Unbeschränkter Zugang zu Wissen?
Leistungsfähigkeit und Grenzen von Suchdiensten im Web. Zwischen informationeller Absicherung und manipulierter Information

Joachim Griesbaum
Universität Konstanz
Informationswissenschaft
Fach D 87
D-78457 Konstanz
Joachim.Griesbaum@uni-konstanz.de

Zusammenfassung
Das Internet bietet Zugriff auf Informationen in bisher nicht bekannten Ausmaß. Die Nutzer beschränken sich bei der Suche nach Informationen allerdings auf einige wenige dominierende Suchdienste. Fraglich ist, ob und inwieweit diese Dienste geeignete Instrumente der informationellen Absicherung darstellen. Um dies in einer ersten Näherung zu beantworten werden die grundlegende Leistungsfähigkeit der dominierenden Suchdienste anhand einer Analyse der prinzipiellen Verfahren bzw. Kriterien der Indexierung und Ergebnissortierung skizziert. Als Ergebnis lässt sich festhalten, die populären Suchdienste bieten Zugriff auf eine Vielzahl von Informationen im Netz, sie sind allerdings nicht in der Lage die gesamten Wissensbestände zu erfassen. Die Suchergebnisse unterliegen zumindest z.T. dem Einfluss kommerzieller Interessen, die sich auch in den als „neutral“ und „objektiv“ geltenden Suchergebnissen von Suchmaschinen als unsichtbares Sortierungskriterium zumindest mittelbar auswirken. Erste Ideen die Qualität dieser Suchdienste als Instrumente der informationellen Absicherung zu erhöhen, bestehen in der Offenlegung dieser Problemfelder und Grenzen sowie in der Integration von (Deep-Web) Suchdiensten, die helfen, auf nicht erschlossene bzw. nicht erschließbare Inhalte dennoch zugreifen zu können.

1. Google und Co.: Pfadfinder oder Flaschenhälse im Datendschungel des Internets?

Mit Hilfe des Internet ist inzwischen für einen Großteil der Bevölkerung[1] Zugriff auf Information in bisher nicht gekanntem Ausmaß möglich. Die Chancen zur informationellen Absicherung, verstanden als die Summe der Möglichkeiten sich umfassend informieren zu können und damit die eigenen Entscheidungen bestmöglich abzusichern, waren damit theoretisch nie so groß wie heute.

Zur Orientierung in den riesigen Datenräumen[2] steht eine Vielzahl verschiedenster Suchdienste als Wegweiser bereit, um aus den Wissensbeständen des Netzes die tatsächlich benötigte Information zu erschließen. OPACs[3] ermöglichen die umfassende Recherche in den Beständen öffentlicher Bibliotheken. Professionelle Informationsanbieter ermöglichen, i.d.R. kostenpflichtig, Zugriff auf die Inhalte fachspezifischer Datenbanken.[4] Eine große Zahl allgemeiner oder spezialisierter

[1] Nach Angaben einer Studie von @facts E-Commerce sind im Januar 2003 rund 52% der Bundesbürger online. [http://www.welt.de/data/2003/01/16/33293.html] 09.03.03. Die Studie findet sich unter [http://www.atfacts.de/001/pdf_studies/atfacts_extra_eCommerce_200203.pdf] 10.03.03.
[2] Der Internet Domain Survey Report des Internet Software Consortiums vom Januar 2003 verzeichnet rund 170 Millionen Server, die Informationen und Dienste im Internet bereit zu stellen [http://www.isc.org/ds/WWW-200301/index.html] 09.03.03.
[3] Online Public Access Catalog.
[4] Beispielweise bietet die Dialog Corporation, einer der weltweit führenden Informationsanbieter, online Zugriff auf eine Vielzahl fachspezifischer Datenbanken [http://library.dialog.com/bluesheets/html/blf.html] 10.03.03.

Suchmaschinen[5] und Webverzeichnisse erlauben unmittelbaren Zugriff auf Inhalte des World Wide Web.[6]

Faktisch beschränkt sich der übliche Gebrauch von Suchdiensten allerdings auf einige wenige marktführende, kommerziell ausgerichtete Suchmaschinen und Webverzeichnisse. Als führend können insbesondere Google, Yahoo, Msn, Lycos, Altavista bezeichnet werden.[7] Diese dominieren als Suchinstrumente selbst in Recherchekontexten - wie z.B. der Suche nach wissenschaftlicher Information - in denen systematischere und effektivere Suchwerkzeuge zur Verfügung stehen.[8]

Die Möglichkeiten der Nutzer zur informationellen Absicherung werden also in vielen Fällen de facto von Google & Co. präterminiert. Nichts spricht die Macht und Verantwortung des Marktführers Google klarer aus als die Tatsache, dass im amerikanischen Sprachgebrauch zunehmend „googling“ als allgemeine Bezeichnung für Suchen im Internet verwendet wird.[9] Diese Sprachentwicklung benennt offen die Tatsache, dass die führende Suchmaschine Google von einer Großzahl der Nutzer als der zentrale Pfadfinder zu den Inhalten des Netzes eingestuft wird. Unklar bleibt, inwieweit dies tatsächlich zutrifft, bzw. zutreffen kann.

Beschreibt „googling“ den Tatbestand, dass der bequeme und einfache Zugriff auf Informationen via einiger prominenter Websuchdienste den Nutzer letztlich zu einem eingeengten Verhaltensspektrum führt, bzw. im Sinne einer Selbstbeschränkung dazu verleitet, andere, potenziell adäquatere Informationsquellen, zunehmend auszublenden? Bestimmen die Suchmaschinen damit die herrschenden Meinungen?[10] Führt so der neue Informationsreichtum letztlich gar zu einer neuen Informationsarmut? Ob und inwieweit und in welchen Kontexten diese plakativ-provokante Fragestellung zutrifft, kann hier nicht beantwortet werden. Wichtig ist es festzuhalten, dass das theoretische Potenzial des Internet, die informationelle Absicherung zu verbessern, durch die Dominanz weniger Suchdienste faktisch ins Gegenteil degenerieren kann: zu einer beschränkten, vorsortierten und unter Umständen manipulierten Sicht der Welt.

