**In:** Hammwöhner, Rainer; Rittberger, Marc; Semar, Wolfgang (Hg.): Wissen in Aktion. Der Primat der Pragmatik als Motto der Konstanzer Informationswissenschaft. Festschrift für Rainer Kuhlen. Konstanz: UVK Verlagsgesellschaft mbH, 2004. S. 35 – 50

# Von der Kommerzialisierung bis zum Deep Web: Problemfelder der Internetsuche

*Joachim Griesbaum*[1], *Bernard Bekavac*[2]

[1]Universität Konstanz
Informationswissenschaft
Fach D 87
D-78457 Konstanz
griesbau@inf.uni-konstanz.de

[2]HTW Chur
Hochschule für Technik
und Wirtschaft
CH-7004 Chur
bernard.bekavac@fh-htwchur.ch

**Abstract**
Die Suchdienste des Internet entwickeln sich nicht nur im privaten sondern z.T. auch im professionellen Bereich zu den zentralen Hilfsmitteln zur Befriedigung von Informationsbedürfnissen. Dabei stellt sich die Frage inwieweit Google und andere populäre Suchdienste tatsächlich die Fähigkeit besitzen, Qualität und Umfang an Informationen zu referenzieren, die zu einem gegebenen Informationsbedürfnis im Internet momentan abrufbar sind. Ausgehend von einer Analyse der aktuellen Suchdienstetypen werden die unterschiedlichen Verfahren der Dokumentbeschaffung und Sortierung von Ergebnislisten dargestellt. Damit wird gezeigt, auf welche Inhalte aus den Wissensbeständen des Internets überhaupt zugegriffen werden kann und welche Eigenschaften für die Sichtbarkeit in vorderen Positionen der Suchergebnisse maßgeblich sind. Abschließend wird die reale Ausprägung des Suchdienstemarktes skizziert und die primären Defizite der populären Suchdienste zusammengefasst sowie aktuelle Entwicklungstendenzen angeführt.

## 1 Einleitung

Das Internet entwickelt sich, neben traditionellen Medien wie Zeitungen und Fernsehen sowie klassischen Informationsdiensten wie etwa Bibliotheken, zunehmend zu der zentralen Quelle zur Befriedigung von Informationsbedürfnissen. Die Art und Weise, wie Menschen auf Informationen zugreifen, hat sich, im Vergleich zu den Vorgehensweisen bei

Dieses Dokument wird unter folgender creative commons Lizenz veröffentlicht:
http://creativecommons.org/licenses/by-nc-nd/2.0/de/

traditionellen Medien, grundlegend verändert[1]: Sowohl Gelegenheitssurfer wie auch Informationsexperten nutzen Google und andere bekannte Suchdienste wie etwa Yahoo, Msn, Lycos, Altavista als primäres Instrument, um gezielt im Internet die gewünschten Informationen ausfindig zu machen [Klatt et al. 2001]. Der Grund für die Popularität der genannten Suchdienste ist offensichtlich: Sehr häufig werden Nutzer schon nach Eingabe von ein bis zwei Suchbegriffen mit den Ergebnissen auf der ersten Trefferseite befriedigt. Solange relevant scheinende Treffer zurückgegeben werden, wird Aktualität und inhaltliche Richtigkeit (Validität) der gefundenen Suchergebnisse nur selten in Frage gestellt.[2] Das Vertrauen in die Suchdienste ist hoch, die Frage nach der Objektivität und Validität der gefundenen Informationen scheint immer mehr in den Hintergrund zu geraten.

Betrachtet man die den Suchmaschinen zu Grunde liegenden Suchverfahren und Rankingalgorithmen, so ist die Qualität der ausgewiesenen Trefferlisten zumindest hinterfragbar: „It´s very easy to go into google and get an answer, but it´s fairly easy to get a bad answer or mythological answer. Google represents an illusion of ease of search, it´s easy to use, it´s quick and it´s free but it´s not the whole picture".[3] Dem entgegen gesetzt steht der formulierte Anspruch vieler Suchdienste, dem Nutzer stets die benötigten Informationen bereit zu stellen. Das Leitbild der Suchmaschine Google besagt etwa: "Our mission is to organize the world's information, making it universally accessible and useful."[4]

Letztlich stellt sich also die Frage, ob Google und Co. tatsächlich die Fähigkeit besitzen Qualität und Umfang an Informationen referenzieren zu können, die zu einem gegebenen Informationsbedürfnis im Internet tatsächlich verfügbar sind oder ob nur versucht wird durch eine zumindest teilweise Befriedigung des Informationsproblems auch eine gewisse Befriedigung des Nutzers zu erzielen. Über die den Suchdiensten zu Grunde liegenden genauen Verfahren bei der Erschließung der Wissensbestände und der Sortierung der Ergebnismengen ist nur wenig Detailinformation bekannt.

---

[1] Als führend bzw. dominant kann gegenwärtig insbesondere Google bezeichnet werden URL http://www.wordspy.com/words/google.asp (letzter Zugriff 11.03.2004). Die „American Dialect Society" nominierte das „Verb" "google" gar als nützlichstes Wort des Jahres 2002. URL http://www.americandialect.org/woty.html (letzter Zugriff 11.03.2004).

[2] Seattle Times, Cynthia Flash, Google for a grade: UW class to study popular search engine, 02.02.2004. URL http://seattletimes.nwsource.com/html/businesstechnology/2001848831_google02.html (letzter Zugriff 11.03.2004)

[3] Ebd.

[4] URL http://www.google.com/contact/search.html (letzter Zugriff 11.03.2004).

