Handbuch Internet-Suchmaschinen
D. Lewandowski (Hrsg.)

Typologie der Suchdienste im Internet

Joachim Griesbaum [a], Bernard Bekavac [b], Marc Rittberger[c]
[a] *Universität Hildesheim*
Marienburger Platz 22
31141 Hildesheim

[b] *Hochschule für Technik und Wirtschaft Chur*
Ringstraße
CH-7000 Chur
bernard.bekavac@fh-htwchur.ch

[c] *Deutsches Institut für Internationale Pädagogische Forschung*
Schloßstraße 29
60486 Frankfurt
rittberger@dipf.de

Abstract. Der folgende Beitrag strukturiert das Themenfeld Web Information Retrieval anhand einer Typologie der Suchdienste im WWW. Damit soll ein grundlegender Überblick über die derzeit vorhandenen konzeptionellen und technologischen Ansätze der Informationssuche im Internet gewonnen werden. Vor diesem Hintergrund werden die elementaren Suchdienstetypen hinsichtlich ihrer methodischen Verfahren zur Dokumenterschließung, der Informationsaufbereitung und der Dokumentselektion bzw. der Rankingfaktoren vorgestellt. Dabei werden auch aktuelle Entwicklungstendenzen diskutiert, in denen es derzeit primär darum geht, maschinelle Algorithmen und das Wissen der bzw. über die Nutzer gewinnbringend miteinander zu kombinieren. Abschließend wird ein Einblick in die gegenwärtige Ausprägung des Suchdienstemarktes gegeben und somit insgesamt ein konzeptueller Überblick zu Suchdiensten im Web erschlossen.

Keywords. Suchdienste im Internet, Suchmaschinen, Typologie

Einleitung

Internetsuchdienste können nach vielfältigen Kriterien differenziert werden. Eine grundlegende und prinzipielle Unterscheidung ist die nach intellektuell und manuell erstellten Dokumentsammlungen (Katalogen) auf der einen und algorithmenbasierten automatischen Systemen (Suchmaschinen) auf der anderen Seite [1]. Ebenso ist es möglich, zwischen „Universal-“ und Spezialsuchdiensten zu differenzieren[1], sei es nun bezüglich thematischer, dokumenttypbezogener/dokumentformatbezogener oder geografischer Hinsicht.

Die folgende Darstellung orientiert sich an der grundlegenden Unterscheidung zwischen manuell und automatisch erstellten Dokumentsammlungen. Zunächst werden Webkataloge und Dienste, die auf einer gemeinschaftlichen Indexierung durch die Nutzer beruhen (Social Tagging), insbesondere Social Bookmarkdienste, dargestellt.

Im nächsten Schritt werden automatische Systeme, d.h. die grundlegende Funktionsweise von Suchmaschinen dargestellt. Anschließend werden Ansätze von Spezialsuchdiensten und von Metasuchmaschinen skizziert. Auf dieser Grundlage wird eine Einschätzung des Suchdienstemarktes gegeben und aufgezeigt, wie Suchmaschinenwerbung, die derzeitige finanzielle Grundlage der Websuche, in Form von „Pay per Click"-Diensten funktioniert. Zum Schluss wird eine zusammenfassende Einschätzung zum derzeitigen Stand und der weiteren Entwicklung des Web Information Retrieval vorgenommen.

Den Lesern soll einerseits ein Überblick über den State of the Art und wichtige Entwicklungsansätze bei Websuchdiensten verschafft werden, und zugleich sollen sie dazu angeregt werden, jenseits von Google und Yahoo!, weitere Suchdienste auszuprobieren und gegebenenfalls in ihr Rechercheportfolio aufzunehmen.

1. Webkataloge

Webkataloge oder -verzeichnisse stellen manuell ausgewählte bzw. zusammengestellte Linksammlungen dar [2]. Die Verweise werden in einer mono- oder polyhierarchischen Struktur eingeordnet. Zu den einzelnen Links werden ein Titel, die URL, ein Beschreibungstext und gegebenenfalls weitere Zusatzinformationen, oft geografische Angaben, erfasst.

Einträge in Kataloge sind oftmals kostenpflichtig und werden i.d.R. von Websitebetreibern vorgenommen, die sich davon eine bessere Sichtbarkeit für Internetnutzer und Suchmaschinen und damit eine höhere Besucherzahl erhoffen. Abbildung 1 illustriert den konzeptuellen Aufbau von Webkatalogen.

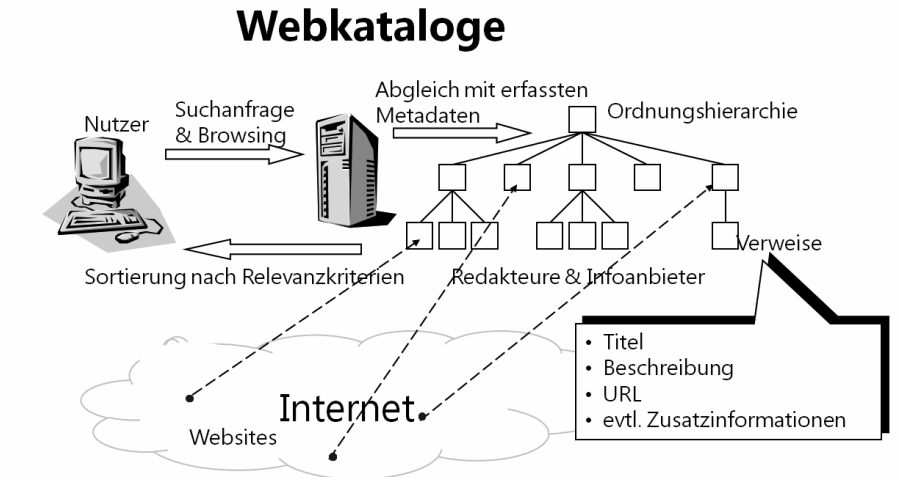

Abbildung 1. Aufbau von Webkatalogen

Im Allgemeinen entscheidet eine redaktionelle Begutachtung durch Mitarbeiter des Katalogbetreibers über die Aufnahme in den Katalog und die Zuordnung in der Hierarchie. Meist werden nur Homepages bzw. Einstiegsseiten und keine einzelnen (Unter-)Seiten angenommen, d.h. die detaillierten Inhalte der einzelnen Webseiten einer Domain werden nicht erfasst. Vielmehr gibt der jeweilige Beschreibungstext das Gesamtthema einer Website wieder. Kataloge bieten also keinen Zugriff auf die

Volltexte von Dokumenten, sondern ermöglichen die Suche nach Websites durch die Navigation in der hierarchisch aufgebauten Rubrikstruktur bzw. mittels einer Stichwortsuche in den erfassten Metadaten (Titel, URL, Beschreibungstext) [2].

Je nach Zugriffsweise werden die Treffer nach unterschiedlichen Kriterien sortiert. Navigiert der Nutzer durch den Katalog, ist die hierarchische Position im Verzeichnis bzw. die meist nach Eintragsdatum oder Alphabet vorgenommene Reihenfolge in der jeweiligen Rubrik entscheidend. Führt der Nutzer eine Stichwortsuche durch, so wird i.d.R. ein Relevanzranking mit Hilfe einfacher Gewichtungsverfahren vorgenommen. Ergänzend können, in Abhängigkeit von der Einstufung durch Redakteure oder der Zahlungsbereitschaft des jeweiligen Websitebetreibers, Katalogeinträge höher sortiert oder bezüglich der Darstellung, etwa durch eine farbliche Kennzeichnung, hervorgehoben werden[1].

Will man die Vor- und Nachteile von Katalogen diskutieren, ist zunächst zu konstatieren, dass die redaktionelle Betreuung aus einer theoretischen Perspektive für eine hohe Qualität der Einträge bürgt und aus diesem Grund erwartet werden kann, dass Spam[2] in Webkatalogen eher selten zu finden ist. Des Weiteren gruppieren Kataloge ihre Einträge nach, zumindest im jeweiligen Katalogkontext, konsistenten thematischen und hierarchischen Prinzipien. Damit stellen sie Kontextinformationen für die einzelnen Einträge zur Verfügung, welche vor allem bei unspezifischen Informationsbedürfnissen wie dem Einstieg in ein neues Thema positiv, z.B. über Browsing-Effekte[3], zum Tragen kommen.

Neben diesen Vorteilen weisen Kataloge aber auch Nachteile auf. Zunächst gilt, dass ein hoher „Wartungsaufwand“ notwendig ist, da einmal vorgenommene Katalogeinträge regelmäßig auf Aktualität überprüft werden müssen. Häufig sind die zur Anordnung der Katalogeinträge erstellten Hierarchien zunächst pragmatisch definiert worden und dann organisch gewachsen. Qualitätskriterien, wie sie etwa für hierarchische Klassifikationen im Bibliotheksbereich von Bedeutung sind, spielen dabei eher eine untergeordnete Rolle. Insbesondere bei Katalogen, die den Versuch unternehmen, alle Themen des Web zu erfassen stellt die wohl größte Schwäche gerade im Vergleich zu den weiter unten dargestellten Suchmaschinen, die geringe Abdeckung dar. Der größte Webkatalog, das Open Directory Project, verzeichnet im April 2008 rund 4 600 000 Websites[4]. Im Vergleich zu den Milliarden Einträgen, welche Suchmaschinen aufweisen, eine marginale Größe. Die Kombination von geringer Abdeckung und fehlender Möglichkeit der Volltextsuche hat zur Folge, dass allgemeine Kataloge für die Befriedung spezifischer Informationsbedürfnisse oftmals weniger geeignet sind als Suchmaschinen.

Historisch betrachtet stellen Webkataloge die erste Form der globalen Suche im WWW dar. Yahoo, 1994 gegründet[5], ist wahrscheinlich das prominenteste Beispiel eines Webkatalogs. Daran lässt sich zugleich die in den vergangenen Jahren gesunkene Bedeutung von Webkatalogen veranschaulichen. Ab 1998 wurde die Yahoo-Websuche,

[1] Vgl. etwa die „Premiumdienste“ des Webkatalogs Allesklar.de, URL https: //listing.allesklar.de/listingshop2005/index.php?mid=2 (Letzter Zugriff 24.04.2008).

[2] Im Kontext von Suchmaschinen bezeichnet SEMPO, eine globale Non-Profit-Organization (Dachorganisation) der Suchmaschinenindustrie, Spam als „Any search marketing method that a search engine deems to be detrimental to its efforts to deliver relevant, quality search result“, URL http: //www.sempo.org/learning_center/sem_glossary#s (Letzter Zugriff 10.04.2008).

[3] Der Browsing-Effekt besteht darin, dass der Nutzer beim Explorieren auch solche Informationen entdeckt, mit denen er anfänglich gar nicht gerechnet hat, ähnlich dem Stöbern in einer Bibliothek.

[4] Vgl. URL http: //www.dmoz.org/ (Letzter Zugriff 24.04.2008).

[5] URL http: //yahoo.client.shareholder.com/press/overview.cfm (Letzter Zugriff 09.04.2008).

die bis dahin rein katalogbasiert war, mit Suchmaschinentechnologie ergänzt, die supplementäre Suchergebnisse bereitstellte. Ab dem Jahr 2002 wurden dann Suchmaschinentreffer von Google und Yahoo-Katalogergebnisse in einer Suchergebnisliste zusammengeführt[6].
Nachdem Yahoo 2003 die Suchmaschinentechnologieanbieter Overture und Altavista erwarb, entwickelte es eine eigene Suchmaschine. Diese stellt seit dem Jahr 2004 den Kern der Suchfunktionalität auf Yahoo-Seiten[7] dar. Den Webkatalog findet man bei Yahoo mittlerweile nur noch als Link bei den supplementären Suchdiensten und Spezialindizes.

