer: Die Erneuerung des verteilten Rechnens. CNET Networks Deutschland GmbH. München. Online unter: http://www.zdnet.de/itmanager/tech/0,39023442,2107183,00.htm, letzter Abruf: 20.01.2007.

Fuchs-Kittowski, Klaus; Schewe, Tankred (2002): Informationsverarbeitung, -recherche und - erzeugung in den Biowissenschaften. In: Wissenschaft und Innovation: Wissenschaftsforschung Jahrbuch 2001; (Hg.) Parthey, Heinrich u. Spur, Günther. Berlin: Gesellschaft für Wissenschaftsforschung 2002. S. 185–220.

Garbe, Wolf (2001): Bingooo - Die Transformation des World Wide Web zur virtuellen Datenbank. In: Wirtschaftsinformatik, Jg. 43, Nr. 5, S. 511–515. Online

unter: http://www.wirtschaftsinformatik.de/wi_artikel.php?sid=823, letzter Abruf: 20.01.2007.

Gartz, Joachim (2000): Professionelle Internetrecherche für Wissenschaftler und Online-Profis. Kilchberg: SmartBooks.

Gilder, George (2006): The Information Factories. Wired; Issue: 14.10, Online unter: http://www.wired.com/wired/archive/14.10/cloudware.html, letzter Abruf: 20.01.2007)

Glögger, Michael (2003): Suchmaschinen im Internet. Funktionsweise, Ranking Methoden, Top Positionen. Berlin, Heidelberg, New York (USA): Springer.

Griesbaum, Joachim (2003): Unbeschränkter Zugang zu Wissen? Leistungsfähigkeit und Grenzen von Suchdiensten im Web. Zwischen informationeller Absicherung und manipulierter Information. In: Online-Tagung; ComInfo; Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis (Hg.): Competence in content. Frankfurt am Main: S. 37–50.

Griesbaum, Joachim; Bekavac, Bernard (2004): Web-Suche im Umbruch? Entwicklungstendenzen bei Websuchdiensten. In: Internationales Symposium für Informationswissenschaft (Hg.): Information zwischen Kultur und Marktwirtschaft. Proceedings des 9. Internationalen Symposiums für Informationswissenschaft, Konstanz: UVK-Verl.-Ges.; Information Research, Vol. 9, Nr.4, S. 283–299.

Grietje, Bettin (2005): Suchmaschinen: Das Tor zum Netz, Pressemitteilung, Nr. 220 online unter: http://www.g-bettin.de/cms/default/dok/63/63512. suchmaschinen_das_tor_zum_netz_eine_info.htm, letzter Abruf: 20.01.2007.

Gscheidle, Christoph; Fisch, Martin; In: Media Perspektiven (2005): Der Einfluss der Computerausstattung auf die Internetnutzung. In: Media Perspektiven Nr. 11, S. 570-581.

Hauswirth, Manfred; Dustar, Schahram (2005): Peer-to-Peer: Grundlagen und Architektur. In: Datenbank Spektrum, Jg. 5., Nr. 13, S. 5–13.

Heaps, Harold Stanley (1978): Information Retrieval. Computational and Theoretical Aspects. London: Academic Press.

Hölscher, Christoph (2002): Die Rolle des Wissens im Internet: gezielt suchen und kompetent auswählen. Stuttgart: Klett-Cotta

Ihlenfeld, Jens (2006): Google setzt BMW vor die Tür. Golem.de. Online unter: http://www.golem.de/0602/43155.html, letzter Abruf: 20.01.2007.

Karzauninkat, Stefan (2002): Die Suchfibel. Wie findet man Informationen im Internet? Stuttgart: Ernst Klett Verlag.

Karzauninkat, Stefan; Alby, Tom (2006): Suchmaschinenoptimierung. Professionelles Website-Marketing für besseres Ranking. München: Hanser.

Kim, Amy Jo (2000): Community building on the Web. Secret strategies for successful online communities. Berkeley Californien: Peachpit.

Krömker, Heidi; Klimsa, Paul (Hg.) (2005): Handbuch Medienproduktion. Produktion von Film, Fernsehen, Hörfunk, Print, Internet, Mobilfunk und Musik. Wiesbaden: VS Verl. für Sozialwiss.

Kühl, Stefan (Hg.) (2005): Quantitative Methoden der Organisationsforschung. Ein Handbuch. Wiesbaden: VS Verl. für Sozialwissenschaften.

Lehmann, Kai; Schetsche, Michael (Hg.) (2005): Die Google-Gesellschaft. Vom Wandel des digitalen Wissens. Bielefeld: Transcript.