Dass dabei auch Aspekte der Kommerzialisierung eine immer wichtigere Rolle spielen, ist jedoch unverkennbar. Daher ist also die Frage nach der tatsächlichen Neutralität/Objektivität dieser Verfahren ebenfalls angebracht.

Im Spannungsfeld dieser Fragen wird im Folgenden dargestellt, was Websuchdienste grundsätzlich zu leisten vermögen. Ziel ist es sowohl die Defizite von Websuchdiensten herauszuarbeiten als auch darzulegen, wie die Suchdiensteanbieter momentan versuchen diesen zu begegnen. Ausgangspunkt ist die Analyse der aktuellen Suchverfahren, aus der dann eine Typologie der grundlegenden Suchdienstetypen resultiert. Dabei werden die unterschiedlichen Verfahren der Dokumentbeschaffung und Aufnahmekriterien der Wissensbestände in die Indizes der verschiedenen Suchdienstetypen skizziert, um deutlich zu machen, auf welche Inhalte aus den Wissensbeständen des Netzes überhaupt zugegriffen werden kann. Die Beschreibung aktueller Sortierkriterien gibt Aufschluss darüber, welche Eigenschaften für die Sichtbarkeit in vorderen Positionen der Suchergebnisse maßgeblich sind. Darauf aufbauend wird die reale Ausprägung des Suchdienstemarktes als Markt gegenseitig verflochtener Suchdienste dargestellt und anschließend die primären Defizite der populären Suchdienste zusammengefasst und aktuelle Entwicklungspotenziale angesprochen.

# 2 Typologie der Suchdienste

Suchdienste im Web lassen sich zunächst grundlegend nach Katalogen und Suchmaschinen differenzieren [Ferber 2003] S. 295-306. Kataloge stellen redaktionell ausgewählte Webseiten als geordnete Sammlung von Verweisen mit mono- oder polyhierarchischer Ordnungsstruktur dar. Suchmaschinen sind Systeme, die mit Hilfe automatischer Programme und Algorithmen die Inhalte des Web indexieren und sortieren. 1998 etablierte sich eine zusätzliche Art von Suchdiensten:[5] Pay-per-Click-Suchdienste. Bei diesen Suchdiensten werden die Trefferlisten nicht mehr nach inhaltlichen Kriterien sortiert, sondern rein durch die Vermarktung bestimmt. Dabei ersteigern Informationsanbieter durch Abgabe von Geboten für Suchbegriffe Positionen in der Trefferliste nach dem Motto: Je höher das Gebot, desto höher die Position innerhalb der Trefferliste.

Aufbauend auf diesen Grundtypen werden häufig als weitere Ausprägung von Suchdiensten Metasuchmaschinen angeführt. Metasuchmaschinen fragen

[5] Danny Sullivan, GoTo Sells Positions, From The Search Engine Report, 03. 1998, URL http://searchenginewatch.com/sereport/98/03-goto.html (letzter Zugriff 12.03.2004).

mehrere Suchdienste über eine Eingabemaske ab und stellen die gelieferten Ergebnisse in einer Ergebnisseite zusammen [Ferber 2003] S.307.

Im Folgenden werden die genannten Suchdienstetypen hinsichtlich ihrer Verfahren der Dokumentbeschaffung, Aufnahmekriterien, Inhaltserschließung und Ergebnissortierung grundlegend beschrieben. Dies soll deutlich machen, auf welche Inhalte Suchdienstenutzer tatsächlich zugreifen können und welche Kriterien für die Sichtbarkeit dieser Inhalte auf vorderen Positionen der Suchergebnisse bestimmend sind. Diese zwei Aspekte sind von zentraler Bedeutung, da Suchdienstenutzer meist nur kurze Anfragen formulieren und in den überwiegenden Fällen maximal die ersten drei Ergebnisseiten [Jansen et al. 2000] sichten. Alles, was hinter diesen ersten Trefferseiten kommt, bleibt für die typischen Nutzer quasi unsichtbar.

## 2.1 Kataloge

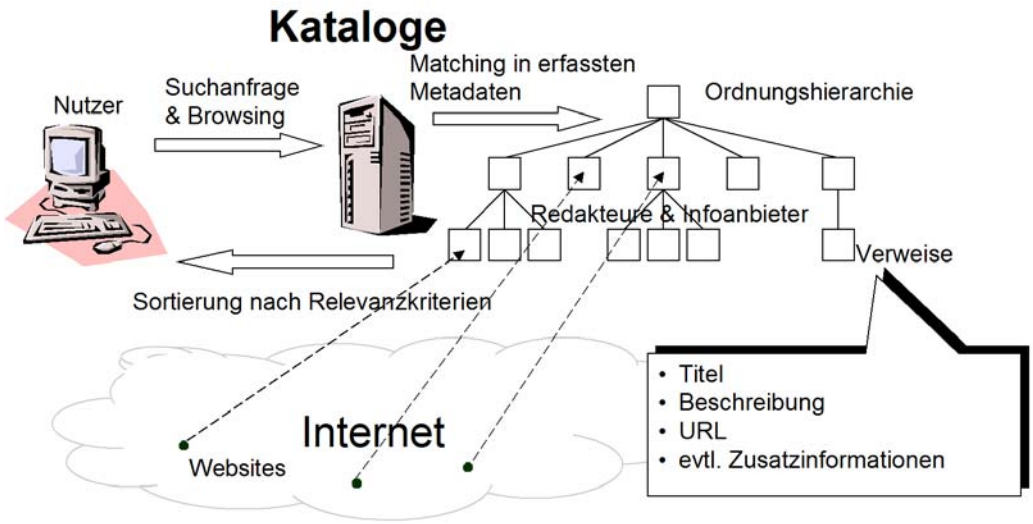


Abbildung 1: Schematische Darstellung Kataloge

Bei Katalogen ist eine redaktionelle Begutachtung entscheidend für die Aufnahme und Zuordnung in die Hierarchie. Dabei werden Metadaten wie Titel, URL und ein Beschreibungstext erfasst. Konzeptionell soll die redaktionelle Begutachtung eine hohe Qualität und semantisch korrekte Einordnung der Einträge sicherstellen. Aufgrund der aufwändigen Erstellung decken die Kataloge nur einen sehr kleinen Teil des Web[6] ab. Zudem basiert die Stichwortsuche innerhalb der Kataloge nur auf dem Beschreibungstext zu den referenzierten Web-Sites und bietet keinen Zugriff auf die Volltexte. Die