Das Beispiel Yahoo veranschaulicht sehr deutlich, dass und wie Webkataloge in den letzten 10 Jahren mehr und mehr von Suchmaschinen verdrängt bzw. ersetzt wurden. Dennoch ist es wenig sinnvoll, die manuelle Erschließung von Dokumenten in Form von Katalogen als überholt zu betrachten. Zunächst ist festzuhalten, dass nach wie vor eine Vielzahl spezialisierter Kataloge, beispielsweise Firmenverzeichnisse und thematisch spezialisierte Verzeichnisse im Web zu finden sind. Aus Nutzersicht sind solche Dienste hervorragende Anlaufstellen, um sich einen Überblick über spezielle Themenbereiche zu verschaffen[8]. Des Weiteren stellen Kataloge zu spezialisierten Suchdiensten derzeit immer noch quasi *das* Meta-Informationsinstrument im Internet dar. Denn derartige Kataloge bieten die Möglichkeit, einen Überblick über vorhandene Suchoptionen im Web zu gewinnen und ermöglichen es, in Abhängigkeit von dem jeweiligen Informationsbedürfnis die geeignetsten Suchdienste auszuwählen. Oftmals erfassen solche Kataloge eine Vielzahl von hochwertigen (Fach-)Datenbanken, deren Inhalte Suchmaschinen verschlossen bleiben und bieten damit indirekt Zugriff auf das für Google & Co. verschlossene Deep Web[9]. Tabelle 1 nennt beispielhaft einige derartige Meta-Verzeichnisse:

Tabelle 1. Kataloge und Katalogrubriken, in denen Websuchdienste verzeichnet sind

Katalog	URL
DBIS Datenbank Infosystem: Listet 5.700 Datenbanken	http://rzblx10.uni-regensburg.de/dbinfo/fachliste.php?lett=l
Bibl. Linksammlung mit ca. 50.000 Internetquellen	http://digilink.digibib.net/wk/links.pl
Inforunner: Datenbank-Verzeichnis mit direktem Zugriff auf die Datenbanken und Archive	http://www.inforunner.de/
Suchfibel: Verlinkt auf rund 2000 katalogisierte, beschriebene und bewertete Suchdienste	http://www.suchfibel.de/

Des Weiteren weisen Kataloge auch für Informationsanbieter Mehrwerte auf. Die weiter oben bereits angesprochene höhere Sichtbarkeit für in Katalogen eingetragene Informationsanbieter resultiert nicht zuletzt aus der Tatsache, dass Kataloge für

[6] Danny Sullivan 09.10.2002, Yahoo Renews With Google, Changes Results URL http://searchenginewatch.com/showPage.html?page=2165081 (Letzter Zugriff 09.04.2008).

[7] Und auch weiterer, meist Yahoo zugehöriger Websuchdienste, wie Altavista oder AlltheWeb.

[8] Ein Beispiel stellt etwa der Suchmaschinen-Tippgeber.de dar. Ein Webkatalog mit rund 1200 Einträgen zum Thema Suchmaschinen und Suchmaschinenmarketing.

[9] Alle Inhalte des Web, auf die aufgrund von Zugangsbeschränkungen durch die Anbieter oder technischer Restriktionen von Suchrobotern nicht über Suchmaschinen zugegriffen werden kann. Zitat: „Text pages, files, or other often high-quality authoritative information available via the World Wide Web that general-purpose search engines cannot, due to technical limitations, or will not, due to deliberate choice, add to their indices of Web pages“[3].

Suchmaschinen ideale Startpunkte zur Dokumentbeschaffung darstellen. Websites, die in Katalogen eingetragen sind, besitzen eine höhere Wahrscheinlichkeit, in die Indizes der Suchmaschinen aufgenommen zu werden, als Sites, die nicht gelistet sind. Zugleich resultiert aus Katalogeinträgen oftmals eine höhere Linkpopularität [4]. Wie weiter unten ausgeführt wird, stellt die Linkpopularität einen wichtigen Rankingfaktor dar. Hinzu kommt, dass es aus Suchmaschinenperspektive sinnvoll ist, in Katalogen eingetragenen Seiten einen Qualitätsbonus zu geben, weil davon ausgegangen werden kann, dass die gelisteten Seiten gewissen editorialen Qualitätskriterien entsprechen bzw. zumindest keinen Spam darstellen.

Als ein Beispiel für einen spezialisierten Webkatalog, der die genannten Positivkriterien in hohem Maße erfüllt und damit auch eine Selektions- und Ordnungsfunktion für seine Domäne übernimmt, gilt der Deutsche Bildungsserver[10]. Er ist mit dem Auftrag versehen, relevante Informationen insbesondere aus den durch die föderale Struktur der Bundesrepublik Deutschland informationell stark zersplitterten öffentlichen Quellen zu strukturieren. Dabei werden, aufsetzend auf einem von einem fachlich und methodisch ausgebildeten Redakteursteam ausgewählten Satz von Informationshinweisen, zu allen Themen des Bildungswesens Links und Informationen angeboten. Jeder Link oder jede Linkgruppe wird inhaltlich erschlossen und durch kurze Zusammenfassungen erläutert. Neben dem thematischen Zugang (z.B. Schule, Erwachsenenbildung…) besteht auch die Möglichkeit, zielgruppenorientiert (z.B. Eltern, Studierende…) die zur Verfügung stehenden Informationen zu erarbeiten. Auch hier ist eine Weiterentwicklung zu beobachten. Zusätzlich zu den beschriebenen Katalogangeboten wird das Angebot durch eine Vielzahl spezialisierter Informationsangebote und -portale (z.B. zur Leseförderung oder zu innovativen Vorhaben von Bund und Ländern im Bildungsbereich), durch Datenbanken und eine höhere Nutzerbeteiligung erlaubende Web 2.0-Angebote (z.B. ein Fork der Wikipedia) ergänzt. Der Deutsche Bildungsserver weist aufgrund seiner zentralen Position in der Bildungsinformation über 1 Million Zugriffe pro Monat auf. Die hohe Popularität des Angebots und der eingebundenen Links, die starke Vernetzung des Deutschen Bildungsservers im Web, die Seriosität des Anbieters und die hohe Anzahl der Benutzer führen dazu, dass Verweise des Deutschen Bildungsservers bei Suchmaschinen sehr hoch gerankt werden. Das zeigt sich beispielsweise an den Analysen von seitwert.de, einem Anbieter automatisierter Bewertungsverfahren für Web-Domains. Seitwert.de rankt den Deutschen Bildungsserver unter den TOP 500 Seiten in Deutschland. Dabei sind besonders die Werte des Suchmaschinenrankings von seitwert.de als sehr positiv bewertet worden (Google: „Sichtbarkeit von www.bildungsserver.de ist extrem hoch (86200)“; Yahoo: „www.bildungsserver.de hat sehr viele Backlinks (189.655)“).[11]

Des Weiteren sind in den letzten Jahren vermehrt Ansätze zu beobachten, die dahin zielen, das Paradigma „Human-powered Search“ auch für Universalsuchdienste zu reetablieren bzw. diesem zu größerer Popularität zu verhelfen. Als Beispiel für Webkataloge lässt sich Mahalo[12] heranziehen. In Mahalo stellen Editoren, sogenannte Guides, „handverlesene“ Ergebnisse zu populären Suchanfragen zusammen. Die Trefferseiten sind nach verschiedenen Kriterien vorstrukturiert. So liefert etwa die in

[10] URL http://www.bildungsserver.de (Letzter Zugriff 09.05.2008).
[11] URL http://www.seitwert.de wurde mit der URL http://www.bildungsserver.de am 10.6.08 getestet.
[12] URL http://www.mahalo.com/ (Letzter Zugriff 24.04.2008).

Abbildung 2 dargestellte Trefferseite zur Anfrage „Search Engine Optimization“ die Gliederung:

- The Mahalo Top 7
- Search Engine Optimization Information
- SEO Firms and Companies
- Prominent People in SEO
- Search Engine Optimization News
- Search Engine Optimization Blogs
- Search Engine Optimization Tips and Tools
- Search Engine Optimization Conferences
- Related Searches.

Abbildung 2. Ergebnisseite des Webkatalogs Mahalo[13]

Damit bietet Mahalo eine deutlich spezifischere Sicht auf die vorhandenen Themen bzw. Rubriken als andere Webkataloge und auch Suchmaschinen. Mahalo nutzt u.a. Suchmaschinen als Backend. D.h. falls in Mahalo keine Treffer gefunden werden, werden Ergebnisse anderer Suchdienste ausgeliefert. Das ist im Prinzip genau das Konzept, welches Yahoo von 1998-2002 verfolgte. Ob und inwieweit sich der Ansatz von Mahalo durchsetzt oder dauerhaft etabliert, bleibt abzuwarten. Beide Beispiele, Deutscher Bildungsserver und Mahalo, verdeutlichen aber, dass die Idee, Inhalte des Web in Form von Katalogen zu erschließen, nach wie vor sinnvoll und lebendig ist.

[13] URL http: //www.mahalo.com/Search_Engine_Optimization (Letzter Zugriff 10.04.2008).

2. Social Bookmarkdienste & Tagging Communities

Während bei Katalogen Inhalte mit Hilfe redaktioneller Kräfte erschlossen oder zumindest geprüft werden, sind Dienste, die auf einer gemeinschaftlichen Indexierung durch die Nutzer aufsetzen, dadurch gekennzeichnet, dass sie die jeweiligen Objekte frei, d.h. weitgehend ohne strukturelle oder inhaltliche Vorgaben und Kontrolle in das System einspeisen und inhaltlich erschließen (verschlagworten). Die Idee, Objekte durch Nutzer zu verschlagworten, wurde mit Diensten wie Flickr.com (Bildercommunity) bzw. Del.icio.us (Social Bookmarks) populär und wird als ein zentraler Entwicklungstrend des Web 2.0 gesehen[14] [6]. Die nachfolgende Darstellung bezieht sich primär auf Social Bookmarkdienste, da diese am ehesten als „Universalsuchdienste" betrachtet werden können, während andere Tagging Communities wie z.B.

- Flickr.com: Bilder
- Slideshare.net: Präsentationen
- Scribd.com: Textdokumente
- Youtube.com: Videos usw.

aus der Sicht eines Informationssuchenden quasi dokumenttypbezogene Spezialsuchdienste darstellen.

Social Bookmarkdienste ermöglichen es allen Internetnutzern, Bookmarks online anzulegen und auf dem Server des jeweiligen Anbieters zu speichern. Populäre Social Bookmarkdienste sind z.B. im internationalen Raum Del.icio.us [15] und im deutschsprachigen Raum Mister-Wong.de[16]. Im Unterschied zu Katalogen werden die Websites nicht in eine, meist vorab kreierte, Ordnungshierarchie eingefügt, sondern mit sogenannten Tags (freie Schlagworte) versehen. Bookmarks können kommentiert, verschlagwortet und anderen Benutzern öffentlich zugänglich gemacht werden. Die Summe aller Tags aller Nutzer wird auch als Folksonomy bezeichnet. Die von den Nutzern vergebenen Schlagwörter unterliegen i.d.R. keinerlei terminologischer Kontrolle oder Struktur. Tags werden so auch dazu benutzt, um neben inhaltsbezogenen auch formale Aspekte (z.B. Datum) oder gar individuelle emotionale Bezüge („cool") auszudrücken [7]. Das gemeinsame Indexieren bzw. die daraus entstehende Folksonomy bildet die Grundlage der in Social Bookmarkdiensten vorhandenen Such- und Browsingfunktionen. Nutzer können über Tagclouds[17] (ähnlich Katalogrubriken) „Populär"- oder „New"-Listen navigieren bzw. über eine Stichwortsuche in den erfassten Daten (Titel/Beschreibung/Tags) auf die vorhandenen Bookmarks zugreifen. Abbildung 3 veranschaulicht den konzeptionellen Aufbau von Social Bookmarkdiensten.