Lessig, Lawrence (2005): Free Culture, The Nature and Future of Creativity; Penguin Goup USA

Lewandrowski, Dirk (2005): Web Information Retrieval. In: Information: Wissenschaft und Praxis, Jg. 2005, Nr. 56, S. 5–12.

Lewandowski, Dirk (2006): Query Types and Search Topics of German Web Search Engine Users. [Preprint]. In: Information Services & Use, Jg. 26. Online unter: http://www.durchdenken.de/lewandowski/doc/isu2006_preprint.pdf, letzter Abruf: 20.01.2007.

Lewandrowski, Dirk (2006a): Aktualität als erfolgskritischer Faktor bei Suchmaschinen, In: Information: Wissenschaft und Praxis, Jg. 2006, Nr. 57/3, S. 141-148. Online unter: http://www.durchdenken.de/lewandowski/ doc/Aktualitaet_IWP.pdf (Abruf: 20.01.2007).

Linde, Frank (2005): Ökonomie der Information. Göttingen: Universitätsdrucke Göttingen.

Loban, Bryan (2004): Between rhizomes and trees: P2P information systems. In: FirstMonday, Nr. 10. Online unter: http://firstmonday.org/issues/issue9_ 10/loban/, letzter Abruf: 20.01.2007.

Machill, Marcel; Welp Carsten (Hg.) (2003): Wegweiser im Netz. Qualität und Nutzung von Suchmaschinen. Gütersloh: Bertelsmann-Stiftung.

Machill, Marcel; Beiler (2006): Internet-Suchmaschinen als neue Herausforderung für die Medienregulierung: Jugendschutz und publizistische Vielfalt als Fallbeispiele für Governance Issues. Zürich: Publizistikwissenschaft und Medienforschung.

Mauthe, Andreas; Heckmann, Oliver (2005): Distributed Computing - GRID Computing. In: Steinmetz, Ralf; Klaus, Wehrle (Hg.): Peer-to-Peer Systems and Applications. State-of-the-Art Survey. Berlin, Heidelberg, S. 193–206.

Mehta, Stephanie N. (2006): Behold the server farm. Online unter: http://money.cnn.com/2006/07/26/magazines/fortune/futureoftech_serverfarm.fortune/index.htm, letzter Abruf: 20.01.2007.

Minar, Nelson; Hedlund, Marc (2001): A Network of Peers. Peer-to-Peer Models Through the History. In: Oram, Andy (Hg.): Peer-to-peer. Harnessing the benefits of a disruptive technology. Beijing: O'Reilly, S. 3–20.

Ntoulas, Alexandros; Cho, Junghoo; Olston, Christopher (2004): What's New on the Web? The Evolution of the Web from Search Engin Perspective. New York, USA. Thirteenth WWW Conference. Online unter: http://www2004.org/proceedings/docs/1p1.pdf, letzter Abruf: 20.01.2007.

Oram, Andy (Hg.) (2001): Peer-to-peer. Harnessing the benefits of a disruptive technology. Beijing: O'Reilly.

ORF.at (2006): Google Orkut vs Brasilien. futurezone. Online unter: http://futurezone.orf.at/it/stories/139712/, letzter Abruf: 20.01.2007.

Panzer, Volker (2006): Big Google - die geheime Macht der Suchmaschinen (ZDF Nachtstudio). Ausgestrahlt am 17.09.2006. ZDF. Online unter: http://www.zdf.de/ZDFde/inhalt/28/0,1872,3974652,00.html, letzter Abruf: 20.01.2007.

Patzwaldt, Klaus: 3. Tagung des SuMa-eV in Berlin. Online unter: http://www.at-web.de/blog/20060929/3-tagung-des-suma-ev-in-berlin.htm, letzter Abruf: 20.01.2007.

Rabe, Lars (2006): Suchmaschinen Marketing im Überblick. In: Handelsblatt 11.09.2006; Online verfügbar unter: http://www.handelsblatt.com/news/ Technologie/Suchmaschinen-Marketing/_pv/_p/302127/_t/ft/_b/1131920/ default.aspx/suchmaschinen-marketing-im-ueberblick.html, letzter Abruf: 20.01.2007.

Rohwedder, Wulf (2005): Die virtuelle chinesische Mauer. Online unter: http://www.tagesschau.de/aktuell/meldungen/0,1185,OID3939762,00.html, letzter Abruf:20.01.2007.

Rose, D. E.; Levinson, D. (2004). Understanding user goals in Web search. Thirteenth International World Wide Web Conference Proceedings, WWW 2004, 13-19.

Rossig, Wolfram; Prätsch Joachim (2005): Wissenschaftliches Arbeiten. Hamburg: Weyhe.