[6] Das Open Directory Project, einer der umfangreichsten Kataloge im Web, umfasst nach eigenen Angaben ca. 4 Millionen Websites. URL http://www.dmoz.org/ (letzter Zugriff 12.03.2004).

detaillierten Inhalte einzelner Webseiten eines Angebots werden also nicht erfasst, vielmehr gibt der Beschreibungstext den Inhalt einer gesamten Site wieder [Ferber 2003] S.295. Zudem ist inzwischen für die redaktionelle Begutachtung und Aufnahme – zumindest für kommerzielle Informationsangebote – bei den meisten Katalogen eine Zahlung erforderlich.

Sortierkriterien bei Katalogen sind von der Art des Zugriffs abhängig. Bei Browsing-Zugriff entscheiden hierarchische Position, Popularitätseinstufung und alphabetische Sortierung über die Referenzeinordnung. Bei Matching-Zugriff über die Suchmaske sind inhaltsbezogene Kriterien in den erfassten Metadaten entscheidend. Aus diesem Grund sind Kataloge für eher unspezifische Informationsbedürfnisse geeignet. Die Suchtreffer sind insofern von einer hohen Qualität, weil die redaktionelle Bewertung zumindest Einschlägigkeit erwarten lässt.

## 2.2 Suchmaschinen

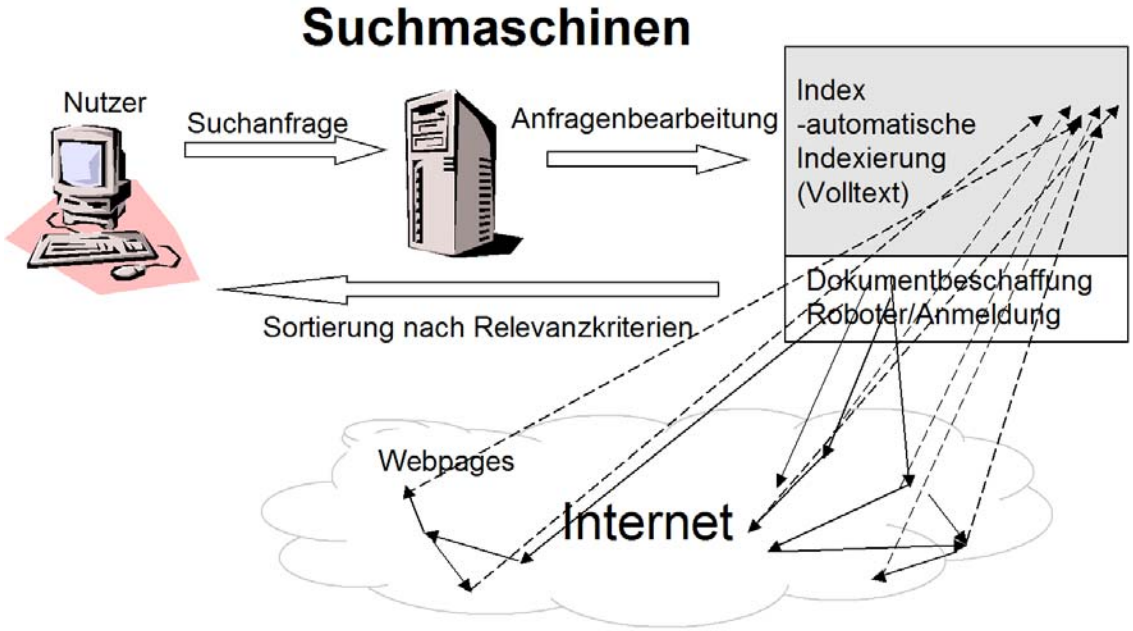


Abbildung 2: Schematische Darstellung Suchmaschinen

Die Dokumentbeschaffung findet bei Suchmaschinen primär mit Hilfe von Roboterprogrammen statt. Diese traversieren rekursiv die Links im Web und erfassen so automatisch die Textinhalte unterschiedlicher Dokumentformate. Suchmaschinen suggerieren dadurch die Inhalte des gesamten Web erfassen

zu können.[7] Dies gelingt jedoch nur ansatzweise [Ferber 2003] S.299-302. Problembereiche stellen insbesondere nicht verlinkte, neu erstellte oder aktualisierte Dokumente und vor allem dynamische Dokumente dar. Die Aktualisierungszyklen bereits erfasster Web-Seiten umfassen Zeiträume zwischen zwei Wochen und mehreren Monaten.[8] Die Indizes der Suchmaschinen sind somit unvollständig und verweisen nicht selten auf Web-Seiten, die nicht mehr erreichbar sind (Dead-Links). Es wird versucht, dieses Aktualitätsproblem durch die Anpassung der Indexierungsfrequenz nach Abfragehäufigkeit bzw. Popularität oder Aktualisierungsfrequenz der Informationsangebote zu lösen.[9]

Schätzungen gehen davon aus, dass insgesamt rund 30-40% des „Surface-Web" von Suchmaschinen erfasst werden [Ferber 2003] S.301. [Bergman 2000] formuliert das grundlegende Problem, dass Suchmaschinen Inhalte des „Deep Web"[10] nicht erfassen. Unter dem Deep Web werden alle Wissensbestände verstanden, auf welche die Roboterprogramme der Suchmaschinen auf Grund von Zugangsbeschränkungen durch die Anbieter oder technischen Restriktionen nicht zugreifen können. In der Regel handelt es sich dabei um anbieterspezifische Datenbanken, die Webseiten erst auf Grund konkreter Nutzeraktionen dynamisch generieren. Beispiele hierfür stellen etwa Web-Shops dar, die erst nach spezifischen Benutzereingaben wie Marken- oder Produktnamen entsprechende Angebotsseiten generieren.