[14] Vgl. URL http: //de.wikipedia.org/wiki/Web_2.0 (Letzter Zugriff 10.04.2008).
[15] 2005 von Yahoo akquiriert.
[16] Für eine Übersicht vgl. den Webartikel „0 Largest Social Bookmarking Sites" URL http: //www.ebizmba.com/articles/social30 (Letzter Zugriff 10.04.2008). Des Weiteren existieren mit Bibsonomy.org und Connotea.org auch Dienste, die speziell auf ein wissenschaftliches Publikum ausgerichtet sind.
[17] Wortwolken.

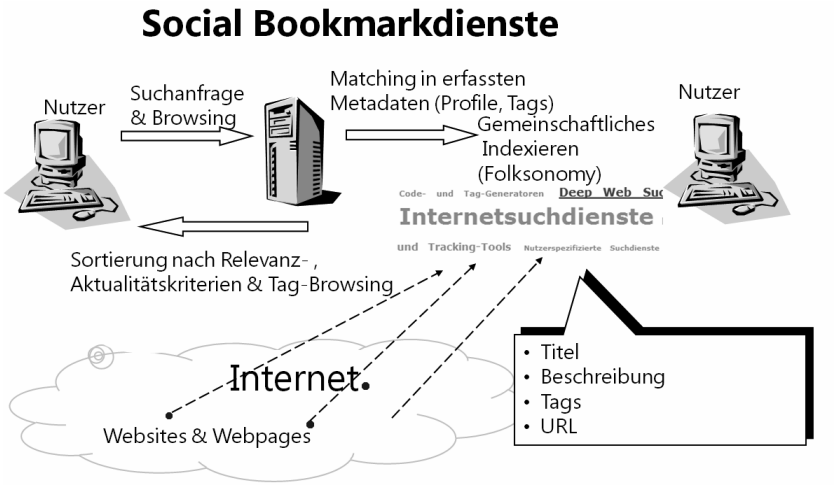

Abbildung 3. Aufbau von Social Bookmarkdiensten

Bei der Ausgabe von Suchergebnissen kann neben einem Termabgleich zwischen Suchanfrage und Bookmarks auch die Popularität[18] der Einträge berücksichtigt werden [8]. Abbildung 4 zeigt die Ergebnisliste in Del.icio.us zur Suchanfrage „Search Engine Optimization".

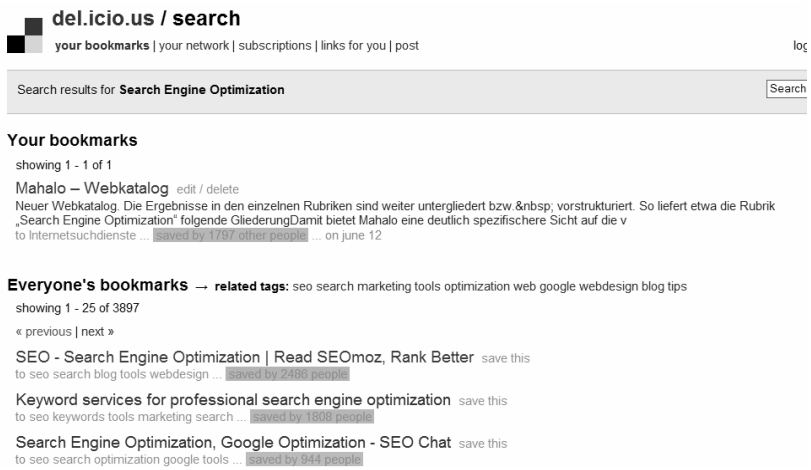

Abbildung 4. Ergebnisliste in Del.icio.us[19]

Die Abbildung veranschaulicht, wie neben Treffern in den selbst eingefügten Bookmarks („Your bookmarks") auch Treffer aus der Gruppe aller Nutzer (Everyone´s bookmarks") zurückgeliefert werden. Des Weiteren ist es möglich, die Nutzer anzuzeigen, welche die jeweiligen Treffer als Bookmarks gespeichert haben. Social Bookmarkdienste stellen damit also sowohl einen thematischen als auch einen

[18] Popularität bedeutet Anzahl der Nutzer, welche das jeweilige Bookmark gespeichert haben.
[19] URL http: //del.icio.us/search/?fr=del_icio_us&p=Search+Engine+Optimization&type=all (Letzter Zugriff 10.04.2008).

personenbezogenen Zugang zu Informationsressourcen bereit und können in diesem Sinne etwa auch zum Aufspüren von „Experten“ verwendet werden.

Tagging-basierte Communities stehen am Beginn ihrer Entwicklung und weisen, wie anhand der eben angeführten Möglichkeit zur „Expertensuche“ angedeutet, vielfältiges Potenzial zur Verbesserung des Information Retrieval im Web auf. Hinsichtlich der Qualität der Erschließung lässt sich aus theoretischer Perspektive konstatieren, dass einerseits durch die hohe Indexierungsbreite bei populären Einträgen in und mit der jeweiligen Folksonomy ein weiter semantischer Raum aufgespannt wird, in welchem die Objekte vielfältig beschrieben und repräsentiert werden. Auf der anderen Seite ergeben sich aus der Sicht Informationssuchender zunächst die typischen Synonym- und Homonymprobleme, die freiem Vokabular zugeschrieben werden [9]. Hinzu kommen Rechtschreibfehler, Singular-, Pluralansetzungen, unterschiedliche Sprachen und auch die Frage der Kombination mehrerer Worte in Tags, die sich negativ auf die Indexierungskonsistenz auswirken. Damit wird vor allem die Wiederauffindbarkeit von Objekten erschwert, die nur wenige Tags aufweisen. Bzgl. der Qualität der erschlossenen Objekte selbst lässt sich einerseits argumentieren, dass die manuelle Dokumentbeschaffung für eine hohe Qualität bürge. Dies ist zunächst sicher richtig. Allerdings sind die Kontrollmechanismen bei Social Bookmarkdiensten im Vergleich zu Webkatalogen stark vermindert [20]. Zugleich werden mit der zunehmenden Verbreitung und Nutzung von Social Bookmarkdiensten, Tagging Communities und anderen Sozialen Medien deren Anwendung bzw. Einbindung in das Suchmaschinenmarketing[21] zunehmend attraktiver. Der Grund hierfür liegt darin, dass der Reputations- und Linkpopularitätsaufbau über solche Dienste auch als Erfolgsfaktor gesehen wird, um in Google & Co. eine hohe Sichtbarkeit zu erreichen. Der niedrigschwellige Zugang erleichtert damit auch Möglichkeiten des Missbrauchs zu Spamzwecken. Derzeit ist weitgehend unerforscht, welches Ausmaß dieses Spamproblem einnimmt.

Zusammenfassend bleibt festzuhalten, dass Tagging Communities, insbesondere Social Bookmarkdienste, als die Katalysatoren eines sich neu etablierenden Paradigmas der inhaltlichen Erschließung im Web betrachtet werden können [10]. Social Bookmarkdienste lassen sich im Kontext des Web Information Retrieval quasi als neue Form von nutzergenerierten Linksammlungen begreifen, die eine große Ähnlichkeit zu Webkatalogen aufweisen, dabei allerdings aufgrund der größeren Zahl von „Editoren“ deutlich umfangreicher und aktueller sind und in ihrer hierarchischen Struktur erheblich freier bzw. chaotischer ausfallen. Im Vergleich zu Suchmaschinen weisen aber auch Social Bookmarkdienste immer noch eine geringe Abdeckung auf. Um diese Aussage zu veranschaulichen, sind in Tabelle 2 die Trefferzahlen des Open Directory-Webkatalogs, des Social Bookmarkdienstes Del.icio.us und der

[20] In [44] werden die intellektuelle inhaltliche Erschließung und social tagging-Mechanismen miteinander verglichen. [44] kommt zu dem Schluss, dass social tagging das inhaltliche Erschließen auf keinen Fall wird ersetzen können, aber eine Kombination oder Ergänzung beider Verfahren für eine bessere Findbarkeit von Dokumenten wünschenswert ist.

[21] Als Suchmaschinenmarketing können alle Maßnahmen verstanden werden, die dazu dienen, in Suchdiensten eine höhere Sichtbarkeit zu erreichen. Die Motivation, Suchmaschinenmarketing zu betreiben, ist i.d.R. ökonomischer Natur und zielt, etwas vereinfacht ausgedrückt, dahin, durch eine größere Besucherzahl höhere Umsätze zu erreichen.

Suchmaschine Google für vier bei „Google Zeitgeist“ gelistete Suchanfragen deutscher Nutzer[22] aufgeführt.

Tabelle 2. Trefferzahlenvergleich Open Directory, Del.icio.us, Google

Suchanfrage	Open Directory	Del.icio.us	Google.de
playstation	2 410	29 787	187 000 000
nail art	1 005	833	3 310 000
pink floyd	238	2219	26 000 000
sonnenbrille	1	108	1 860 000

Der Vergleich ist tentativ, illustriert aber die Vermutung, dass Social Bookmarkdienste einerseits i.d.R. eine höhere Trefferzahl als Kataloge aufweisen und andererseits Kataloge und Social Bookmarkdienste im Vergleich zu Suchmaschinen insgesamt nur eine marginale Abdeckung erreichen. Offen bleibt in diesem Kontext, ob das Suchbedürfnis des Benutzers immer Antworten aus einem vollständigen Spektrum erwartet oder ob er sich in vielen Fällen auch mit einer Auswahl begnügt. Zu vermuten wäre, dass diese Auswahl besonders dann interessant wird, wenn sie qualitätsgeprüfte und hochwertige Ergebnisse zulässt. Bzgl. der qualitativen Einschätzung von Social Bookmarkdiensten wurde das Problem der Indexierungskonsistenz und Spamanfälligkeit bereits angesprochen. Abbildung 5 veranschaulicht anhand der Ergebnismenge zur Suchanfrage „Sonnenbrille“ in Del.icio.us, dass Social Bookmarkdienste, aus der Perspektive eines Suchsystems betrachtet, des Weiteren noch erhebliches Verbesserungspotenzial bei der Darstellung der Suchtreffer aufweisen: Suchtreffer werden oftmals nur sehr knapp mit einem Titel angezeigt, der zudem u.U. im jeweiligen Suchkontext nur wenig aussagekräftig ist.

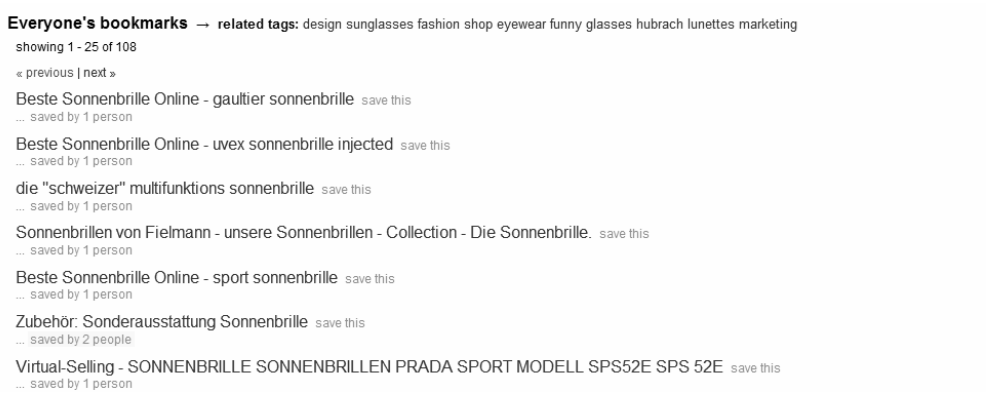

Abbildung 5. Trefferdarstellung in Del.icio.us

Das qualitative Potenzial von Social Bookmarkdiensten für die Websuche ist bislang wenig erforscht. Es finden sich anekdotische Aussagen, z.B. in Diskussionen in Webforen oder Blogs[23], wissenschaftliche Untersuchungen jedoch stehen noch

[22] Es handelt sich dabei um die Top 5 der „Top Gaining Queries“, also der Suchanfragen, die im Vergleich zum Vormonat am stärksten zugelegt haben, für den Monat February 2008, vgl. http://www.google.com/intl/en/press/intl-zeitgeist.html#de [letzter Zugriff 11.04.2008].