Im Folgenden wird versucht den aktuellen Stand der marktführenden kommerziellen Suchdienste zu skizzieren. Eine grundlegende Analyse der Arbeitsweise und Grenzen dieser Dienste soll Hinweise zur Einordnung dieser Dienste liefern. Ziel ist es, die Fähigkeiten aber auch die Beschränkungen und Manipulationspotenziale offen zu legen.

2. Methodik

[5] Beispielsweise auf wissenschaftliche Inhalte ausgerichtete Suchmaschinen wie [www.scirus.com] oder [www.campussearch.de], auf Online-Shops spezialisierte Kataloge wie [www.shop-netz.de], usw. 09.03.03.

[6] Übersichten über solche Suchdienste bieten Suchdiensteverzeichnisse wie [http://dir.yahoo.com/Computers_and_Internet/Internet/World_Wide_Web/Searching_the_Web/] oder [http://lii.org/search/file/netsearch].

[7] Ein genaues Ranking ist schwierig, da verschiedene Quellen im Internet zum Teil veraltet sind, als nicht repräsentativ gelten und/oder auf Angaben beruhen, die ähnliche aber nicht deckungsgleiche Werte wie Marktreichweite präsentieren. Bezogen auf die Informationsangebote des World Wide Web gilt i.d.R.: „Nur zwei dutzend Suchdienste sind für mehr als 90% des Traffic verantwortlich, der von Suchdiensten auf Ihre Seite geleitet wird.“ Klaus Schallhorn, Suchmaschinen bringen Besucher, [http://www.at-web.de/Informationen/suchmaschinen-refers.htm] 12.03.03. Statistische Angaben zu Markt- und Besucherreichweite von Suchdiensten finden sich unter Jupiter MMXI European Search Engine Ratings April 2002 [http://searchenginewatch.com/reports/mmxi-europe.html] 11.03.03.

[8] Vgl. BMBF-Studie zur Nutzung elektronischer Informationen an deutschen Hochschulen. Rüdiger Klatt, Konstantin Gavriilidis, Kirsten Kleinsimlinghaus, Maresa Feldmann u.a.: Nutzung elektronischer wissenschaftlicher Information in der Hochschulausbildung. Barrieren und Potenziale der innovativen Mediennutzung im Lernalltag der Hochschulen. Kurzfassung. Dortmund 2001. [ftp://ftp.bmbf.de/010612_zusfass.pdf] 10.03.03.

[9] [http://www.wordspy.com/words/google.asp] 10.03.03. Die „American Dialect Society" nominierte das „Verb“ "google" gar als nützlichstes Wort des Jahres 2002 [http://www.americandialect.org/woty.html] 12.03.03.

[10] [http://www.heise.de/tp/deutsch/special/auf/12187/1.html Artikel vom 28.03.02] 07.03.03.

Untersuchungen zur Leistungsfähigkeit von Suchdiensten im Web existieren in verschiedenen Ausprägungen. Usabilitytests untersuchen vor allem Einflussfaktoren der Benutzungsoberfläche auf den Rechercheerfolg und Benutzerzufriedenheit [Aurelio & Mourant 2001]]. Retrievaltests prüfen die Leistungsfähigkeit von Suchdiensten anhand der Qualität der Ergebnisse [Griesbaum et al. 2002]. Auch wenn solche Tests zu fundierten qualitativen Urteilen von Suchdiensten führen und, wie zuletzt der Test der Stiftung Warentest,[11] die vergleichsweise hohe Qualität des Marktführers Google bestätigen, lassen sich doch, durch den notwendigerweise eingeschränkten Fokus der jeweiligen Untersuchungen, nur sehr eingeschränkte Aussagen bezüglich der grundlegenden Eignung der untersuchten Suchdienste als Instrumente der informationellen Absicherung ableiten.

Um in einer ersten Näherung die Frage der prinzipiellen Befähigung der dominierenden Suchdienste als primäre Zugangsportale zu den Wissensräumen des Internets zu erörtern, scheint deshalb zunächst ein breiterer und deshalb weniger spezifischer Ansatz sinnvoll. Hier wird der Versuch unternommen, die grundlegende Leistungsfähigkeit der dominierenden Suchdienste anhand einer Analyse der prinzipiellen Verfahren bzw. Kriterien der Indexierung und Ergebnissortierung zu skizzieren.

Eine rein funktionale Analyse ist aufgrund der zunehmenden, mittlerweile stark ausgeprägten Kommerzialisierung unzureichend.[12] Um die Grenzen und vor allem Manipulationspotenziale offen zu legen, sind auch die verschiedenen, u.U. sogar gegenläufigen, Interessen der Partizipanten des Suchdienstemarktes darzustellen [Machill et al. 2002] S.12.

Den Ausgangspunkt der nachfolgenden Analyse bildet eine Kategorisierung der grundlegenden Suchdienstetypen, auf denen die populären Suchdienste, mittlerweile meist kombiniert, aufsetzen. Die Darstellung der verschiedenen Verfahren und Kriterien zur Aufnahme von Informationen in die Indizes der idealtypischen Suchdienstetypen macht deutlich, auf welche Inhalte aus den Wissensbeständen des Netzes theoretisch zugegriffen werden kann. Die Beschreibung fundamentaler Sortierkriterien gibt Aufschluss darüber, welche Eigenschaften für die Sichtbarkeit in vorderen Positionen der Suchergebnisse maßgeblich sind.