Neben der automatischen Erfassung von Webinhalten mit Hilfe von Roboterprogrammen ermöglichen die Suchmaschinen, momentan noch nur mit Ausnahme von Inktomi, eine kostenfreie Anmeldung. Mittels einer Eingabemaske können Informationsanbieter den Suchmaschinen kostenfrei Webseiten zur Indexierung vorschlagen. Die tatsächliche Aufnahme in den Index der Suchmaschinen wird dabei allerdings nicht garantiert. Aufgrund zahlreicher Missbrauchsversuche, z.B. bei täglicher Massenanmeldung, werden direkte Einträge häufig nicht in die Indizes aufgenommen bzw. schlechter bewertet als Inhalte, die durch die Roboterprogramme gefunden werden.

---

[7] 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. URL http://www.alltheweb.com/ (letzter Zugriff 12.03.2004).

[8] Vgl. Search Engine Statistics: Freshness Showdown vom 17.05.2003 URL http://www.searchengineshowdown.com/stats/freshness.shtml (letzter Zugriff 12.03.2004).

[9] Vgl. die verschiedenen Ansätze bei Altavista, URL http://www.at-web.de/altavista/relaunch2002.htm (letzter Zugriff 12.03.2004) und Google URL http://www.webworkshop.net/google_fresh_crawl.html (letzter Zugriff 12.03.2004).

[10] Zum Teil auch als „Hidden" oder „Invisible Net" bezeichnet.

Eine dritte Art der Inhaltserschließung stellen die sogenannten Paid-Inclusion-Programme dar. Diese bieten gegen Bezahlung eine garantierte Aufnahme und die regelmäßige Aktualisierung im Index der jeweiligen Suchmaschine.[11] Aber auch die über Paid Inclusion erfassten Seiten unterliegen den neutralen Rankingverfahren der Suchdienste und werden bei der Sortierung nicht bevorzugt behandelt.[12] 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.[13] Die Informationsanbieter behalten mit Paid Inclusion auch die weit gehende Kontrolle über die Anzeige ihrer Seiten auf den Trefferlisten der Suchmaschinen. Die hohe Indexierungsfrequenz stellt die Aktualität sicher. Paid-Inclusion-Programme haben folglich einige Vorteile, sind jedoch nur für zahlungskräftige und zahlungswillige Informationsanbieter interessant. Tendenziell hat dies jedoch eine Kommerzialisierung der als neutral geltenden Suchmaschinen zur Folge. Dabei wirft sich natürlich die Frage auf, ob und inwieweit die „editoriale" Integrität bzw. kommerzielle Neutralität der Suchmaschinenergebnisse in Zukunft noch erhalten bleibt.

Zur Dokumentbewertung verwenden Suchmaschinen einerseits sichtbare bzw. auf dem Browser unsichtbare Textinformationen, andererseits werden aber auch nicht-dokumentinhärente Metainformationen, vor allem Linkstrukturen, berücksichtigt. Dokumentinhärente Kriterien beruhen auf der Annahme, dass Relevanz als positive Relation der Terme von Suchanfragen und den Web-Seiten operationalisiert werden kann. Die Sortierung wird typischerweise von statistischen Merkmalen, wie z.B. 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 mit einbezogen. Diese Kriterien sind bei der Relevanzbestimmung meist nicht hinreichend, insbesondere wenn bei sehr kurzen Anfragen viele tausende potenziell relevante Dokumente vorhanden sind. Zudem sind sie leicht zu manipulieren: Spam-Techniken wie falsche Inhaltsangaben,

---

[11] Danny Sullivan, Search Engine Watch, The Evolution Of Paid Inclusion, From The Search Engine Report, 07. 2001. URL http://searchenginewatch.com/sereport/01/07-inclusion.html

[12] Danny Sullivan bezeichnet Paid Inclusion deshalb auch als "Paid crawling", URL http://www.clickz.com/search/opt/print.php/870521 (letzter Zugriff 13.03.2004).

[13] Siehe beispielsweise die Informationen von Inktomi unter URL http://www.marketleap.com/services/semarketing/indexconnect.htm (letzter Zugriff 14.03.2004).