[23] Vgl. etwa den Beitrag „The Search Engine That's Already Better Than Google” auf Seomoz.org, URL http://www.seomoz.org/blog/the-search-engine-thats-already-better-than-google (Letzter Zugriff 14.04.2008).

weitgehend aus bzw. fokussieren den Nutzen von Social Bookmarks für die Verbesserung automatischer Retrievalverfahren [11].

3. Websuchmaschinen

Systeme, welche auf der Verwendung maschineller oder auch roboterbasierter Verfahren der Dokumentbeschaffung aufsetzen und bezüglich der Inhaltserschließung und der Spezifizierung der Treffermengen auf Methoden des Information Retrieval beruhen, stellen den dominierenden Typus der Suchdienste im Web dar [2]. Abbildung 6 zeigt die wesentlichen Komponenten einer Suchmaschine.

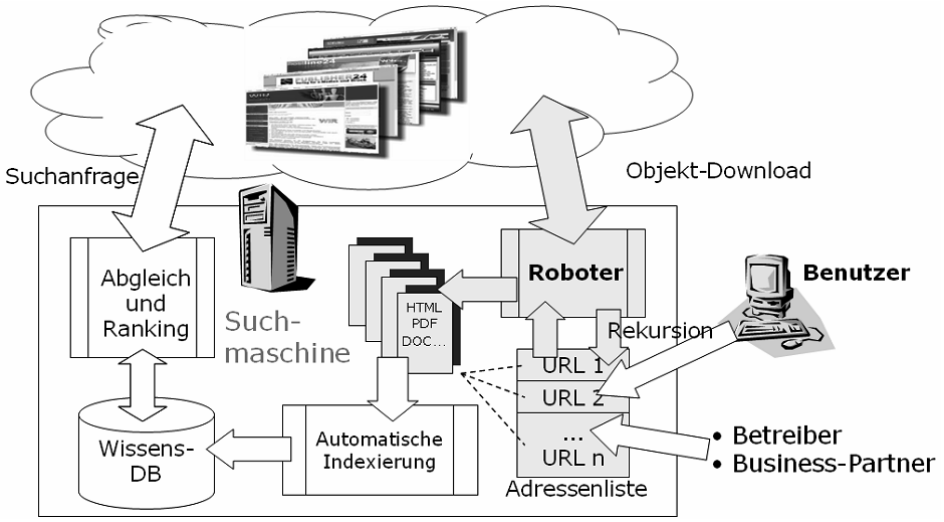

Abbildung 6. Aufbau einer Suchmaschine

Suchmaschinen bestehen im Wesentlichen aus drei Komponenten. Einer Komponente zur Dokumentbeschaffung, einer Komponente zur Inhaltserschließung und Erfassung weiterer struktureller und statistischer Daten sowie einer Komponente, welche die Ergebnismenge und deren Sortierung in Bezug zu den gestellten Suchanfragen determiniert. Die Komponenten werden nachfolgend dargestellt. Dabei ist zu betonen, dass einerseits die methodischen Ansätze, die von Websuchmaschinen genutzt werden, in der Fachwelt zwar bekannt bzw. in der Fachliteratur nachzulesen sind, vgl. u.a. [1], [12], [13], [14], die genaue Funktionsweise aber nicht exakt dargestellt werden kann, da diese von den Suchmaschinenbetreibern geheim gehalten wird. Dennoch ist es möglich, auf der Basis von Fachliteratur, Hinweisen der Suchmaschinenbetreiber[24], Analyse von Patenten[25] und nicht zuletzt mit Hilfe des Erfahrungswissens von Suchmaschinenmarketing- und Suchmaschinenoptimierungsexperten [22] ein vielschichtiges Bild von Suchmaschinen zu gewinnen.

[24] Welche diese z.B. im Web selbst bekannt geben. Vgl. z.B. http://www.ysearchblog.com/, http://googleblog.blogspot.com/ (Letzter Zugriff 24.04.2008).

[25] Kostenlose Patentrecherchen sind z.B. über das Webportal des Europäischen Patentamtes (EPA) möglich. URL http: //ep.espacenet.com/ (Letzter Zugriff 24.04.2008).

3.1. Dokumentbeschaffung

Die Dokumentbeschaffung von Suchmaschinen findet primär über Programme, sogenannte Spider oder Crawler statt, die, ausgehend von einer vorhandenen URL-Liste, rekursiv die Hyperlinks des Web traversieren und die Inhalte von HMTL-Dokumenten und anderen textbasierten Dateiformaten extrahieren. Des Weiteren ist es Websitebetreibern möglich, URLs manuell anzumelden[26]. Wenig thematisiert wird die Tatsache, dass Suchmaschinen zudem teilweise in der Lage sind, Navigationsprofile von Internetnutzern zu erfassen[27]. Gerade im Vergleich zu manuell erstellten Dokumentsammlungen erreichen die automatischen Verfahren der Suchmaschinen eine hohe Abdeckung bezüglich der im Web vorhanden Inhalte. Die letzten veröffentlichten Angaben der Suchmaschinenbetreiber Yahoo und Google aus dem Jahre 2005 geben eine Indexgröße von rund 20 Milliarden Dokumenten an[28].

Ziel von Suchmaschinen ist es, die Inhalte des Web möglichst vollständig aufzuspüren. Bezogen auf die Dokumentbeschaffung mittels Spider- oder Crawlerprogrammen wäre ein vollständiges Auffinden für den Fall denkbar, dass alle Webinhalte über Links miteinander verbunden und frei zugänglich sind. Letzteres ist oft nicht der Fall, ersteres ist ganz sicher nicht erfüllt. Die Inhalte Login-geschützter Seiten oder Ergebnisse formularbasierter Anfragen[29] sind für die linktraversierenden Spider unzugänglich. Des Weiteren kommt eine Untersuchung aus dem Jahr 2000 zu dem Schluss, dass das Web zwar einen Kern stark untereinander verlinkter Seiten aufweist, aber ebenso Bereiche beinhaltet, die nicht miteinander verbunden sind[15]. D.h. Suchmaschinen erreichen zwar eine hohe Abdeckung, zugleich existiert aber ein Deep oder Invisible Web, auf dessen Inhalte aufgrund von Zugangsbeschränkungen durch Anbieter oder aufgrund technischer Restriktionen der Suchmaschinen nicht zugegriffen werden kann. Zur Größe dieses Deep Web gibt es unterschiedliche Schätzungen. So geht ein Whitepaper der Firma Brightplanet aus dem Jahr 2001 davon aus, dass das Deep Web 400 bis 550Mal größer sei als das indexierbare Web und mindestens 550 Milliarden Dokumente umfasse [16]. Eine aktuellere Schätzung kommt für den Wissenschaftsbereich auf eine Größe von zwischen 20 und 100 Milliarden Dokumenten [17].

Aufgrund der hohen Veränderlichkeit des Internet – in dem ständig neue Webseiten und andere Objekte publiziert und vorhandene Dokumente/Daten modifiziert oder entfernt werden – ist es zudem erforderlich, die Indizes der Suchmaschinen fortlaufend zu aktualisieren. D.h. die Dokumentbeschaffung durch Suchmaschinenroboter ist ein zyklischer Prozess, darauf angelegt, die Veränderungen des Dokumentraums Internet möglichst zeitnah zu erfassen. Neben Crawling-Heuristiken, die darauf abzielen, die Inhalte einer Domain möglichst vollständig zu erfassen, und denen, die das Ziel verfolgen, eine möglichst hohe Zahl von Domains zu erfassen, nutzen Suchmaschinen auch weitere Informationen wie Besuchshäufigkeit bzw. Aktualisierungsfrequenz von

[26] Bei Google z.B. unter http://www.google.com/addurl/?hl=de&continue=/addurl (Letzter Zugriff 24.04.2008).

[27] Das ist dann der Fall, wenn die Nutzer sogenannte Suchmaschinentoolbars verwenden und derart konfigurieren, dass Daten der im Browser aufgerufenen Seiten an den jeweiligen Dienst übermittelt werden. Informationen zur Google-Toolbar finden sich unter http://www.google.com/support/toolbar/?hl=de (Letzter Zugriff 24.04.2008).

[28] Vgl. "The size of the World Wide Web", http://www.pandia.com/sew/383-web-size.html (Letzter Zugriff 14.04.2008).

[29] Dabei handelt es sich meist um anbieterspezifische Datenbanken, die Webseiten erst aufgrund konkreter Nutzeraktionen dynamisch generieren.

Webseiten, um das Verhalten von Spiderprogrammen zu optimieren. Websitebetreiber wiederum verfügen über mehrere Optionen, das Verhalten von Suchmaschinenrobotern zu beeinflussen. Zunächst können Websitebetreiber über Meta-Tags[30] Suchmaschinen Informationen zur Indexierung bereitstellen. Tabelle 3 nennt auf Suchmaschinenroboter zielende Metaangaben in HTML-Seiten[31].

Tabelle 3. Meta-Tags für Suchmaschinenroboter

Tag	Bedeutung
index	Indexieren
noindex	Nicht Indexieren
follow	Verweisen folgen
nofollow	Verweisen nicht folgen
noodp	Bei der Ergebnisanzeige keine Beschreibung aus dem Open Directory Webkatalog verwenden
noydir	Bei der Ergebnisanzeige keine Beschreibung aus dem Yahoo Webkatalog verwenden
noarchive	Webseite nicht archivieren

Des Weiteren führte Yahoo 2007 mit dem „robots-nocontent"-Tag eine Möglichkeit ein, auch Textinhalte im sichtbaren Bereich einer Webseite von der Indexierung auszuschließen.

Seit 2005 kommunizieren die Suchdienstebetreiber Google, Yahoo und Microsoft dass Linkattribut „Nofollow". Dessen Gebrauch bewirkt, dass derart gekennzeichnete Links bei der Sortierung nicht mehr berücksichtigt werden. Neben diesen granularen Steuerungsmöglichkeiten auf Ebene der einzelnen Seiten existiert mit dem „Robots exclusion standard" [32] auch eine Konvention, um das Verhalten von Suchmaschinenrobotern auf Domainebene zu spezifizieren. Websitebetreiber können Suchmaschinenrobotern mitteilen, dass ihre Domain bzw. Teilbereiche davon nicht indexiert werden sollen. Die entsprechenden Anweisungen werden in einer Textdatei namens Robots.txt hinterlegt. Abbildung 7 zeigt ein Beispiel für die Website xyz.com, in der der Websitebetreiber für alle Roboter („User-agent: *") spezifiziert, dass die Unterverzeichnisse „Templates" und „CGI" nicht indexiert werden sollen.