3. Partizipanten des Suchdienstemarktes

Das ursprüngliche Bild des World Wide Web als Netz des wissenschaftlichen Austauschs freier Information lässt sich seit geraumer Zeit nicht aufrecht erhalten. Die zunehmende private und geschäftliche Nutzung des Netzes haben zu einer anhaltenden Kommerzialisierung insbesondere auch des Suchdienstemarktes geführt. Die Auffindbarkeit ist ein zentraler Erfolgsfaktor vieler kommerziell ausgerichteter Informationsanbieter. Suchmaschinenmarketing (SEM), häufig von speziellen Search Engine Optimization (SEO) Dienstleistern umgesetzt, gilt als die effektivste Marketingmethode im Web. Schließlich bildet die Monetarisierung der Suchergebnisse zunehmend die finanziellen Grundlage der Suchdienste.[13] Die Nutzer sind primär an relevanten Suchergebnissen interessiert. Neben „klassischen Informationsbedürfnissen“ im Sinne von „Suche Informationen zum Thema x“ werden häufig auch sogenannte „navigational“ oder „homepage target queries“ formuliert, die das Ziel haben, bekannte oder vermutete Angebote im Web (wieder)

[11] Stiftung Warentest, Internet Suchmaschinen, 02/2003, S.38-41.
[12] Danny Sullivan schreibt schon Ende 2000: “In the search engine business, this is known as ‘monetizing the search,’ which means making money in some way off the search results you present. More than ever, monetizing the search is a concern for the search engines.”Danny Sullivan, Monetizing The Search, From The Search Engine Report, 09. 2000. [http://searchenginewatch.com/sereport/00/09-money.html] 17.03.03.
[13] Danny Sullivan, The End For Search Engines?, From The Search Engine Report, 02. 2001. [http://searchenginewatch.com/sereport/01/02-theend.html] 23.03.03.

zu finden. Darüber hinaus werden sehr häufig auf Transaktionen zielende Anfragen eingegeben, die auf die Suche nach Dienstleistungen oder Produkten oder Informationen darüber schließen lassen.[14]

Damit besteht für Informationsanbieter im WWW oft ein direktes kommerzielles Interesse den Suchdiensten vorhandene Informationen nicht nur bereit zu stellen, sondern deren Auffindbarkeit aktiv zu fördern, um nach der Formel bessere Sichtbarkeit bedeutet mehr Besucher und Kunden oder zumindest Werbeeinnahmen einen höheren Umsatz zu erzielen.

Um dieses Bedürfnis zu befriedigen verwendet die SEO Industrie verschiedene Platzierungstechniken. Suchmaschinenoptimierung als „klassische Methode“ umfasst alle Techniken, die darauf zielen innerhalb der neutralen Bewertungskriterien der Suchdienste möglichst gute Platzierungen zu erhalten. Die Grenze der Interpretation von Suchmaschinenoptimierung ist fließend. Die zugrundeliegende Fragestellung lautet, ist Suchmaschinenoptimierung im Einzelfall eher als Unterstützung oder eher als Manipulation der automatischen oder redaktionellen Bewertungsmechanismen von Suchdiensten zu interpretieren? Entscheidend ist, Manipulationspotenzial ist vorhanden. Die Anwendung solcher Techniken zu Spam-Zwecken stellt für die Suchdienste ein zentrales Problem dar. [15]
Spamming beeinträchtigt die Geschäftsgrundlage der Suchdienste, denn hinreichend relevante Suchergebnisse bilden die Basis für hohe Nutzerzahlen. Die Marktreichweite der Suchdienste bestimmt wiederum deren Attraktivität als Promotionsinstrumente für die Informationsanbieter. Für die Suchdienste ist es lukrativ, mit Hilfe von Platzierungs- oder Inklusionsprogrammen die Treffer der Ergebnisseiten, oder Teile davon, zu vermarkten. Die Suchdienste stehen damit im Spannungsfeld zwischen den Erwartungen der Nutzer und den finanziellen Möglichkeiten, die sich durch die Generierung von Besucherströmen für kommerzielle Informationsanbieter eröffnen.

4. Suchdienstetypen, Suchdiensteanbieter

Suchdienste im Web werden überwiegend nach Katalogen und Suchmaschinen differenziert [Ferber 2003] S.295-306. Als Kataloge werden Verzeichnisse bezeichnet, in denen von Menschen ausgewählte Webseiten als geordnete Sammlung von Verweisen mit mono- oder polyhierarchischer Ordnungsstruktur zusammengestellt werden. Unter Suchmaschinen werden Systeme verstanden, die mit Hilfe automatischer Programme und Algorithmen die Inhalte des Web indexieren und sortieren. Aufbauend auf diesen zwei Grundtypen werden häufig als dritte Ausprägung von Suchdiensten Metasuchmaschinen angeführt. Metasuchmaschinen sind Dienste, die mehrere Suchdienste über eine Eingabemaske abfragen und die gelieferten Ergebnisse in einer Ergebnisseite zusammen stellen [Ferber 2003] S.307. Mit GOTO[16] etablierte sich 1998 erstmals erfolgreich eine neue Art von Suchdiensten.[17] Sogenannte Pay-per-Click-Suchdienste. Dies sind Suchdienste, deren Trefferlisten nach Gebotshöhe sortiert werden. Informationsanbieter ersteigern durch Abgabe von

[14] Ungefähr 35-50% aller Anfragen laut [http://www.clickz.com/feedback/buzz/print.php/2168281 21.03.03] 26.03.03.
[15] Vgl. die Diskussion um Cloaking. Unter Cloaking versteht man das Ausliefern unterschiedlicher Inhalte unter derselben URL in Abhängigkeit von vorgegebenen Kriterien. Cloaking verfolgt meist den Zweck, in Suchmaschinen hohe Positionen in den Trefferlisten zu erreichen und wird angewendet, um zwischen Suchmaschinen und menschlichen Besuchern zu differenzieren. Dazu werden den Suchmaschinen beim Spidern Seiten ausgeliefert, die auf die jeweiligen Rankingmechanismen optimiert sind, während menschlichen Besuchern unter derselben URL andere, i.d.R. optisch attraktiv aufbereitete, Seiten präsentiert werden. Befürworter vertreten die Ansicht, Cloaking sei kritisch zu betrachten, der Gebrauch dieser Positionierungstechnik für gewisse Zwecke aber notwendig und damit für Suchdienstenutzer wie Suchmaschinen selbst hilfreich. Gegner hingegen betrachten Cloaking stets als Spam. Befürworter sind eher Vertreter der SEO-Industrie, Suchmaschinenbetreiber, vor allem Google, lehnen Cloaking i.d.R. ab [http://www.enforum.net/www/inf/iwk/enforum.nsf/7e40294922effa30c1256ae1004456c2/0ce285604a80971ec1256b690058f509?OpenDocument] 28.03.03.
[16] Mittlerweile umbenannt in Overture [http://www.overture.com].
[17] Danny Sullivan, GoTo Sells Positions, From The Search Engine Report, 03. 1998 [http://searchenginewatch.com/sereport/98/03-goto.html] 21.03.03.