Termwiederholungen, unsichtbarer Text haben inzwischen bewirkt, dass z.B. dokumentinhärente Metaangaben bei der Sortierung nur zur Bestimmung der Grundtreffermenge verwendet werden. Seit geraumer Zeit gebrauchen Suchmaschinen zusätzlich andere Kriterien bei der Sortierung dieser Grundmenge. Zentrales Element ist dabei die Analyse von Referenzstrukturen, die erstmals in Form von Pagerank bei Google verwendet wurde [Brin & Page 1998]. Die zu Grunde liegende Idee ist es, die Bedeutsamkeit der Dokumente durch die Auswertung der Verweisstrukturen zu ermitteln und dies bei der Sortierung mit zu berücksichtigen. Je häufiger auf ein Dokument von anderen Web-Seiten aus verwiesen wird, desto höher wird dieses gerankt. Gerade bei kurzen Anfragen lässt sich auf diese Weise häufig eine sehr hohe Qualität erreichen. Ein weiterer Vorteil ist, dass Missbrauch relativ schwierig ist, weil die 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]. Die Berücksichtigung von Referenzstrukturen bei der Ergebnissortierung bewirkt mittlerweile eine Veränderung im Linkverhalten der Informationsanbieter.[14] Solche Anpassungen drohen die Vorteile dieser Sortierungsverfahren obsolet werden zu lassen. Kennzeichnend für die Ergebnissortierung bei Suchmaschinen ist also, dass verschiedene Sortierungsverfahren verwendet und miteinander kombiniert werden. Dabei werden nicht-dokumentinhärente Verfahren in zunehmendem 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 Verfahren erreichen oft eine hohe Qualität. Aber insbesondere bei häufig verwendeten Suchanfragen muss, nach wie vor, mit Spam gerechnet werden.

## 2.3 Pay-per-Click-Suchdienste

Dokumentbeschaffung und Indexierung sind bei Pay-per-Click-Suchdiensten ähnlich wie bei den Katalogen. Im Rahmen vorgegebener Richtlinien, welche die Qualität der Einträge gewährleisten sollen, können Informationsanbieter Suchanfragen buchen und die Indexierungsangaben, i.d.R. Titel, Beschreibungstext und URL, weit gehend frei bestimmen. Indizes von Pay-per-Click-Diensten bestehen nahezu vollständig aus Geboten kommerzieller

[14] URL http://www.intern.de/news/3614.html (letzter Zugriff 14.03.2004).

Informationsanbieter und umfassen häufig sogenannte Deep Links, die Suchdienstenutzer direkt auf Produkte und Dienstleistungen verweisen.

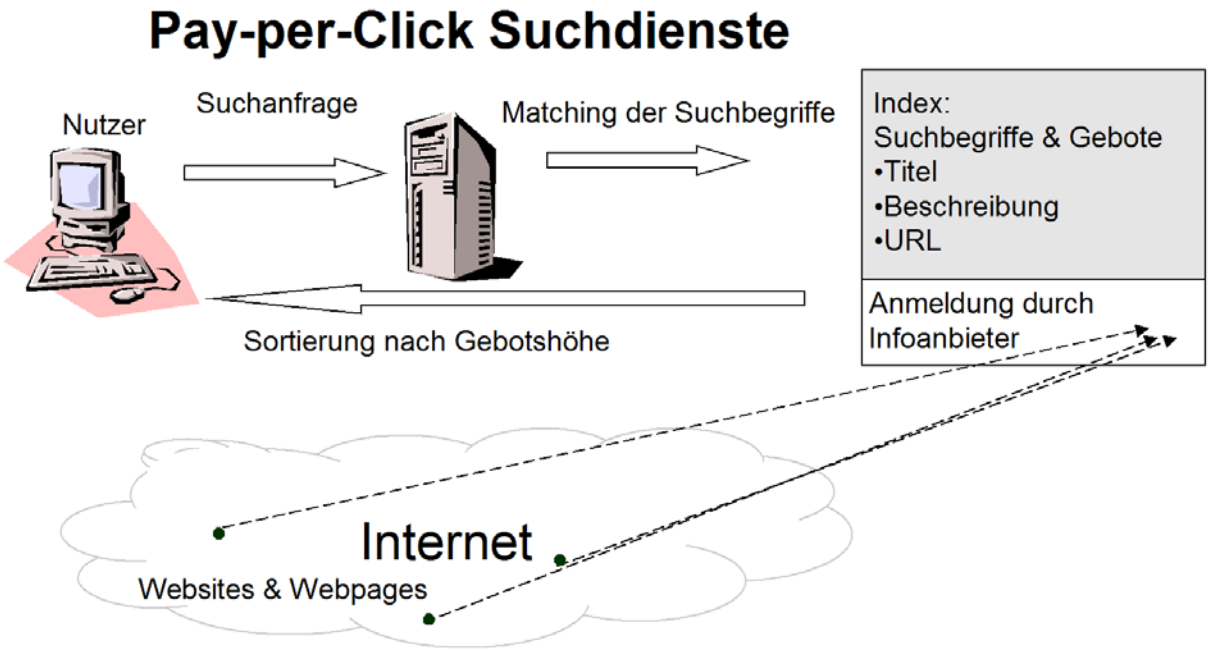


Abbildung 3: Schematische Darstellung Pay-per-Click, Suchdienste

Pay-per-Click-Suchdienste sortieren ihre Trefferlisten nach Gebotshöhe (Overture, Espotting) oder nach einer Kombination von Gebotshöhe und Klickhäufigkeit (Google). Die vorhergehende 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 rein ökonomisches Interesse daran haben, nur relevante Suchergebnisse zu positionieren. Die zu Grunde liegende Annahme ist, dass die Zahlungsbereitschaft des Informationsanbieters mit der Relevanz seines Angebots für den Informationsnachfrager korrespondiert.