RRZN - Regionales Rechenzentrum für Niedersachsen (2001): Suchen & Finden im Internet. oder: Die Nadel im Heuhaufen. Universität Hannover.

Schmidt-Mänz, Nadine; Bomhardt, Christian (2005): Wie suchen Onliner im Internet? Online unter: http://www.absatzwirtschaft.de/pdf/sf/Maenz.pdf, letzter Abruf: 20.01.2007.

Schmidt-Mänz, Nadine (2006): Erkenntnisse aus dem Suchverhalten im Web. "Muster in Suchanfrage". Fachkonferenz: Suchen und Finden im Internet. München: Veranstalter: Münchner Kreis; Übernationale Vereinigung für Kommunikationsforschung e.V. Online unter: http://www.muenchner-

kreis.de/pdfs/SuchenundFinden/Schmidt-Maenz.pdf, letzter Abruf: 20.01.2007.

Schmitz, Henrik (2006): Aufklärung statt Nazi-Hetze. In: Frankfurter Rundschau; FR-online.de, Online unter: http://www.fr-aktuell.de/in_und_ausland/ multimedia/aktuell/?em_cnt=979344, letzter Abruf: 20.01.2007.

Schmücker, Jörg; Müller; Wolfgang (2003): Praxiserfahrungen bei der Einführung dezentraler Wissensmanagement-Lösungen. In: Wirtschaftsinformatik, Jg. 45., Nr. 3, S. 307–311.

Schnell; Hill; Esser (2005): Methoden der empirischen Sozialforschung. 7. Aufl. München: Oldenbourg.

Schoder, Detlef; Fischbach, Kai; Teichmann, René (Hg) (2002): Peer-to-Peer. Ökonomische, technologische und juristische Perspektiven. Berlin, Heidelberg, New York (USA).

Schollmeier, Rüdiger (2001): A Definition of Peer-to-Peer Networking for the Classification of Peer-to-Peer Architectures and Applications. In: Proceedings of the IEEE 2001 International Conference on Peer-to-Peer Computing. Linköping, Sweden.

Schollmeier, Rüdiger (2005): Signaling and Networking in unstructured Peer-to-Peer Networks. Dissertation. TH München.

Schwärze, Marcus (2004): Jedermanns Suchmaschine. In: Hannoverschen Allgemeinen Zeitung, 2004, S. 23. Online unter: http://suma-ev.de/downloads/yacy-haz.pdf, letzter Abruf: 20.01.2007.

Shapiro, Carl; Varian, Hal R. (1999): Information rules. A strategic guide to the network economy. Boston Mass.: Harvard Business School Press.

Sherman, Chris; Price, Gary (Hg.) (2002): The invisible web. Uncovering information sources search engines can't see. Medford NJ: Information Today

Shirky, Clay (2001): Listening to Napster. In: Oram, Andy (Hg.): Peer-to-peer. Harnessing the benefits of a disruptive technology. Beijing: O'Reilly, S. 21–37.

Sietmann, Richard (2005): Wider die Monokultur. P2P-Strategien gegen die Suchmaschinen-Monopolisierung. In: c't, S. 52. Online unter: http://www.heise.de/ct/05/16/052/, letzter Abruf: 20.01.2007.

Statistisches Bundesamt (2005): Informationstechnologie in Unternehmen und Haushalten 2005. Online unter: http://www.destatis.de/download/d/veroe/ Tabellenanhang_Haushalte_IKT2005.pdf, letzter Abruf: 2007.

Steinmetz, Ralf; Wehrle, Klaus (Hg.) (2005): Peer-to-Peer Systems and Applications. State-of-the-Art Survey. Berlin, Heidelberg.

Stieler, Wolfgang (2006): Die Kultur des Mitmachens. Interview mit Bradley Horowitz Yahoo. In: Technology Review, Jg. 2006.

Stock, Wolfgang G. (2000): Informationswirtschaft. Management externen Wissens. München: Oldenbourg (Managementwissen für Studium und Praxis).

Stock, Mechthild; Stock, Wolfgang G. (2004): Trend des Jahres: Kooperation und Konkurrenz auf Märkten elektronischer Informationsdienste: Mit dem Wettbewerber zusammenarbeiten?; In: Password 01/2004; Bredemeier: Hattingen/Ruhr

Stock, Wolfgang G.; Lewandowski, Dirk (2005): Suchmaschinen und wie sie genutzt werden. Preprint: WISU 35(2006); 8-9, 1078-1083. Online unter: http://www.durchdenken.de/lewandowski/doc/wisu_preprint.pdf, letzter Abruf: 20.01.2007.