Geboten für Suchbegriffe Positionen in der Trefferliste. Dabei gilt: je höher das Gebot für eine Suchanfrage, desto höher die Position in der Ergebnisliste.

Faktisch ist die idealtypische Trennung in drei oder vier Suchdienstetypen nicht aufrecht zu erhalten. Kataloge greifen meist auf Suchmaschinenergebnisse zurück, wenn sie selbst zu einer Suchanfrage keine Treffer liefern können. Umgekehrt integrieren Suchmaschinen häufig Kataloge und stellen als zusätzliche Option eine Suche im Verzeichnis bereit. Einige Suchdienste integrieren Katalogeinträge und roboterbasierten Suchergebnisse in einer Trefferliste. Suchdienste verwenden also die Ergebnisse anderer Suchdienste, die wiederum auf andere Suchdienste zurück greifen und diese integrieren usw.. Die Paid Listings des Anbieters Overture werden beispielsweise bei Yahoo, Msn, Lycos, Altavista eingeblendet.[18] Overture selbst verwendet wiederum die Suchmaschinentreffer von Inktomi als sekundäre Ergebnisse. Die Suchtreffer von Google bilden die primären Suchergebnisse von Aol.com, Netscape.com, Yahoo.com. Die Katalogeinträge des Open Directory Projects werden unter anderem von Google, Lycos, Hotbot verwendet. Der Suchdienstemarkt ist also eng verflochten.[19] Dabei bildet die Suchtechnologie einiger weniger Suchmaschinen - Google, Altavista, Inktomi und Fast - das technologische Rückgrat quasi aller bedeutenden Suchdienste. Der Markt für direkt bezahlte Einträge wird von Google und Overture, in Europa auch von Espotting, dominiert. Mit der Akquise von Inktomi durch Yahoo und dem Kauf von Altavista und Alltheweb durch Overture reduziert sich die Zahl bedeutender Suchtechnologieanbieter auf drei Akteure von globaler Bedeutung. Google, Yahoo und Overture.

5. Erschließung der Wissensbestände des Internet

Welche Inhalte des Internets werden durch die verschiedenen Suchdienste nun auf welche Weise erschlossen? Auf welche Datenbestände des Netz können die Nutzer mit Hilfe der dominierenden Suchdienste theoretisch zugreifen?

1. Bei Katalogen entscheidet eine redaktionelle Begutachtung über die Aufnahme und Zuordnung in die Hierarchie. Dabei werden i.d.R. Titel, URL und ein Beschreibungstext erfasst. Kataloge bieten Browsing-Zugriff über die hierarchischen Rubriken oder Keyword-matching in den Metadaten. Die redaktionelle Begutachtung sichert eine hohe Qualität und semantisch korrekte Einordnung der Einträge. Bedingt durch die aufwändige Erstellung erschließen sie nur einen kleinen Teil des Web und bieten keinen Zugriff auf die tatsächlichen Informationen. Detaillierte Inhalte einzelner Webseiten eines Angebots werden nicht erfasst, vielmehr gibt der Beschreibungstext den Inhalt einer gesamten Site wieder [Ferber 2003] S.295. Mittlerweile ist für die redaktionelle Begutachtung und Aufnahme i.d.R. eine einmalige, jährliche oder monatliche Zahlung erforderlich.[20] Für die zuvor kostenlos aufgenommenen Einträge besteht meist Bestandschutz. Die Aufnahme nicht zahlungsfähiger oder zahlungsbereiter Informationsanbieter ist inzwischen i.A. unwahrscheinlich. Damit besteht die Tendenz, dass nichtkommerzielle Informationsangebote zunehmend ausgeblendet werden und die Verzeichnisse sich immer mehr zu Branchenbüchern des Netzes entwickeln.[21]

2. Die Dokumentbeschaffung bei Suchmaschinen findet primär mit Hilfe von Roboterprogrammen

[18] Stand vom19.03.03 [http://overture.com/d/USm/about/news/uspart.jhtml].
[19] Danny Sullivan, Who Powers Whom? Search Providers Chart, SearchEngineWatch.com 02.2003 [http://searchenginewatch.com/reports/alliances.html] 19.03.03.
[20] Letzteres ist seit April 2002 bei Looksmart Bedingung. Vgl. [20] Danny Sullivan, Search Engine Watch LookSmart Changes To Cost-Per-Click Listings, From The Search Engine Report, 05. 2002. [http://searchenginewatch.com/sereport/02/05-looksmart.html] 19.03.03
[21] Der Katalog Web.de war beispielweise nicht bereit, einen kostenlosen Eintrag für ein Forschungsprojekt der Informationswissenschaft der Universität Konstanz, ENFORUM – ein virtuelles kollaboratives Wörterbuch – vorzunehmen.