## 2.4 Metasuchmaschinen

Metasuchmaschinen verfügen im Gegensatz zu den bislang vorgestellten Suchdienstetypen über keinen eigenen Index, sie leiten die Suchanfragen an andere Suchdienste und z.T. auch andere Quellen z.B. Lexika weiter und führen die zurückgelieferten Treffer in einer Trefferliste zusammen. Dabei werden Duplikate i.d.R. eliminiert und eine fusionierte Relevanzbewertung durchgeführt. Die Qualität der Ergebnisse von Metasuchmaschinen ist also direkt abhängig von der Qualität der Treffermengen der abgefragten

Suchdienste. Obwohl also Metasuchmaschinen über keinen eigenen Indexbestand verfügen, wird ihnen durch die parallele Abfrage mehrerer anderer Suchdienste die höchste Abdeckung des Web zugeschrieben.[15] Problematisch ist allerdings, dass diese theoretisch höhere Abdeckung von Metasuchmaschinen in der Praxis häufig keinen Nutzen bzw. Mehrwert bringt. Metasuchdienste sollten vor allem dann zum Einsatz kommen, wenn die einzelnen abgefragten Suchdienste zu wenige Treffer liefern. Jedoch leiten Suchdienste z.T. nur eine begrenzte Zahl ihrer Ergebnisse an Metasuchdienste weiter und haben so selbst höhere Trefferzahlen als die Metasuchmaschinen.[16] Ein direkter Nachteil ist auch, dass sich die Suchoptionen der zu Grunde liegenden Suchdienste nur teilweise nutzen lassen, denn Metasuchdienste können hier selbst nur den kleinsten gemeinsamen Nenner der abgefragten Suchdienste anbieten. Die Effektivität und Transparenz der Sortierkriterien von Metasuchmaschinen ist natürlich schwer einzuschätzen, da die Ergebnismengen den Treffermengen mehrerer heterogener Suchverfahren entstammen.

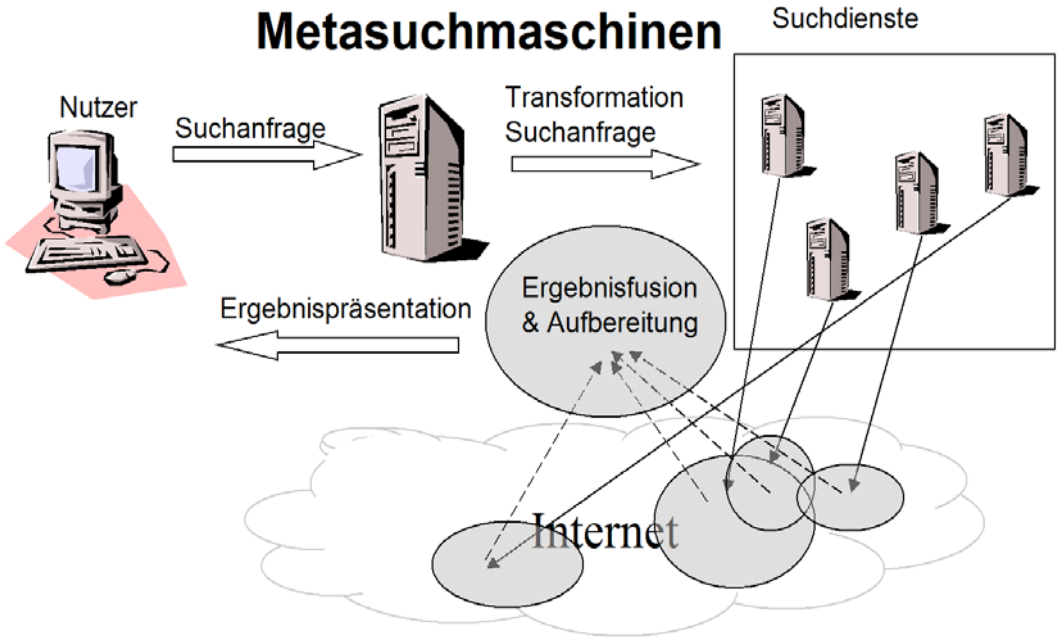


Abbildung 4: Schematische Darstellung Metasuchmaschinen

[15] Vgl. Suchmaschinenlabor RRZN & RVS Universität Hannover, URL http://metager.de/suma.html (letzter Zugriff 14.03.2004).

[16] So liefert die Metasuchmaschine Dogpile am 24.02.2004 zur Suchanfrage „information retrieval“ (als Phrase formuliert) 81 Treffer, URL http://www.dogpile.com/info.dogpl/search/web/%2522information%2Bretrieval%2522 (letzter Zugriff 14.03.2004), während der auch von Dogpile abgefragte Suchdienst Google.com zur selben Suchanfrage eine Zahl von 1 600 000 Treffern angibt, URL http://www.google.com/search?hl=en&ie=UTF-8&oe=UTF-8&q=%22information+retrieval%22&btnG=Google+Search (letzter Zugriff 14.03.2004).

# 3 Suchdienstemarkt - Verflechtung der Suchdienste

Die hier vorgenommene typologische Differenzierung nach Katalogen, Suchmaschinen, Pay-per-click-Diensten und Metasuchmaschinen findet in der gegenwärtigen Ausprägung des Suchdienstemarktes keine ausschließliche Entsprechung. Vielmehr integrieren die bekannten Suchdienste mehrere Suchdienstetypen.

Das heißt, 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 roboterbasierte 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. Overture selbst verwendet wiederum die Suchmaschinentreffer von Inktomi als sekundäre Ergebnisse. Die Suchtreffer von Google bilden die primären Suchergebnisse von Aol.com, Netscape.com und Google selbst. Die Katalogeinträge des Open Directory Projects werden unter anderem von Google, Lycos, Hotbot verwendet. Der Suchdienstemarkt ist also eng verflochten. Dabei bildet die Suchmaschinentechnologie zweier Anbieter - Google, und Yahoo - das technologische Rückgrat quasi aller bedeutenden Suchdienste. Der Markt für direkt bezahlte Einträge wird von Google und Overture - im Besitz von Yahoo - dominiert.