Bei den genannten roboterspezifischen (Meta-)Tags und dem „Robots exclusion standard" handelt es sich um Konventionen und keine für Suchmaschinen verbindliche Maßnahmen: D.h. Suchmaschinen können, müssen sich aber nicht an die Vorgaben halten. 2006 einigten sich Google, Yahoo und Microsoft weiterhin auf ein „Standard Sitemap Protokoll" [33] . Sitemaps gestatten es, in Form eines XML-Files, Metainformationen zum letzten Aktualisierungszeitpunkt, zur Aktualisierungsfrequenz und zur Priorität der aufgelisteten URLs einzutragen. Dies erleichtere es

[30] Meta-Tags sind für den Nutzer unsichtbare Metainformationen, die als Text im Quellcode von Webseiten eingetragen werden.

[31] Vgl. auch http://en.wikipedia.org/wiki/Meta_element#The_robots_attribute (Letzter Zugriff 14.04.2008).

[32] Vgl. URL http://www.robotstxt.org/ (Letzter Zugriff 24.04.2008).

[33] Vgl. "Google, Yahoo and Microsoft Agree to Standard Sitemaps Protocol" URL http://www.techcrunch.com/2006/11/15/google-yahoo-and-microsoft-agree-to-standard-sitemaps-protocol/ (Letzter Zugriff 14.04.2008).

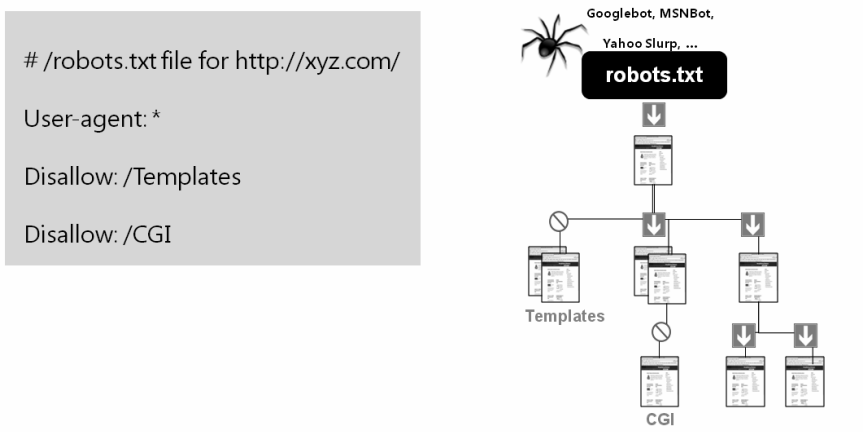

Abbildung 7. Beispiel einer Robots.txt

Suchmaschinen, Webseiten intelligenter zu indexieren[34]. Google bietet des Weiteren mit den „Webmaster-Tools“ Websitebetreibern die Option, umfangreiche Crawling-Informationen zur Indexierung durch Google zu erhalten[35].

Sitemaps und „Webmaster-Tools“ können als eine Weiterentwicklung der Steuerungsmöglichkeiten durch Meta-Tags und des „Robot Exclusion Standards“ betrachtet werden. Alle genannten Möglichkeiten tragen dazu bei, die Ressourcen der Suchmaschinen zur Dokumentbeschaffung effizienter zu nutzen und Probleme wie die Mehrfachindexierung gleicher Inhalte oder niedrige Aktualitätsfrequenzen zu minimieren. Zugleich wird mit Diensten wie den Webmaster-Tools die roboterbasierte Dokumentbeschaffung auch für Websitebetreiber transparenter. Damit ist zu erwarten, dass z.B. technische Problembereiche beim Crawling, die etwa durch nicht verfolgbare Links, dynamische oder dynamisch erzeugte Seiten auftreten, (tendenziell früher) entdeckt und behoben werden.

Zusammenfassend sind also vielfältige Bemühungen erkennbar, die Dokumentbeschaffung von Suchmaschinen zu verbessern. Dabei zeigt die über die Jahre erheblich zunehmende Größe der Suchmaschinenindizes sowohl in Bezug auf die Zahl der erfassten Dokumente als auch in Bezug auf die Zahl der unterstützten Dokumentformate, dass es den Suchmaschinen in zunehmendem Maße gelingt, die Inhalte des „Indexable Web“ durchsuchbar zu machen[36]. Eine erhebliche Leistung, denn nach einer Schätzung aus dem Jahre 2004 werden pro Woche ca. 300 Millionen Webseiten erstellt [18]. Eine aktuelle Studie deutet zudem darauf hin, dass die populären Suchmaschinen Google, Yahoo und MSN einen Großteil von Webseiten innerhalb weniger Tage reindexieren [19]. Suchmaschinen erreichen also eine hohe Abdeckung und sind i.d.R. relativ aktuell. Dennoch ist festzuhalten: Die Größe des (indexierbaren) Web ist nicht bekannt, ebenso ist unbekannt, welcher Anteil durch Suchmaschinen abgedeckt wird [20]. Für den Suchmaschinennutzer ist es deshalb

[34] „In its simplest form, a Sitemap is an XML file that lists URLs for a site along with additional metadata about each URL (when it was last updated, how often it usually changes, and how important it is, relative to other URLs in the site) so that search engines can more intelligently crawl the site.“ URL http://sitemaps.org/ (Letzter Zugriff 14.04.2008).

[35] URL https://www.google.com/webmasters/tools/docs/de/about.html (Letzter Zugriff 14.04.2008).

[36] Vgl. u.a. http://www.boutell.com/newfaq/misc/sizeofweb.html und http://www.worldwidewebsize.com/ (Letzter Zugriff 14.04.2008).

wichtig, sich zu vergegenwärtigen, dass Suchmaschinen zwar große Teilbestände des indexierbaren Web nachweisen, aber Wissensbestände des sogenannten Deep Web, oft umfangreiche Wissensbasen professioneller Anbieter, nur zu einem geringen Teil erfassen (können).

3.2. Erschließung und Spezifikation der Ergebnismenge

Im Information Retrieval bestimmen die Verfahren, die zur Repräsentation der erfassten Wissensobjekte verwendet werden, weitgehend die Optionen der Anfragenbearbeitung bzw. die Möglichkeiten, welche zur Spezifikation und Sortierung der Suchergebnisse zur Verfügung stehen. Deshalb werden nachfolgend beide Bereiche gemeinsam betrachtet.

Den Kern der Inhaltserschließung bei Suchmaschinen stellen zunächst die von den Suchmaschinenrobotern erfassten Inhalte der gefundenen Dokumente dar. Abbildung 8 zeigt auf der linken Seite eine Webseite[37], wie sie durch Webbrowser dargestellt wird, und veranschaulicht auf der rechten Seite mit Hilfe eines „Robot-Simulators[38]“ die Sicht einer Suchmaschine.

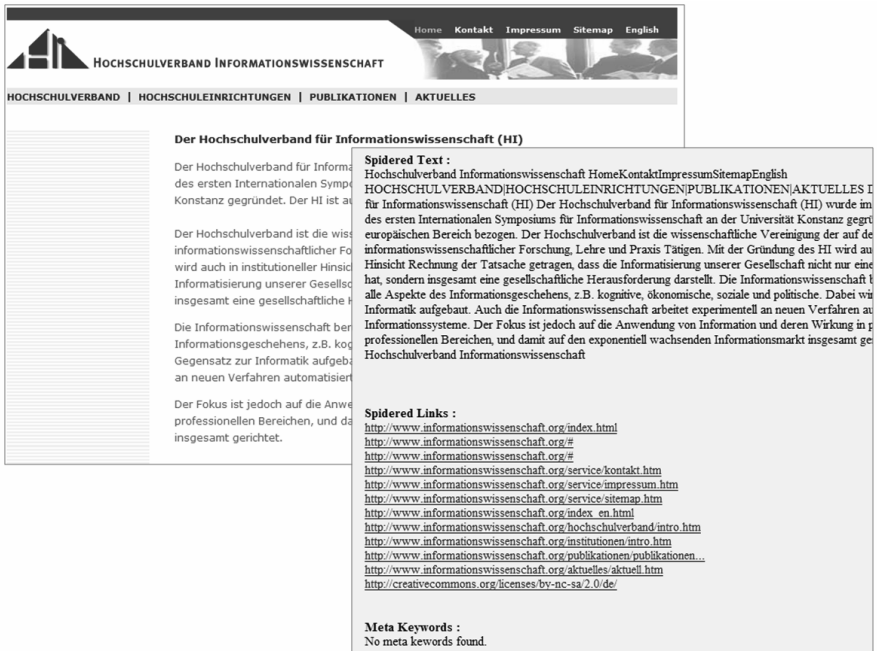

Abbildung 8. Browser vs. Suchmaschinendarstellung

Die Abbildung illustriert, wie Suchmaschinen den Volltext von Webseiten erschließen. Dabei werden exakte Schreibweisen erfasst und Groß- und Kleinschreibung sowie Umlaute meist normalisiert. Morphologische und syntaktische Verfahren der Textanalyse, wie z.B. Grund- und Stammformreduktion, Kompositazerlegung oder die

[37] URL http://www.Informationswissenschaft.org (Letzter Zugriff 24.04.2008).
[38] URL http://www.webconfs.com/search-engine-spider-simulator.php (Letzter Zugriff 24.04.2008).

Erkennung von Mehrwortbegriffen, finden derzeit meist keine Anwendung. Neben den Stichwörtern werden auch

- HTML-Strukturinformationen (HTML-Tags),
- ausgehende Links,
- dokumentinhärente Metainformationen (Meta-Tags),
- weitere formale Elemente (z.B. Dateigröße, Änderungsdatum),
- eingebettete Elemente (z.B. Dateinamen von Bildern)
- und teilweise auch Formatelemente (z.B. Schriftgröße, Farbe)

erfasst.

Diese Art der Volltextindexierung stellt die Grundlage zur Anwendung klassischer termbasierter Abgleichs- und Sortierverfahren dar, die auf der Analyse von Wortvorkommen in Dokumenttext und Metainformation aufsetzen. Die Analyse ausgehender Verweise bildet die Basis linktopologischer Rankingverfahren.

Diese aus den Dokumentinhalten extrahierten Informationen werden mit weiteren Faktoren ergänzt. Google spricht derzeit von über 200 „Signalen“, die beim Ranking berücksichtigt werden[39]. Diese lassen sich im Wesentlichen vier zentralen Bereichen zuordnen:

- On-Page-Faktoren
- On-Site-Faktoren
- Linkfaktoren
- Eigenschaften und Verhalten der Benutzer.

Abbildung 9 stellt die Faktoren in einer grafischen Übersicht dar.

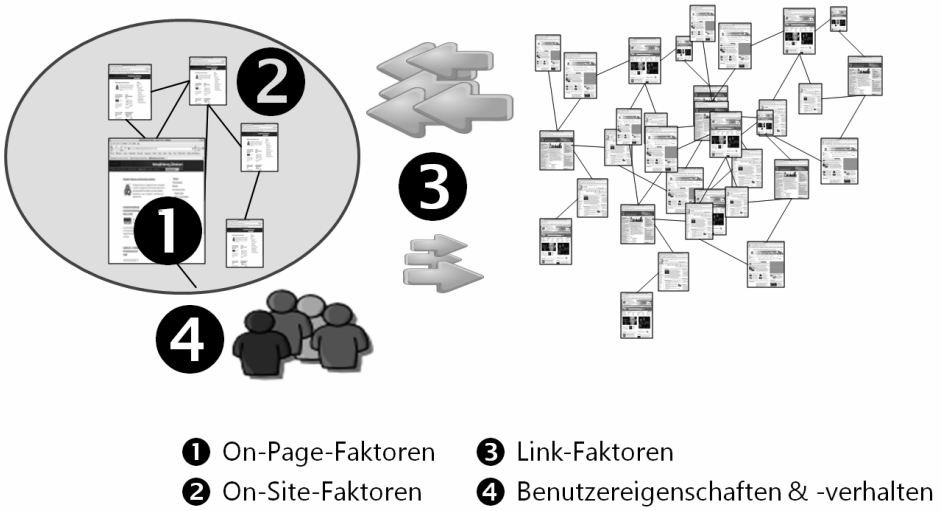

Abbildung 9. Ranking-Faktoren bei Suchmaschinen

[39]Vgl. den Artikel „Google Keeps Tweaking Its Search Engine” in der New York Times vom 03.06.2007, URL http://www.nytimes.com/2007/06/03/business/yourmoney/03google.html (Letzter Zugriff 15.04.2008).