statt. Sogenannte Spider traversieren, ausgehend von einer vorhandenen URL-Liste, rekursiv die Hyperlinks des Web und parsen die (Text) Inhalte unterschiedlicher Dokumentformate. Die Informationsanbieter können i.A. das Verhalten der Spider mit Hilfe des Robots Exclusion Standard und teilweise auch über verschiedene Metaangaben in Webseiten steuern. Suchmaschinen erheben prinzipiell den Anspruch die Inhalte des gesamten Webs zu erfassen.[22] Dies gelingt (nur) teilweise [Ferber 2003] S.299-302. Problembereiche stellen insbesondere nicht verlinkte, neu erstellte oder aktualisierte Dokumente und vor allem dynamische Dokumente dar. Die Aktualisierungszyklen umfassen Zeiträume zwischen zwei Wochen und mehreren Monaten.[23] Die Indizes der Suchmaschinen sind somit unvollständig und dabei permanent veraltet. Das Aktualitätsproblem versuchen die Suchmaschinenbetreiber durch die Anpassung der Indexierungsfrequenz, nach der Abfragehäufigkeit bzw. der Popularität oder der Aktualisierungsfrequenz der Informationsangebote zu lösen.[24] Schätzungen gehen davon aus, dass insgesamt rund 30-40% des „Surface-Web" von Suchmaschinen erfasst werden [Ferber 2003] S.301. [Bergman 2000] formuliert das grundliegende Problem, dass Suchmaschinen Inhalte des „Deep-Web"[25] nicht erfassen. Unter dem Deep-Web lassen sich alle Inhalte verstehen, auf die aufgrund von Zugangsbeschränkungen durch die Anbieter oder technischen Restriktionen seitens der Suchmaschinen nicht zugegriffen werden kann. Meist handelt es sich um anbieterspezifische Datenbanken, die Webseiten erst aufgrund konkreter Nutzeraktionen dynamisch generieren. Beispiele für Deep-Web-Inhalte stellen etwa die Ergebnisseiten der Suchmaschinen oder zugangsgeschützte Datenbanken dar. Schätzungen gehen davon aus, dass rund 200 000 Deep-Web-Sites vorhanden sind. Die Datenmenge und die Anzahl der Dokumente des Deep-Web soll die des Surface-Web um ein Vielfaches übertreffen. [BERGMAN 2000]. Da insbesondere sehr umfangreiche Informationsangebote datenbankbasiert sind, werden gerade die Wissensbasen größerer Informationsanbieter über die dominierenden Suchdienste unzureichend referenziert. Unterstellt man aufgrund des erforderlichen technischen Know-hows und der zu tragenden Kosten eines solchen Informationsangebot eine höhere Professionalität solcher Informationsanbieter, so lässt sich die Schlussfolgerung ziehen, dass die Suchmaschinen tendenziell eher den Zugriff auf eher unspezifische und weniger professionelle Informationen ermöglichen, während die Inhalte eher spezifischerer und professionellerer Informationsangebote in den Indizes der Suchmaschinen eher nicht zu finden sind. Ansätze zur Dokumentbeschaffung und Inhaltserschließung des Deep-Web stehen noch ganz am Anfang. Erste Konzepte beruhen auf Überlegungen die Inhalte durch dynamische Anfragen an Datenbanken der Deep-Web-Sites zu erschließen [Weber 2001]. Neben der automatischen Erfassung von Webinhalten mit Hilfe von Roboterprogrammen ermöglichen die genannten Suchmaschinen, mit Ausnahme von Inktomi, eine kostenfreie Anmeldung. Mittels einer Eingabemaske können Informationsanbieter den Suchmaschinen kostenfrei Webseiten zur Indexierung vorschlagen. Eine Aufnahme in den Index der Suchmaschinen wird nicht garantiert. Aufgrund der zahlreichen Missbrauchsversuche, z.B. durch tägliche Massenanmeldung, werden solche Einträge häufig nicht in die Indizes aufgenommen bzw. schlechter bewertet, als Inhalte die durch die Roboterprogramme gefunden werden.

3. Eine dritte Art der Inhaltserschließung stellen die sogenannten Paid-Inclusion-Programme dar. Diese bieten gegen Bezahlung die Garantie der Aufnahme und zweitägige oder wöchentliche Aktualisierung im Index der jeweiligen Suchmaschine.[26] Die durch Paid-Inclusion übermittelten

[22] Plakativ sei hier nur der ehemalige Slogan der Suchmaschine Alltheweb angeführt „all the web, all the time" der mittlerweile durch die Formulierung „find it all" abgelöst wurde. [http://www.alltheweb.com/] 18.03.03.
[23] Vgl. Search Engine Statistics: Freshness Showdown vom 20.10.2002 [http://www.searchengineshowdown.com/stats/freshness.shtml] 18.03.03.
[24] Vgl. die verschiedenen Ansätze bei Altavista [http://www.at-web.de/altavista/relaunch2002.htm] 18.03.03 und Google [http://www.webworkshop.net/google_fresh_crawl.html] 18.03.03.
[25] Zum Teil auch als „Hidden" oder „Invisible Net" bezeichnet.
[26] Danny Sullivan, Search Engine Watch, The Evolution Of Paid Inclusion, From The Search Engine Report, 07. 2001. [http://searchenginewatch.com/sereport/01/07-inclusion.html] 19.03.03

Seiten unterliegen den neutralen Rankingverfahren der Suchdienste.[27] Die Vorteile für Informationsanbieter bestehen z.B. darin, dass die Möglichkeit besteht, Inhalte bereit zu stellen, die normalerweise von Suchmaschinen nicht gefunden oder nur schwer indiziert werden können, insbesondere auch Deep-Web-Inhalte.[28] Die Informationsanbieter behalten eine weitgehende Kontrolle über die Anzeige ihrer Seiten auf den Ergebnisseiten der Suchmaschinen. Durch die hohe Indexierungsfrequenz wird Aktualität sichergestellt. Die Inhaltserschließung mit Hilfe von Paid-Inclusion birgt also Potenziale, die geeignet scheinen, die obengenannten Problembereiche der Suchmaschinen bei der Inhaltserschließung zu mindern. Die grundlegende Einschränkung besteht wiederum in der Tatsache, dass dies nur für kommerziell potente und zahlungsbereite Informationsanbieter gilt. Eine Kommerzialisierung der als neutral geltenden Indizes ist die Folge. Es stellt sich die Frage, ob und inwieweit die „editoriale" Integrität der Suchmaschinenergebnisse dabei erhalten bleibt. Das Argument diese bliebe unangetastet, weil die auf diese Weise erschlossenen Inhalte keinerlei "Rankingvorteile" genießen, ist hinterfragbar. Zunächst lässt sich bestreiten, dass dies faktisch zutrifft. Beispielweise wird die Anwendung klassischer Suchmaschinenoptimierungstechniken durch die kürzeren Feedbackzeiten erheblich erleichtert. Grundsätzlich lässt sich anführen, dass die Gefahr besteht, dass sich eine Tendenz dahingehend herausgebildet, andere, nicht durch Paid-Inclusion übermittelte Seiten, zu vernachlässigen. Dies ist z.B. der Grund dafür, dass Google kein Paid-Inclusion-Programm anbietet. In einem Interview hat sich Sergey Brin folgendermaßen geäußert: "'We don't want to incentivize ourselves to do a worse job on the nonpaying sites' - and so, he says, Google is not considering indexing fees. 'You're putting yourself in a very difficult situation. Suppose there's some very good Web site on cancer, but this Web site hasn't paid you. Are you going to give the user a worse site and a worse source of information just because the site hasn't paid? I think it's an ethically difficult matter.'"[29]