# 4 Zusammenfassung der Defizite von Suchdiensten

Bekannte bzw. populäre Suchdienste wie z.B. Google oder Yahoo liefern auf Suchanfragen Ergebnisse aller drei grundlegenden Suchverfahren aus. Dabei werden Katalogeinträge und roboterbasierte Suchergebnisse häufig in einer Trefferliste integriert. Diese nach Relevanzkriterien ausgelieferten Treffer aus Katalogen und Suchmaschinen werden als sogenannte „editorial“ oder „organic results“ benannt, wohingegen von Suchdiensten verkaufte Einträge als „paid listings“ bezeichnet werden.[17] Paid Listings sind Werbeeinträge, die durchaus relevant sein können, deren Erscheinen aber dennoch nicht auf den „neutralen“ Dokumenterfassungs- und Sortierkriterien beruht. Diese Tatsache

[17] Search Engine Optimization & Marketing Glossary, URL http://www.sempo.org/search-engine-marketing-glossary.php (letzter Zugriff 14.03.2004).

ist Nutzern häufig nicht bewusst,[18] obwohl Paid Listings inzwischen bei den Suchdiensten relativ eindeutig als Werbung gekennzeichnet werden - meist als „Sponsoren-Links“ oder „Gesponserte Websites“. Die editorial results geben nur ein unvollständiges Bild der Inhalte des Web wieder. 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.

Damit lassen sich also folgende grundlegende Defizite der populären Websuchdienste identifizieren:

**Validität/Qualität:** Die inhaltliche Validität/Qualität von Suchdienstergebnissen kann nur unzureichend gewährleistet werden**.** Eine redaktionelle Prüfung findet nur bei den Katalogeinträgen und den Paid Listings statt. Die Dokumentbeschaffungs-, Indexierungs- und Sortieralgorithmen von Suchmaschinen sind zudem manipulationsanfällig und können gezielt zur Platzierung wenig relevanter und irrelevanter Ergebnisse genutzt werden. Eine Prüfung der Validität der Suchmaschinenergebnisse kann mit Hilfe automatischer Bewertungskriterien, wenn überhaupt, dann nur ansatzweise bestimmt werden.

**Abdeckung/Deep-Web:** Die aktuellen Suchdienste im Internet ermöglichen keinen umfassenden Zugriff auf die Datenbestände des Web. Ein Großteil der vorhandenen Inhalte kann u.a. aus technischen Gründen nicht erfasst werden und taucht somit nicht in den Ergebnislisten der Suchdienste auf. 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 sich insbesondere sehr umfangreiche Informationsangebote traditioneller professioneller Informationsanbieter wie z.B. STN, Dialog oder Lexis-Nexis in geschützten Bereichen des Web befinden, werden diese qualitativ hochwertigen Wissensbasen durch Suchmaschinen kaum erschlossen. Natürlich würden die kostenpflichtigen Dokumente dieser Anbieter auch gar nicht in das Raster heutiger Suchdiensteanbieter wie z.B. Google passen –

[18] Laut einer Umfrage von Consumer WebWatch im Januar 2002 ist nur 43% der Suchmaschinenbenutzer bekannt, dass Suchmaschinenergebnisse z.T. aus kommerziellen Einträgen bestehen. Princeton Survey Research Associates, A Matter of Trust: What Users Want From Web Sites, Results of a National Survey of Internet Users for Consumer WebWatch, URL http://www.consumerwebwatch.com/news/report1.pdf. (letzter Zugriff 14.03.2004).

denn auf alles worauf eine Suchmaschine zugreifen kann, kann auch vom Benutzer zugegriffen werden. Ob aber Benutzer wirklich bereit sind für qualitativ hochwertige Informationen zu bezahlen, wird sich erst zeigen, wenn interessante Angebote auf Basis der gleichen einfachen Zugriffskompetenz, wie dies bei den aktuellen Suchdiensten im Internet der Fall ist, vorhanden sind.

**Neutralität/Objektivität:** Für die Dokumenterfassung, Indexierung und Sortierung sind inzwischen nicht nur technische oder inhaltliche Kriterien maßgeblich. Nach Zusammenbruch der Internet-Aktienmärkte und der damit verbundenen Investitionsflaute müssen Suchdiensteanbieter andere Finanzquellen erschließen. Bei Katalogen ist beispielsweise die Zahlungsbereitschaft der Informationsanbieter meist notwendige Voraussetzung einer redaktionellen Bewertung. Bei Suchmaschinen muss z.T. ebenfalls für die Aufnahme in den Index bezahlt werden. Die redaktionelle Neutralität von Treffermengen vermindert sich schon durch die Tatsache, dass die Angebote solventer Anbieter präferiert werden. Bei den Paid Listings schwindet diese sogar ganz zu Gunsten wirtschaftlicher Interessen. Obwohl bezahlte Suchmaschineneinträge häufig gekennzeichnet werden, ist es fraglich, ob den Benutzern die nicht mehr vorhandene inhaltliche Objektivität bei der Sortierung der Trefferlisten bewusst ist (vgl. [Consumer Webwatch 2002]).

# 5 Entwicklungstendenzen bei Websuchdiensten

Neue Suchtechnologien bzw. Retrievalansätze werden nicht von den bedeutenden Suchdiensten und Technologieanbietern entwickelt, sondern häufig zuerst von Suchdiensten umgesetzt, die im Vergleich zu den populären Suchdiensten eher ein Nischendasein führen und nur einer relativ geringen Zahl von Nutzern bekannt sind.