Strauch, Dietmar; Kuhlen, Rainer; Laisiepen, Klaus (2004): Grundlagen der praktischen Information und Dokumentation; Handbuch: Bd. 1; München: K G Saur

Strauch, Dietmar; Kuhlen, Rainer; Laisiepen, Klaus (2004a): Grundlagen der praktischen Information und Dokumentation; Glossar: Bd. 2; München: K G Saur

Thiedeke, Udo (Hg.) (2003): Virtuelle Gruppen. Charakteristika und Problem-dimensionen. 2., überarbeitete und aktualisierte Auflage. Wiesbaden: Westdeutscher Verlag

Thiele, Frédéric Phillip; Speck, Hendrick (2004): Suchmaschinenpolitik - Google is watching you! Online unter: http://www.ccc.de/congress/2004/fahrplan/files/461-94-suchmaschinen-politik-paper.pdf, letzter Abruf: 20.01.2007.

Wilkens, Andreas (2006): AOL wegen Veröffentlichung von Suchanfragen verklagt. Heise online. Online unter: http://www.heise.de/newsticker/meldung/76474, letzter Abruf: 20.01.2007.

Wolling, Jens (2005): Suchmaschinen? - Selektiermaschinen? In: Krömker, Heidi; Klimsa, Paul (Hg.): Handbuch Medienproduktion. Produktion von Film, Fernsehen, Hörfunk, Print, Internet, Mobilfunk und Musik. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 529–537.

Zerdick, Axel (2001): Die Internet-Ökonomie. Strategien für die digitale Wirtschaft. Berlin: Springer

Homepages[169]

AllPeers. http://www.allpeers.com

Berliner Beauftragter für Datenschutz und Informationsfreiheit. http://www.datenschutz-berlin.de/

Bigchampagne. http://www.bigchampagne.com

Bundesamt für Sicherheti in der Informationstechnik. http://www.bsi-fuer-buerger.de/suchmaschinen/index.htm

Chaos Computer Club. http://www.ccc.de/congress/

Die Suchfibel. http://www.suchfibel.de/

Directory of open access journals. http://www.doaj.org

Duden. http://www.duden.de

Faroo. http://www.faroo.com[170]

Fight Aids at home. http://fightaidsathome.scripps.edu

Gemeinnütziger Verein zur Förderung der Suchmaschinen-Technologie und des freien Wissenszugangs. http://suma-ev.de

Gesellschaft für deutsche Sprache. http://www.gfds.de/

Google. http://www.google.com

Googlewhacking. http://www.googlewhacking.de

Gridcafe. http://gridcafe.web.cern.ch/gridcafe

Heise Online. http://www.heise.de/

Institut für Deutsche Sprache. http://www.ids-mannheim.de

Internet World Stats. http://www.internetworldstats.com

ITWissen. Online Lexikon. http://www.itwissen.info

Macwelt. http://www.macwelt.de

Münchener Kreis. http://www.muenchner-kreis.de

Netcraft. http://news.netcraft.com

Netzpolitik. Weblog. http://www.netzpolitik.org/

Netzzeitung. http://www.netzeitung.de

Open Directory Project. http://www.dmoz.de

OpenNet Initiative. http://www.opennetinitiative.org/

O'Reilly Networks. http://www.oreillynet.com

Planetmath. http://planetmath.org/encyclopedia

---

169 Die Adressen der aufgeführten Hompages von Organisationen, Personen und Projekten wurden zuletzt am 20.01.2007 auf ihre Gültigkeit geprüft.

170 Die Faroo Homepage war zum Zeitpunkt als diese Diplomarbeit verfasst wurde noch nicht für die Öffentlichkeit zugänglich. Die Webseiten sind deshalb als PDF-Dokument auf der beigefügten CD einsehbar.

Searchengineland. Weblog. http://searchengineland.com/

Searchenginewatch. http://searchenginewatch.com/

Seekport. http://www.seekport.de/q?liveseek

SinnerSchrade. http://www.nexttenyears.de

Sistrix. http://www.sistrix.com/news/

Suchmaschinen Magazin. http://www.at-web.de/ und http://www.at-web.de/blog/E10

The Search Engine Marketing Glossary. http://www.seobook.com/glossary/

Webhits. http://www.webhits.de/

Whatis. Online Lexikon. http://whatis.com/

Wikipedia (deutsch und englisch). http://de.wikipedia.org

Wired News. http://www.wired.com/

Wortschatz Portal. http://wortschatz.uni-leipzig.de/

Eidesstattliche Erklärung

Hiermit erkläre ich, Britta Jerichow, an Eides Statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungskommission vorgelegt und auch nicht veröffentlicht.

Köln, 14. Mai 2007