4. „Indexierungsverfahren" von Pay-per-Click-Suchdiensten sind der editorialen Begutachtung bei Katalogen recht ähnlich. Der zentrale Unterschied liegt darin, dass die Positionierung in den Ergebnislisten explizit über die Gebotshöhe festgelegt wird. Im Rahmen vorgegebener Richtlinien,[30] welche die Qualität der Einträge sicherstellen sollen, können Werbetreibende Suchanfragen buchen und die Indexierungsangaben i.d.R. Titel, Beschreibungstext und URL weitgehend frei bestimmen. Paid-Listings ermöglichen ebenfalls die Erschließung von Inhalten, die Suchmaschinen ansonsten verborgen bleiben. Die Einträge bestehen fast vollständig aus Geboten kommerzieller Informationsanbieter und umfassen häufig sogenannte Deep-Links, die Suchdienstenutzer direkt auf Produkte und Dienstleistungen verweisen.

6. Sortierverfahren, Rankingkriterien

Suchdienstenutzer formulieren überwiegend kurze Anfragen und sichten in der überwiegenden Zahl der Fälle maximal die ersten drei Ergebnisseiten [Jansen et al. 2003]. Das bedeutet die Einträge der Suchdienste, die nicht auf den vorderen Plätzen gefunden werden (können) bleiben i.d.R. quasi ebenso unsichtbar, als ob sie überhaupt nicht erfasst worden wären. Die Rankingkriterien der Suchdienste sind somit zentrale Informationssortierungsalgorithmen, die letztlich vorentscheiden, welche Informationen wichtig und welche unwichtig sind.

---

[27] Danny Sullivan bezeichnet Paid-Inclusion deshalb auch als "Paid crawling", [http://www.clickz.com/search/opt/print.php/870521] 13.03.03.

[28] Siehe beispielsweise die Informationen von Inktomi unter [http://www.marketleap.com/services/semarketing/indexconnect.htm] 28.03.03

[29] [http://www.thestandard.com/article/0,1902,18023,00.html?body_page=2] 19.03.03.

[30] Listing Guidelines von Overture finden sich unter [http://www.overture.com/d/USm/about/advertisers/relevancy.jhtml] 20.03.03.

1. Kataloge verwenden vielfältige Sortierkriterien, die auch von der Zugriffsart abhängig sind. Bei Browsing-Zugriff sind die hierarchische Position im Verzeichnis, die Popularitätseinstufung durch die Redakteure und die alphabetische Sortierung primäre Kriterien. Bei Matching-Zugriff über die Suchmaske sind inhaltsbezogene Kriterien in den erfassten Metadaten entscheidend. Eine Volltextsuche über tatsächliche Seiteninhalte ist nicht möglich. Kataloge eignen sich für eher unspezifische Informationsbedürfnisse. Die Suchtreffer sind insofern von einer hohen Qualität, dass durch die redaktionelle Einordnung von den Ergebnissen zumindest thematische Einschlägigkeit erwartet werden kann.

2. Suchmaschinen nehmen bei der Inhaltserschließung meist eine Volltextinvertierung vor. Zur Dokumentbewertung werden einerseits sichtbare und unsichtbare Textinformationen herangezogen und andererseits auch nicht-dokumentinhärente Metainformationen, vor allem Linkstrukturen berücksichtigt. Dokumentinhärente Kriterien beruhen auf der Annahme, dass Relevanz als positive Relation der Terme von Suchanfragen und informationellen Einheiten operationalisiert werden kann. Um diese zu bestimmen werden Suchanfragen und Dokumentrepräsentationen hinsichtlich lexikalischer Ähnlichkeit abgeglichen. Die Sortierung wird typischerweise von der Position und Häufigkeit der Suchterme bestimmt. Dabei werden Faktoren wie Funktion der Wörter (URL, Titel, Überschrift, Link), Formatelemente (Schriftgröße, Farbe), HTML-Elemente (z.B. Dateinamen von Bildern, Kommentare) in die Berechnung miteinbezogen. Diese Kriterien sind bei der Relevanzbestimmung häufig nicht hinreichend, insbesondere wenn bei sehr kurzen Anfragen viele tausende potenziell relevante Dokumente vorhanden sind, und sind leicht zu manipulieren. Spam-Techniken wie falsche Inhaltsangaben, Termwiederholungen, unsichtbarer Text haben inzwischen bewirkt, dass z.B. dokumentinhärente Metaangaben bei der Sortierung kaum noch berücksichtigt werden. Aus diesem Grund gebrauchen Suchmaschinen seit geraumer Zeit zusätzliche Sortierungskriterien. Dabei spielt die Analyse von Referenzstrukturen, die erstmals in Form von Pagerank bei Google verwendet wurde, [Brin & Page 1998] eine zentrale Rolle.[31] Die zugrundeliegende Idee ist es, die Bedeutsamkeit der Dokumente durch die Auswertung der Verweisstrukturen zu ermitteln und dies bei der Sortierung mit zu berücksichtigen. Gerade bei kurzen Anfragen lässt sich auf diese Weise häufig eine sehr hohe Qualität erreichen. Ein weiterer Vorteil ist, dass Missbrauch relativ schwierig ist, da die hierzu erforderlichen Spam-Techniken einen vergleichsweise hohen Aufwand erfordern. Nachteilig ist, dass neue Dokumente prinzipiell benachteiligt werden und somit eine Tendenz zur Verstetigung der Suchergebnisse besteht [Feldman 2002] S.187-188. Wie dies zu bewerten ist kann hier nicht beantwortet werden. Wichtig ist es festzuhalten, dass die Berücksichtigung von Verweisstrukturen bei der Sortierung durchaus problematisch sein kann. Während beispielweise Google argumentiert, Pagerank berücksichtige das demokratische Stimmverhalten des Internets,[32] übt z.B. Daniel Brandt von Googlewatch harte Kritik: „In other words, the rich get richer, and the poor hardly count at all. This is not ‘uniquely democratic,’ but rather it's uniquely tyrannical. It's corporate America's dream machine, a search engine where big business can crush the little guy.“ Fakt ist, die Berücksichtigung von Referenzstrukturen bei der Ergebnissortierung bewirkt mittlerweile eine Veränderung im Linkverhalten der Informationsanbieter. Beispielsweise nutzte der Suchdienst Searchking ab Sommer 2002 Pagerankeinstufungen von Google zu Werbezwecken.[33] Solche Anpassungen der Informationsbieter drohen die Vorteile dieser Sortierungsverfahren obsolet werden zu lassen. Zudem lässt sich nicht ausschließen, dass sich durch ein solches verändertes Linkverhalten, das sich unter anderem im Anlegen sogenannter Linkfarmen zeigt,[34] mittelfristig die Netzstruktur selbst signifikant verändert. Kennzeichnend für die Ergebnissortierung bei Suchmaschinen ist,