Eine neue Generation von Metasuchdiensten versucht seit einigen Jahren Wissensbestände des Deep Web zu erschließen. Spezielle Deep-Web-Verzeichnisse, etwa Completeplanet.com[19] katalogisieren ähnlich Webverzeichnissen Deep-Web-Ressourcen. Weiter gehend versuchen neue Metasuchdienste, etwa Turbo10.com[20] die direkte Abfrage einer vorgegebenen Kollektion von Deep-Web-Datenbanken zu ermöglichen. Die populären Suchdienste gehen in zunehmendem Maße dazu über, zusätzliche

[19] URL http://aip.completeplanet.com/ (letzter Zugriff 07.03.2004).
[20] URL http://turbo10.com/ (letzter Zugriff 07.03.2004).

Wissensbestände, insbesondere Produktdatenbanken oder geografische Suchoptionen,[21] bereit zu stellen. Damit lässt sich festhalten, dass es den Suchdiensten nicht nur gelingt mit dem Wachstum des Surface Web mitzuhalten, sondern zusätzlich mehr und mehr Wissensbestände bereitgestellt und insbesondere auch Inhalte des Deep Web recherchierbar werden. Damit sind Web-Suchdienste derzeit zwar noch sehr weit entfernt davon, umfassenden Zugriff auf die offen zugänglichen Wissensbestände des Surface und insbesondere des Deep Web zu ermöglichen, dennoch werden sie bezüglich der Abdeckung stetig besser.

Die Sicherstellung der Qualität der Suchergebnisse bleibt weiterhin hochproblematisch. [Machill & Welp 2003] konstantieren sogar eine Zunahme der Spamproblematik, so dass es eher unwahrscheinlich ist in diesem Bereich auf kurze Frist eine grundsätzliche Verbesserung zu erhoffen. Neue Ansätze von Suchdiensten wie Askjeeves „Smart Search“[22] und Yahoo Shortcuts,[23] die darauf abzielen, das Informationsbedürfnis automatisch zu antizipieren und in Abhängigkeit von der Art des Informationsbedürfnisses passende Dokumenttypen als Ergebnis auszugeben und somit den Informationsbedarf direkt zu befriedigen, funktionieren bislang nur für einen kleinen Teil, relativ klar typisierbarer, Suchanfragen. Diese Ansätze deuten jedoch an, dass die technologischen Entwicklungsbemühungen zukünftig weniger auf der (Weiter)entwicklung als neutral geltender und deshalb vermeintlich objektiver Sortierverfahren gelegt werden, sondern sich mehr und mehr an den subjektiven Bedürfnissen des Nutzers orientieren.

Ob sich die oft geforderte Kennzeichnung kommerzieller Ergebnisse tatsächlich derart manifestiert, dass dem typischen Suchdienstenutzer bewusst wird, dass die Sichtbarkeit, Rangfolge der Suchergebnisse, zunehmend durch Marketingstrategien und ökonomische Kompetenz der Informationsanbieter sowie der Vermarktungsbereitschaft der Suchdienste beeinflusst werden [Griesbaum 2003], bleibt noch abzuwarten. Nach den neuesten Entwicklungen zu urteilen sieht es eher nicht danach aus: Yahoo´s neuartiges

[21] Danny Sullivan, Local Search Part 1: New Developments In Local Search & Moves By Overture. URL http://searchenginewatch.com/searchday/article.php/3091341 (letzter Zugriff 11.03.2004).
[22] Chris Sherman, Ask Jeeves Serves Up New Answers. URL http://searchenginewatch.com/searchday/article.php/3067941 (letzter Zugriff 09.03.04).
[23] URL http://help.yahoo.com/help/us/ysearch/tips/tips-01.html (letzter Zugriff 08.03.04).

Paid-Inclusion-Programm[24] weckt weitere Befürchtungen hinsichtlich der Glaubwürdigkeit und Neutralität relevanzsortierter Suchmaschinentreffer.[25]

# 6 Literaturverzeichnis

Hinweise zu einschlägigen Online-Artikeln sind in den Fußnoten zu finden.

Ohne Autor (2002). What Users Want From Web Sites. Results of a National Survey of Internet Users for Consumer WebWatch. URL http://www.consumerwebwatch.com/news/report1.pdf.

Bergman, M. (2000). The Deep Web: Surfacing Hidden Value. White paper. URL http://www.brightplanet.com/pdf/deepwebwhitepaper.pdf.

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.); Medford NJ: CyberAge, 175-195.

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

Griesbaum, J. (2003). Unbeschränkter Zugang zu Wissen? Leistungsfähigkeit und Grenzen von Suchdiensten im Web. Zwischen informationeller Absicherung und manipulierter Information. Frankfurt, M.: 37-50. URL http://www.inf.uni-konstanz.de/%7Egriesbau/files/unbeschraenkter_zugang_zu_wissen_dgi_03.pdf.

Jansen, B., Spink, A., und Saracevic, T. (2000). 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.

Klatt, R., Gavriilidis, K., Kleinsimlinghaus, K., Feldmann, M., u.a. (2001). Nutzung elektronischer wissenschaftlicher Information in der Hochschulausbildung : Barrieren und Potenziale der innovativen Mediennutzung im Lernalltag der Hochschulen: Kurzfassung. URL http://www.stefi.de/download/kurzfas.pdf.

[24] Yahoo Site Match. URL http://www.content.overture.com/d/USm/ays/bjump/sm.jhtml (letzter Zugriff 13.03.2204).

[25] Pandia Search Engine News, Yahoo! and Overture's new paid inclusion program 01.03.2004. URL http://www.pandia.com/sw-2004/09-overture.html (letzter Zugriff 13.03.2004).

Machill, M. & Welp, C. (2003). Wegweiser im Netz. Qualität und Nutzung von Suchmaschinen. Bertelsmann Stiftung.