Neben der Relevanzeinstufung der Dokumente an sich ist die Zusammensetzung bzw. die Vielfältigkeit der jeweiligen Ergebnismenge ein wichtiges Kriterium für ihre weitere Spezifikation der Ergebnismenge[40].

3.2.1. On-Page-Faktoren

Die inhaltliche Erschließung auf der Basis dokumentinhärenter Daten wurde bereits dargestellt. Die wesentlichen Rankingfaktoren in diesem Bereich werden u.a. in [1], [2] aufgeführt. Insbesondere von Bedeutung sind bezüglich der Terme

- Häufigkeit, Position (Dichte, Abstand)
- Funktion (URL, HTML-Auszeichnungen: Titel, Überschriften, Linktexte…)
- Format von Termen (Schriftgröße, Farbe).

Dabei gilt, je öfter Anfrageterme in einem Dokument vorkommen, je dichter sie zueinander bzw. je weiter sie am Anfang des Dokuments stehen, umso relevanter wird ein Dokument bewertet. Ebenso werden hervorgehobene Terme oder Terme in spezifischen Feldern höher gewichtet. Weitere eher formale Faktoren, z.B. das Entstehungsdatum oder die Änderungsfrequenz, können beispielsweise bei zeitbasierten Anfragekriterien berücksichtigt werden.
On-Page-Faktoren stellen den Kern jeder inhaltsbasierten Bewertung von Suchmaschinen dar. Sie werden aber für das Ranking als nicht hinreichend erachtet. Dies hat zunächst zwei Gründe. Erstens das Suchverhalten der Nutzer: Internetnutzer stellen überwiegend kurze Suchanfragen, d.h. Anfragen mit nur wenigen Termen, oftmals auch nur sogenannte Einwort-Anfragen [21]. Suchmaschinen weisen zu derartigen Suchanfragen i.d.R. Tausende bzw. Millionen potenziell relevanter Dokumente nach, von denen die Nutzer dann meist nur wenige Treffer sichten. Ein Beispiel für dieses Problem stellt etwa die in Tabelle 2 aufgeführte Anfrage „playstation“ dar. Diese liefert in Google knapp 200 Millionen Ergebnisse. Es ist schwierig diese hohe Anzahl von Dokumenten allein mittels der Analyse dokumentinhärenter Termokurrenzen sinnvoll zu sortieren.

Der zweite Grund, warum On-Page-Kriterien für sich betrachtet als nicht hinreichend erachtet werden, liegt in dem Missbrauchspotenzial durch Websitebetreiber. So lässt sich z.B. die Häufigkeit von Termen in Webdokumenten sehr einfach manipulieren. Etwa indem Textpassagen mittels Farbauszeichnung so formatiert werden, dass sie für Nutzer unsicht-, für die Suchmaschine aber sichtbar sind. Oder indem Metainformationen (Meta-Tags) gezielt mit inhaltlich „falschen“ aber populären Termen angereichert werden[41]. Suchmaschinen verwenden zwar schon seit den 1990er Jahren inhaltsbezogene Filter, etwa bezüglich einer maximal tolerierten

[40] “The sites with the 10 highest scores win the coveted spots on the first search page, unless a final check shows that there is not enough “diversity” in the results. “If you have a lot of different perspectives on one page, often that is more helpful than if the page is dominated by one perspective, ” Mr. Cutts says. “If someone types a product, for example, maybe you want a blog review of it, a manufacturer’s page, a place to buy it or a comparison shopping site.” Ebd.

[41] Letzteres hat dazu geführt, dass die Inhalte des Meta-Tags „Keywords“, in dem Websitebetreiber die jeweiligen Inhalte über Schlagworte beschreiben können, z.B. bei Google nicht mehr berücksichtigt, d.h. beim Ranking ignoriert werden .Vgl. http://googlewebmastercentral.blogspot.com/2007/12/answering-more-popular-picks-meta-tags.html (Letzter Zugriff 15.04.2008).

Wortdichte bzw. der Zahl von Wortwiederholungen, um manipulierte Seiten aus dem Ergebnis auszusortieren bzw. mit einem Rankingmalus zu versehen. Dennoch ist festzuhalten, dass die Anwendung zusätzlicher Rankingfaktoren, welche auch Kriterien außerhalb der Dokumentinhalte berücksichtigen, den Missbrauch bzw. die Manipulation der Suchmaschinen zu Spamzwecken erschwert bzw. erheblich aufwändiger gestaltet [2].

3.2.2. On-Site-Faktoren

Die Analyse globaler Faktoren der jeweiligen Domain, auf der sich die Dokumente befinden, stellt einen weiteren wichtigen Faktor zu Bewertung von Suchergebnissen dar. Die Art der verwendeten On-Site-Faktoren und ihre reale Bedeutung sind aber weitgehend unbekannt. D.h. rankingbezogene Aussagen sind gerade in diesem Bereich hochgradig spekulativ. So gibt es z.B. seit mehreren Jahren Diskussionen zu vermuteten Sandbox- oder „trust rank"-Effekten, die zur Folge haben sollen, dass neuen Websites insbesondere für kompetitive Suchanfragen ein Rankingmalus zugeordnet werde[42]. Inhaltlich untermauern lässt sich diese Annahme u.a. dadurch, dass Google 2005 selbst Domain Name-Registrar wurde und diesen Schritt damit begründete, dass Registrarinformationen dazu genutzt werden sollen, um Suchergebnisse zu verbessern[43]. Denkbar ist u.a., dass neben dem Alter der Domain auch Faktoren wie die Art der Domain, ihre Linkpopularität, die thematische Ausrichtung der Gesamtsite, die Gesamtzahl der indexierten Seiten usw. bereits jetzt oder künftig herangezogen werden [22].

3.2.3. Link-Faktoren

Linktopologische Sortierverfahren beruhen auf der Analyse der Referenzstrukturen im Web. Die Idee ist, aus diesen Strukturen Kriterien zur Bewertung von Webdokumenten abzuleiten. Grundlage ist die These, dass Links nicht zufällig gesetzt werden, sondern ein Qualitätsurteil, d.h. eine Empfehlung aussprechen. Erstmalige Umsetzung fand dieser Ansatz 1998 in der damals neu entstandenen Suchmaschine Google. Das ursprünglich verwendete Pagerank-Verfahren ist in [23] und [41] dokumentiert. Derartige linktopologische Verfahren setzen auf Ansätzen der Zitationsanalyse wissenschaftlicher Arbeiten auf [24]. Zitationsanalysen fußen auf der Annahme, dass sich die Bedeutung wissenschaftlicher Arbeiten durch die Zahl der zitierenden Arbeiten abschätzen lässt. Die Anwendung derartiger Verfahren im Web Information Retrieval lässt sich u.a. dadurch begründen, dass die Grundidee plausibel und einfach klingt und Links auch technisch relativ einfach extrahiert und analysiert werden können [25]. Das bekannteste linktopologische Verfahren, das von Google verwendete Pagerank-Verfahren, ermittelt die Wichtigkeit einzelner Dokumente durch die Analyse der Verweisstrukturen aller indexierten Webseiten. Dabei gilt: Je größer die Zahl eingehender Links auf eine Seite, umso höher der Pagerank. Pagerank ist ein themenunabhängiges Qualitätsmaß und weist in seiner ursprünglichen Form jedem erfassten Objekt einen „Wichtigkeitsfaktor" zu. Neben der Anzahl der Links fließt auch deren Gewicht in die Berechnung mit ein. Dieses bestimmt sich durch den Pagerank

[42] URL http://searchengineland.com/070206-101047.php (Letzter Zugriff 18.04.2008).
[43] URL http://www.news.com/Google-gets-rights-as-Web-site-registrar/2100-1032_3-5559164.html?tag=item (Letzter Zugriff 18.04.2008).

der Webseite, von der der jeweilige Link ausgeht, und wird gleichmäßig zwischen allen ausgehenden Links dieser Seite aufgeteilt. Abbildung 10 aus [1] veranschaulicht diesen Zusammenhang und zeigt z.B., wie etwa eine Internetseite mit einem (fiktiven) Pagerank von 100 den zwei von ihr ausgehenden Links jeweils ein Pagerankgewicht von 50 vererbt.

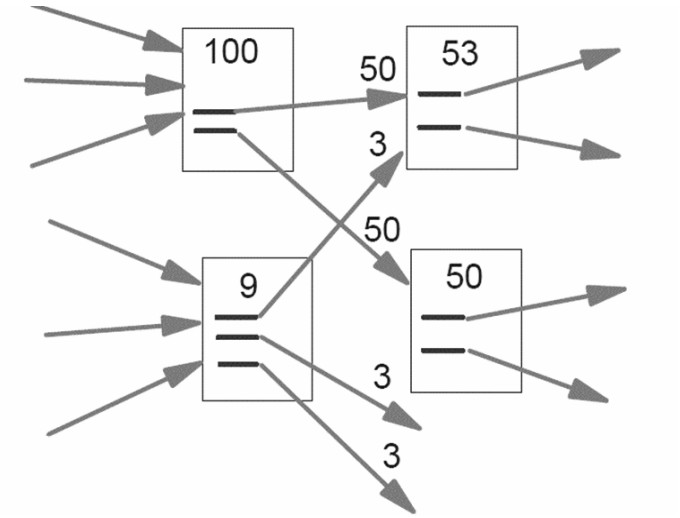

Abbildung 10. Pagerank – Google: ursprüngliches Modell von 1998, Quelle [1]

Neben dem Pagerank-Verfahren existieren weitere linktopologische Algorithmen, die z.B. in [1] und [25] dargestellt sind. Das von Kleinberg entwickelte „Hyperlink-Induced Topic Search“ (HITS)-Verfahren [26] berücksichtigt im Unterschied zu Pagerank auch den Kontext von Suchanfragen. Das HITS-Verfahren floss in die Entwicklung der Suchmaschine Teoma mit ein[44]. Linkfaktoren stellen derzeit ein zentrales Kriterium dar, um Suchergebnisse zu bewerten. Dabei ist davon auszugehen, dass die vor rund 10 Jahren dokumentierten Algorithmen mittlerweile vielfältig modifiziert und weiterentwickelt wurden, nicht nur im wissenschaftlichen Bereich, sondern gerade auch im realen Einsatz bei Google und Co.

Einen weiteren Aspekt linktopologischer Verfahren stellt die Analyse des Verweistexts ausgehender Verweise dar. Wird dieser, wie etwa in [23] beschrieben, dem Inhalt der Objekte zugeschlagen, auf die verwiesen wird, so ist es möglich, diese Objekte auch für Terme nachzuweisen, die gar nicht in ihnen vorkommen. Das ermöglicht u.a. den Nachweis nicht indexierter Dokumente oder auch zunächst nicht-indexierbarer Dokumenttypen (z.B. Bilder), kann aber auch zu inhaltlich verfälschten Suchergebnissen führen[45].

[44] Welche mittlerweile in die Suchmaschine Ask.com integriert wurde; URL http://about.ask.com/en/docs/about/webmasters.shtml (Letzter Zugriff 16.04.2008).