[31] Andere Ansätze, wie die von Direct Hit verwendete Clickpopularity, sind mittlerweile fast unbedeutend.
[32] [http://www.google.com/technology/] 21.03.03.
[33] [http://www.intern.de/news/3614.html 21.10.2002] 21.03.03.
[34] Heise News vom 04.12.2002, Verlinken oder bezahlen – Suchmaschinen im Internet. [http://www.heise.de/newsticker/data/anw-04.12.02-004/] 23.03.03.

dass sie verschiedene Sortierungsverfahren verwenden und miteinander kombinieren. Dabei werden nicht-dokumentinhärente Verfahren in zunehmenden Maße berücksichtigt. Diese gelten als zentraler Qualitätsfaktor. Die grundlegenden Bewertungskriterien werden offen gelegt und sind dadurch für den Nutzer prinzipiell nachvollziehbar. Die konkrete Zusammensetzung und Gewichtung gelten als primärer Erfolgsfaktor und werden nicht transparent gemacht. Die verwendeten statistischen Verfahren erreichen mittlerweile häufig eine hohe Qualität. Aber gerade bei häufig genutzten Suchbegriffen muss mit Spamming-Versuchen gerechnet werden. Deshalb ist es möglich, dass bei populären Suchanfragen wie z.B. Moorhuhn pornographische Ergebnisse zurückgegeben werden können [Machill et al. 2002] S.42-43.

3. Pay-per-Click-Suchdienste sortieren nach Gebotshöhe (Overture)[35] oder nach einer Kombination von Gebotshöhe und Klickhäufigkeit (Google).[36] Eine redaktionelle Kontrolle sichert eine grundlegende Einschlägigkeit bzw. inhaltliche Relevanz. Der Grad der Wichtigkeit wird nicht nach inhaltlichen Kriterien sondern nach Zahlungsbereitschaft und Clickpopularität festgelegt. Die Sortierung nach Gebotshöhe lässt sich qualitativ dann begründen, wenn man davon ausgeht, dass die Informationsanbieter selbst ein ökonomisches Interesse daran haben, nur relevante Suchergebnisse zu positionieren. Die zugrundeliegende Annahme ist, dass die Zahlungsbereitschaft des Informationsanbieter mit der Relevanz seines Angebots für den Informationsnachfrager korrespondiert. Overture-Gründer Bill Gross argumentiert, diese Rankingmethode sei den inhaltsbasierten Sortierverfahren bei Informationsbedürfnissen, die auf der Suche nach Dienstleistungen und Produkten beruhen, qualitativ zumindest ebenbürtig. "All you're really getting back is how good people are at tricking the search engines, not how good people are, Gross said." [37] Welche Qualität Paid Listings im Vergleich zu roboterbasierten Suchergebnissen oder den editorialen Einträgen tatsächlich aufweisen ist bislang nicht bekannt.[38]

7. Ergebnis

Die skizzenhafte Darstellung des Suchdienstemarktes zeigt die unterschiedlichen, u.U. konfligierende Ziele der Partizipanten auf. Zentraler Punkt sind die ökonomischen Interessen, die sich gegenwärtig dergestalt ausprägen, dass die Sichtbarkeit, Rangfolge der Suchergebnisse, zunehmend durch Marketingstrategien und ökonomische Kompetenz der Informationsanbieter sowie der Vermarktungsbereitschaft der Suchdienste beeinflusst werden. Durch die zu Jahresbeginn getätigten oder angekündigten Akquisitionen von Overture und Yahoo ist mit weitreichenden Umstrukturierungen auf dem Suchdienstemarkt zu rechnen. Die Zahl globaler bedeutender Suchtechnologieanbieter reduziert sich gegenwärtig auf drei Akteure, damit kann zumindest auf Anbieterseite gegenwärtig weniger von einem Google Monopol als vielmehr von einem Google-, Yahoo-, Overture-Oligopol gesprochen werden.

Die populären Suchdienste alleine ermöglichen in keinem Fall ausreichenden Zugriff auf die Wissensbestände des Netzes. Suchmaschinen, die über das größte Abdeckungspotenzial verfügen, erfassen zwar große Teilbestände des Surface-Web, die meist datenbankbasierten Inhalte größerer Wissensbasen professioneller Anbieter werden allerdings nur zu einem geringen Teil erfasst. Die Zusammensetzung der Indizes der Suchdienste wird neben technischen und inhaltlichen Kriterien in

[35] [http://www.overture.com/d/USm/learning/bidding_three.jhtml] 23.03.03.
[36] [https://adwords.google.com/select/overview.html] 23.03.03.
[37] Danny Sullivan, GoTo Sells Positions, From The Search Engine Report, 03. 1998 [http://searchenginewatch.com/sereport/98/03-goto.html] 21.03.03.
[38] Diese Fragestellung wird erstmals in der Abschlussarbeit von Caroline Berns, „Ein Vergleich der Retrievaleffektivität von Einträgen bei Paid-Listing-Diensten, redaktionell erstellten Katalogen und roboterbasierten Suchmaschinen“ systematisch evaluiert. Ergebnisse liegen noch nicht vor.