[45] Das Missbrauchspotenzial dieser Linktextanalyse wurde unter dem Schlagwort „Google Bombing“ bekannt. „Google Bombing“ lässt sich als vielfaches Setzen von Links mit einem gemeinsamen Verweistext beschreiben. Ziel ist es, Webseiten bestimmter Organisationen oder Personen mit dem Verweistext in Verbindung zu setzen, oft mit einer diffamierenden Motivation. Vgl. URL http://googlewebmastercentral.blogspot.com/2007/01/quick-word-about-googlebombs.html (Letzter Zugriff 16.04.2008).

Einerseits werden Linkfaktoren als zentraler Erfolgsfaktor bei der Sortierung von Suchergebnissen betrachtet[46], andererseits qualitätssteigernde Effekte aber auch stark in Zweifel gezogen [27]. Letzteres beruht auf der Beobachtung, dass Systeme, welche linktopologische Verfahren nutzen, bei Tests keine besseren Leistungen zeigen als Systeme, die derartige Verfahren nicht verwenden [27]. Hier ist allerdings zu konstatieren, dass diese Tests in Umgebungen durchgeführt wurden, welche webspezifische Bedingungen, z.B. in Bezug auf Spamseiten, nicht vollständig widerspiegeln (können). So lässt sich argumentieren, dass Linkfaktoren gerade bei kurzen Anfragen, die auf semantischer oder pragmatischer Ebene oft vieldeutig sind, eine Verbesserung der Qualität bewirken, weil sie dazu führen, populäre Seiten höher zu ranken. Des Weiteren erhöhen diese Verfahren, im Vergleich zu Rankingverfahren, die ausschließlich auf On-Page-Faktoren beruhen, den Aufwand für eine erfolgreiche Manipulation von Suchmaschinenrankings in erheblichem Maße[47], so dass sie gezielt in Richtung Spamreduzierung wirken. Insofern lassen sich für die Anwendung von Linkfaktoren starke Argumente finden.

Allerdings ist sehr kritisch zu hinterfragen, inwieweit die grundsätzliche Annahme, auf der Linkfaktoren beruhen, nämlich dass das Setzen eines Links ein Qualitätsmerkmal darstellt, derzeit tatsächlich (noch) valide ist. So lässt sich beobachten, dass Webseiten, welche bereits eine hohe Zahl eingehender Links aufweisen, eine überproportional höhere Wahrscheinlichkeit besitzen, neue Links zu erwerben, als Webseiten, welche eine geringe Zahl eingehender Links aufweisen [28]. D.h. linktopologische Verfahren benachteiligen unpopuläre Seiten in einem überproportionalen Ausmaß. Dies betrifft insbesondere neue Seiten (Inhalte), die noch wenig Zeit hatten, „Linkpopularität“ aufzubauen. Damit wirken diese Verfahren in Richtung Verstetigung der Suchergebnisse. Noch weitergehend ist zu konstatieren, dass die Anwendung linktopologischer Verfahren durch die Suchmaschinen wiederum einen Rückkoppelungseffekt auf die Linkstruktur des Web selbst nach sich zieht. Zunächst führt die Bevorzugung populärer Inhalte in den Suchergebnissen dazu, dass sich deren Sichtbarkeit erhöht, was die soeben darstellten Verstetigungseffekte noch weiter verstärkt. Darüber hinaus beeinflusst das Wissen um die Verwendung von Linkfaktoren durch Suchmaschinen bei Websitebetreibern die Motivation bezüglich des Setzens von Links. Dadurch, dass eine hohe Zahl eingehender Links sich positiv auf die Sichtbarkeit des eigenen Webangebots auswirkt, bestehen starke Anreize, Links aktiv „einzuwerben“. Das ist ein qualitativer Effekt, der die Linkstruktur des Web insgesamt beeinflusst. Dass diese mittlerweile auch in hohem Maße direkt von Marketinginteressen (mit)bestimmt wird, zeigt sich u.a. daran, dass Links mittlerweile auch ein kommerzielles Gut geworden sind, das oft auch käuflich erworben werden kann.

Die Auswirkungen dieser Effekte auf die Qualität von Linkfaktoren und das Internet insgesamt sind weitgehend unbekannt. Zumindest aus Sicht des Suchmaschinenbetreibers Google stellen kommerziell erworbene Links, die das Ziel verfolgen, die Linkpopularität zu erhöhen, ein Problem dar[48].

---

[46] Vgl. URL http://www.google.com/technology/ (Letzter Zugriff 16.04.2008).

[47] Denn es ist nicht mehr hinreichend, die Inhalte einzelner Seiten zu verändern, vielmehr muss die Linkstruktur einer Vielzahl von Seiten manipuliert werden.

[48] Vgl. URL http://www.google.de/support/webmasters/bin/answer.py?answer=66736&topic=8524 (Letzter Zugriff 16.04.2008).

3.2.4. Eigenschaften und Verhalten der Benutzer

Die Eigenschaften und das Verhalten der Nutzer beeinflussen in vielfältiger Weise die Spezifikation und Reihenfolge der ausgelieferten Suchergebnisse. Zunächst sind dabei individuelle Eigenschaften von Bedeutung. Derartige Eigenschaften lassen sich zunächst über die IP-Adresse des verwendeten Rechners, durch im Browser (z.B. über Cookies) oder auf dem Server gespeicherte Informationen (Sessions) sowie durch eine Authentifizierung über einen Login ermitteln. Darüber hinaus ist es Suchmaschinen aber auch möglich, aggregierte Interaktionsdaten einer Vielzahl von Nutzern zu berücksichtigen.

Auf individueller Ebene kann zunächst über die jeweilige IP-Adresse eine geografische Zuordnung vorgenommen werden. Damit lässt sich beispielsweise die vom Nutzer vermutlich verwendete Sprache ermitteln und bei der Ergebnisausgabe berücksichtigen [49]. Des Weiteren ist es möglich, bei Suchanfragen mit einem geografischen Bezug – etwa der Anfrage „Umzugservices" – Anbieter, die aufgrund ihrer geografischen Zuordnung „nahe" liegen, zu bevorzugen[50]. Außerdem erleichtert eine geografische Zuordnung auch die Umsetzung rechtlicher Regelungen bezüglich der Auslieferung von Inhalten (Zensur)[51].

Neben der Erfassung der zugreifbaren Daten der jeweils verwendeten Rechner können auch Nutzerprofile verwendet werden, um die Suchergebnisse anzupassen. Hier lässt sich zwischen einer expliziten und einer impliziten Erfassung derartiger Daten differenzieren. Einerseits können Nutzer explizit Präferenzen hinsichtlich der Eingrenzung von Suchergebnissen (z.B. Sprachraum) oder der Darstellung von Ergebnissen (Trefferanzahl) usw. angeben, andererseits ist es möglich, das Verhalten der Nutzer implizit zu analysieren und auf dieser Basis Suchergebnisse zu modifizieren.

Das bekannteste Beispiel einer derartigen Personalisierung stellt derzeit das Webprotokoll von Google dar[52]. Es beruht auf der Aufzeichnung des Such- und Surfverhaltens von authentifizierten Google-Nutzern[53]. Im Webprotokoll werden Suchanfragen und selektierte Ergebnisse festgehalten sowie das Surfverhalten im Web über die Google-Toolbar – sofern vorhanden und aktiviert – erfasst. Diese Daten werden von Google genutzt, um Suchergebnisse zu personalisieren, d.h. einen Teil der Ergebnismenge neu zu sortieren. Abbildung 11 zeigt die Benutzerschnittstelle des Webprotokolls.

---

[49] So dass etwa ein deutschsprachiger Nutzer keine spanischsprachigen Ergebnisse zur Anfrage „playstation" ausgeliefert bekommt.

[50] Dieser Punkt ist allerdings deutlich schwieriger umzusetzen, da es wenig sinnvoll ist, aus der IP-Adresse eines Informationsanbieters dessen geografische Lage abzuleiten. Die geografische Zuordnung muss vielmehr mit Hilfe der in 4.2.1. – 4.2.3. geschilderten Verfahren rückerschlossen werden. Zwar können Informationsanbieter ihre geografische Lage auch mit Hilfe von sogenannten Geo-Tags kennzeichnen, vgl. URL http://www3.tools.ietf.org/html/draft-daviel-html-geo-tag-08 (Letzter Zugriff 16.04.2008), ob und inwieweit diese von Suchmaschinen derzeit interpretiert werden, ist aber unklar.

[51] Google wurde und wird bezüglich der Kooperation mit der staatlichen Zensur in China heftig kritisiert, vgl. u.a. http://www.spiegel.de/netzwelt/web/0, 1518, 397285, 00.html (Letzter Zugriff 16.04.2008).

[52] URL www.google.com/history/?hl=de (Letzter Zugriff 17.04.2008).

[53] D.h. Nutzer, die über ein Google-Konto verfügen, über das Web bzw. die Google-Toolbar in Google angemeldet sind und das Webprotokoll aktiviert haben.

Abbildung 11. Google-Webprotokoll

Mit dem Webprotokoll stellt Google Funktionen zur Verfügung, die weit über die bislang bei Suchmaschinen verfügbare Suchfunktionalität hinausreichen[54]. So ist es möglich, vorherige Suchvorgänge retrospektiv nachzuvollziehen und die gespeicherten Inhalte im Volltext zu durchsuchen. Außerdem bilden die erfassten Nutzungsdaten die Grundlage für neue Empfehlungsdienste. Derartig weitreichende Personalisierungsoptionen werden einerseits als wichtiger Schlüssel zur Verbesserung der Internetsuche betrachtet[55], auf der anderen Seite werden datenschutzrechtliche Aspekte thematisiert. Nach wie vor bleibt unklar, ob und inwieweit sich die implizite Erfassung von Nutzerdaten über Dienste wie Googles Webprotokoll dauerhaft durchsetzen oder aufgrund der Gefahr der Preisgabe sensibler Daten nicht angenommen wird [30].

Neben der expliziten und impliziten Erfassung der Präferenzen der Nutzer auf individueller Ebene wird der aggregierten Analyse des Verhaltens einer Vielzahl bzw. aller Nutzer hohes Potenzial für die Ergebnisspezifikation und Sortierung von Suchergebnissen zugeschrieben. So beruhen z.B. die oben genannten Empfehlungsdienste Googles auf der Analyse des Such- und Surfverhaltens vieler Google-Nutzer. Hier werden nach dem Prinzip des kollaborativen Filterns [12] Interaktionsdaten aggregiert und daraus Empfehlungen abgeleitet[56]. Historisch betrachtet lassen sich Ansätze der aggregierten Auswertung von Interaktionsdaten zunächst der inzwischen von Ask.com aufgekauften Suchmaschine Directhit zuordnen. Directhit verwendete erstmals die Anzahl der von den Nutzern getätigten Klicks auf

[54] Für eine detaillierte Darstellung und Diskussion des Google-Suchprotokoll vgl. „Google Search History expands, becomes web History, URL searchengineland.com/070419-181618.php (Letzter Zugriff 17.04.2008).

[55] Zitat aus dem Webartikel "Just Behave: Google's Marissa Mayer on Personalized Search" vom 23.02.2007: "The actual implementation of personalized search is that as many as two pages of content, that are personalized to you, could be lifted onto the first page and I believe they never displace the first result, because that's a level of relevance that we feel comfortable with. So right now, at least eight of the results on your first page will be generic, vanilla Google results for that query and only up to two of them will be results from the personalized algorithm. I think the other thing to remember is, even when personalization happens and lifts those two results onto the page, for most users it happens one out of every five times." URL http: //searchengineland.com/070223-090000.php (Letzter Zugriff 17.04.2008).

[56] Google offers „Queryless Search"& personalized recommendations, URL http: //searchengineland.com/070418-153238.php (Letzter Zugriff 17.04.2008).