immer höherem Maße von ökonomischen Interessen bestimmt. Auch für die Ergebnisreihenfolge sind neben inhaltlichen Kriterien ökonomische Faktoren immer maßgeblicher. Bei Katalogen bildet die Zahlungsbereitschaft mittlerweile meist die notwendige Voraussetzung einer redaktionellen Bewertung. Bei Suchmaschinen können (finanzielle) Ressourcen dabei behilflich sein innerhalb der statistischen Sortierungsmethoden durch Suchmaschinenoptimierung bessere Positionen zu erreichen. Je nach Einzellfall kann sich dies positiv oder negativ auf die Qualität der Suchergebnisse auswirken. Da es für Informationsanbieter weiterhin attraktiv sein kann Optimierungsverfahren entgegen den Intentionen der Suchmaschinen anzuwenden ist zu erwarten, dass der Wettlauf um die Manipulation bzw. den Schutz der Sortierverfahren weitergeht. Insofern ist auch davon auszugehen, dass Spam weiterhin zumindest ein konzeptionelles Problem für die Suchmaschinen darstellen wird. Bei Paid-Listings sind inhaltliche Kriterien bei der Sortierung quasi nur noch eine notwendige Nebenbedingung. Entscheidender Faktor der Sortierung ist hier alleine die Zahlungsbereitschaft der Informationsanbieter.

Die Auswirkungen der Monetarisierung auf die Qualität der Suchergebnisse einzuschätzen ist schwierig. Es ist durchaus möglich, bzw. nicht unwahrscheinlich, dass sie in vielen Suchkontexten eine Qualitätssteigerung bewirkt. Wichtig scheint es, keine feste Koppelung zwischen finanzieller Potenz und Zahlungsbereitschaft einerseits sowie Indexierung und Relevanzeinstufung andererseits zu schaffen. Ist die kommerzielle Vermarktung von Suchergebnissen mittlerweile auch weitgehend eine notwendige Voraussetzung für das Weiterbestehen der für den Nutzer kostenfreien Dienste, so muss dennoch sichergestellt werden, dass auch nicht kommerziell ausgerichtete Inhalte bzw. nicht zahlungsbereite Informationsanbieter bei der Inhalterschließung und Ergebnissortierung hinreichend berücksichtigt werden.

8. Schlussfolgerungen

Für den Nutzer ist letztlich entscheidend, was für Suchtreffer auf welche Weise auf den Ergebnisseiten der Suchdienste ausgegeben werden. Damit er den Grad der Vollständigkeit, Genauigkeit und Glaubwürdigkeit der Ergebnisse in Bezug auf die vorhandenen Informationsressourcen des Netzes überhaupt einschätzen kann, ist es notwendig, die Einflussfaktoren auf die Suchergebnisse hinreichend kenntlich zu machen. Daraus lässt sich die Forderung ableiten Indexierungs- und Sortierkriterien transparent offen zu legen und die korrespondierenden Einträge auf den Ergebnisseiten entsprechend zu kennzeichnen. Ansonsten besteht z.B. die Gefahr, dass "such concealment may mislead search engine users to believe that search results are based on relevancy alone, not marketing ploys."[39] Eine systematische Evaluation des quantitativen Verhältnisses zwischen Werbung und "neutralen" Ergebnissen sowie der Positionierung und Kennzeichnung (Benennung) kommerzieller Inhalte wäre hilfreich, um in diesem Problemfeld ein erstes Bild hinsichtlich der "kommerziellen" Ausprägung der Trefferlistenseiten zu bekommen.

Ein weitere Möglichkeit die informationelle Absicherungsfähigkeit der Suchdienstenutzer zu erhöhen und die Qualität der dominierenden Suchdienste zu verbessern könnte darin bestehen auf den Ergebnisseiten neben den Suchtreffern zusätzlich eine Auswahl geeigneter Suchdienste, z.B. zur Erschließung von Deep-Web-Inhalten vorzuschlagen. Ansätze hierzu finden sich seit längerer Zeit in Form von „Deep-Web-Gateways“ wie z.B. [http://www.completeplanet.com/], [http://www.invisibleweb.com/]. Bislang wurden solche Ideen von den dominierenden Suchdiensten nicht aufgegriffen.

[39] UNITED STATES OF AMERICA FEDERAL TRADE COMMISSION WASHINGTON, D.C. 20580 June 27, 2002 [http://www.ftc.gov/os/closings/staff/commercialalertattatch.htm] 24.03.03.

8. Literatur

Aurelio, D. N. & Mourant, R. R. (2001). Effect of Web Search Engine Design on User Performance and Preference. In: Proceedings Ninth International Conference on Human-Computer Interaction.

Bergman, M. (2000). The Deep Web: Surfacing Hidden Value. White paper. http://www.brightplanet.com/deepcontent/tutorials/DeepWeb/index.asp.

Brin, S. & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Proceedings of the Seventh International World Wide Web Conference. Ashman, H. & Thistlewaite, P. (eds.); Elsevier, 107-117.

Feldman, S. (2002). This is what I asked for? The searching Quagmire. In: Web of Deception. Misinformation on the internet. Mintz, A. P. et al. (ed.); CyberAge, 175-195.

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

Griesbaum, J., Rittberger, M. und Bekavac, B. (2002). Deutsche Suchmaschinen im Vergleich: AltaVista.de, Fireball.de, Google.de und Lycos.de. Hammwöhner, R., Wolff, C., und Womser-Hacker, C. (eds.); UVK, 201-223.

Jansen, B., Spink, A., und Saracevic, T. (2003). Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. Information Processing & Management, 36 Nr.2, 207-227.

Machill, C., Neuberger, C. und Schindler, F. (2002). Transparenz im Netz. Funktion und Defizite von Internet-Suchmaschinen. Gütersloh: Bertelsmann Stiftung.

Weber, G. (2001). Integration von Datenbanken in Suchmaschinen bei unterschiedlichen Kooperationsgraden. In: Informatik 2001: Wirtschaft und Wissenschaft in der Network Economy - Visionen und Wirklichkeit. Tagungsband der GI/OCG-Jahrestagung, 25.-28. September 2001. Bauknecht, K., Brauer, W., und Mück, T. (eds.); Österreichische Computer-Gesellschaft, 345-352.