Suchergebnisse als Rankingkriterium. Demnach wird Suchergebnissen, die häufiger selektiert werden, ein Rankingbonus zugeteilt. Problematisch bezüglich der Klickhäufigkeit ist zunächst das Manipulationspotential durch Nutzer und automatische Programme. Zum anderen ist die Verwendung von Klickdaten als Rankingkriterium umso weniger sinnvoll, je polysemantischer bzw. inhaltlich vieldeutiger Suchanfragen ausfallen (z.B. Java). In diesem Bereich sind Entwicklungsansätze zu sehen, die dahin zielen, durch den Aufbau von Communities mit Teilnehmern ähnlicher Interessen bzw. durch die Treffermengenbeschränkung auf inhaltlich relativ homogene Themengebiete zu disambiguieren[57] [31]. Über Toolbars oder andere Webanalyse-Tools ist es vielen Suchdiensten darüber hinaus möglich, auch das globale Navigations- bzw. Browsingverhalten einer Vielzahl von Nutzern zu erfassen. Insbesondere Google ist durch die hohe Verbreitung seiner Toolbar und durch weitere Dienste wie dem kostenlosen Analytics[58] bestens gerüstet, derartige globale Strukturmuster des Internet für die Ergebnissortierung anzuwenden. Inwieweit diese Daten derzeit genutzt werden, bleibt weitgehend spekulativ.

Nicht spekulativ sind hingegen die seit wenigen Jahren existierenden sogenannten „Custom Search Engine-Dienste", die es Nutzern auf individueller oder kollaborativer Ebene ermöglichen, eigene Suchmaschinen zu definieren. Im Prinzip der bereits oben erwähnten Suchraumeinschränkung ähnlich, sind diese Dienste wesentlich mächtiger und gestatten es auf der Basis von auf der jeweiligen Suchmaschine aufsetzenden, selbst spezifizierten Dokumentraumsubsets, eigene Suchdienste zu kreieren und diese anderen Nutzern zur Verfügung zu stellen. Beispiele für solche spezifizierbaren Suchmaschinen stellen „MSN-Suchmakros" [59] und „Google Custom Search Engines" dar [60] . Mit „MSN-Suchmakros" können bis zu 30 Websites als zu durchsuchendes Subset bzw. nutzerspezifizierte Suchmaschine definiert werden. Im erweiterten Modus können alle Anfrageparameter der MSN-Suche zur Definition eines Suchmakros verwendet werden. Mit „Custom Search Engines" können bis zu 5000 zu durchsuchende oder im Ranking zu bevorzugende Domains, Pages und Verzeichnisse festgelegt werden. Ergänzend können Terme zur thematischen Ausrichtung definiert werden.

Obwohl es derzeit keine direkte Verbindung bzw. Schnittstelle zwischen den „Custom Search Engine-Diensten" und den jeweiligen Suchportalen bzw. Standardzugängen bei Google und Microsoft gibt, können diese Dienste als Umsetzung von Personalisierungs- bzw. Social Search-Ansätzen begriffen werden. Zum einen gestatten sie auf der Nutzerseite weitreichende Personalisierungsoptionen, zum anderen wird durch die Nutzeraktivitäten eine Vielzahl von Daten generiert, die von den Suchmaschinen zur Verbesserung der Erschließung und der Spezifikation von Ergebnismengen verwendet werden können. So sprechen die Nutzer bzw. die Gestalter von „Custom Search Engine-Diensten" hinsichtlich der spezifizierten URLs zunächst ein positives Qualitätsurteil aus. Weitergehend wird auch eine inhaltliche Zuordnung zum Thema der jeweiligen Sub-Suchmaschine vorgenommen. An dieser Stelle soll nicht tiefer auf ähnlich gelagerte Dienste und Optionen wie "Suscribed Links"[61] oder

[57] Vgl. www.eurekster.com/aboutswickis/technology (Letzter Zugriff 17.04.2008).

[58] Ein Website-Tracking-Tool, mit dem Websitebetreiber Besucherzahlen und -verhalten erfassen können; URL http://www.google.de/analytics/de-DE/ (Letzter Zugriff 18.04.2008)

[59] URL http://gallery.live.com/results.aspx?bt=13&pl=4 (Letzter Zugriff 17.04.2008).

[60] URL http://www.google.com/coop/cse/ (Letzter Zugriff 17.04.2008).

[61] URL www.google.com/coop/subscribedlinks/ (Letzter Zugriff 17.04.2008).

„Bookmarks“[62] des Marktführers Google eingegangen werden. Deutlich wird aber, und das ist der entscheidende Punkt, dass Suchmaschinen – eingangs als algorithmenbasierte automatische Systeme intellektuell und manuell erstellten Dokumentsammlungen dichotom gegenübergestellt – zunehmend Dienste und Optionen bereitstellen, in denen Nutzer ihr Wissen einbringen und Dokumente auf unterschiedliche Arten inhaltlich und qualitativ kennzeichnen können. Suchmaschinen ergänzen ihre automatischen Verfahren verstärkt mit editorialen Komponenten, in denen Nutzer intellektuellen Input zur Verbesserung algorithmenbasierter Suchverfahren bereitstellen. Die Idee bzw. der Ansatz, menschliches Wissen mit maschinellen Algorithmen gewinnbringend zu kombinieren, kennzeichnet damit den derzeit wichtigsten Entwicklungstrend bei Suchmaschinen und lässt für die nahe und mittlere Zukunft erhebliche Verbesserungspotenziale für das Web Information Retrieval erhoffen.

Unmittelbaren und wohl prägnantesten Ausdruck findet dieser Trend derzeit in Diensten wie Wikia Search[63]. Wikia Search, eine Initiative des Wikipedia-Gründers Jimmy Wales, setzt auf frei zugänglichen Suchmaschinentechnologien auf und intendiert, durch menschliche Urteile über die Qualität von maschinellen Suchergebnissen ein besseres Ranking zu erreichen. Die Suche im Web soll damit revolutioniert werden.

Obwohl Wikia Search ein sehr ambitioniertes Projekt ist, wurde seine Qualität nach dem Start im Januar 2008 überwiegend negativ beurteilt[64]. Dennoch gibt dieser Suchdienst einen Ausblick darauf, wie weitreichend maschinelle Verfahren und editoriale Komponenten bzw. der direkte Input von Nutzern künftig miteinander verwoben werden können. Grundlage der Wikia Search bilden Suchmaschinenergebnisse, die auf Basis der in diesem Kapitel bereits skizzierten Algorithmen generiert werden. Nutzer können in einer sogenannten „Whitelist“, eine Liste, in der Webseiten als Startpunkte für die Spiderprogramme eingetragen sind, Ergänzungen vornehmen. Des Weiteren ist es möglich, mit Hilfe eines integrierten Wikis sogenannte „Mini-Artikel“ zu Suchanfragen zu verfassen, die zusätzliche Informationen zu Suchanfragen liefern und etwa mehrdeutige Begriffe („Java“) disambiguieren. Schließlich ist in Wikia Search ein soziales Netzwerk integriert. Melden sich Nutzer in diesem sozialen Netzwerk an, können sie nicht nur soziale Kontakte mit anderen Teilnehmern knüpfen, sondern u.a. auch ein Interessenprofil definieren. Entspricht eine Suchanfrage eines Wiki-Nutzers derartigen Profileinträgen, so werden neben den Dokumenttreffern auch die entsprechenden Teilnehmer des Wikia-Netzwerk zurückgeliefert (vergleiche auch die obenstehende Abbildung). Auf diese Weise führt Wikia Search die bereits in Social Tagging Communities implizit vorhandenen Möglichkeiten der Expertensuche weiter und setzt diese direkt um. Aufsetzend auf den bis hier hin beschriebenen Komponenten ist des Weiteren geplant, Nutzern eine unmittelbare Möglichkeit zur Bewertung der Güte von Suchergebnissen zur Verfügung zu stellen.

[62] URL www.google.com/bookmarks/ (Letzter Zugriff 17.04.2008).
[63] URL http://alpha.search.wikia.com/ (Letzter Zugriff 17.04.2008).
[64] Vgl. “Search Wikia: Not Even A Remote Threat To Google”, URL http: //searchengineland.com/080107-131756.php (Letzter Zugriff 18.04.2008).

3.3. Zusammenfassung Websuchmaschinen

Zusammenfassend bleibt festzuhalten, dass sich hinsichtlich der Arbeits- und Funktionsweise von Suchmaschinen ein komplexes Bild ergibt. Ab Mitte der 1990er Jahre in Form einfacher Verfahren gestartet, die zunächst (Teile der) Volltexte der erfassten Webseiten invertierten und mit Hilfe klassischer Retrievaltechniken die auf Termokurrenzen aufsetzenden Ergebnisse sortierten, entwickelten sie sich kontinuierlich weiter bis zu den heute deutlich komplexer arbeitenden und leistungsfähigeren Systemen. Neben der fortlaufenden Optimierung bestehender Verfahren sind insbesondere ab Ende der 1990er Jahre die Anwendung von Linkfaktoren sowie die zunehmende Nutzung nutzergenerierter oder nutzerbezogener Daten als wichtige Entwicklungsschritte zu kennzeichnen [32]. Anhand neuer Google-Dienste und der Wikia Search lassen sich weitergehende Entwicklungstendenzen verdeutlichen, die aufzeigen, dass auch bei algorithmisch arbeitenden Websuchmaschinen in zunehmendem Maße das Wissen der und das Wissen über die Nutzer zur Verbesserung des Web Information Retrieval genutzt wird. Mit der Wikia Search kann bereits derzeit eine Entwicklungslinie ausgemacht werden, in der intellektuelles und automatisches Retrieval miteinander verschmelzen.

Aus der Sicht von Recherchierenden eignen sich Websuchmaschinen durch ihre hohe Abdeckung und die Volltextinvertierung im Vergleich zu manuell erstellten Dokumentsammlungen insbesondere für spezifische Suchanfragen bzw. Informationsbedürfnisse. Aufgrund der häufig sehr hohen Trefferzahlen ist es i.d.R. lohnenswert, Anfragen möglichst spezifisch zu formulieren und neben geeigneten Suchbegriffen dazu auch die von der jeweiligen Maschine bereitgestellten Operatoren und Limits zu nutzen[65]. Nach wie vor problematisch sind der weitgehend fehlende Kontext in den Suchmaschinenergebnisseiten, die nur rudimentär vorhandenen Browsingoptionen („Ähnliche Seiten“) sowie die nach wie vor existierende Spamproblematik.

4. Spezialsuchdienste

Websuchmaschinen indexieren zwar z.T. viele Milliarden Dokumente, erfassen damit aber derzeit nur einen Teil des Internet. Dabei sind die erfassten Ressourcen inhaltlich, strukturell und qualitativ sehr heterogen, so dass trotz der hohen Abdeckung und der oben dargestellten aufwändigen Sortiermechanismen eine große Unsicherheit bezüglich der Vollständigkeit und Qualität der Ergebnisse besteht. Hinsichtlich der Repräsentation der indexierten Objekte ist festzuhalten, dass diese meist nur einen geringen, je nach Dokumenttyp auch unterschiedlichen Strukturierungsgrad aufweisen und dass, sofern vorhanden, Metadaten nicht a priori als verlässlich einzustufen sind. Nicht zuletzt aus diesem Grund weisen Suchmaschinen nur rudimentäre Optionen zur Suchraumbegrenzung auf: meist Dateiformat, Domain, Datum [33], Sprachraum, Region. D.h. es bestehen erhebliche Defizite bzw. nur geringe Möglichkeiten zur Durchführung strukturierter Anfrageformulierungen.

[65] Vgl. URLs: http://www.google.de/support/bin/topic.py?topic=352, http://www.weboptimierung-griesbaum.de/wissen/google-suche--11-suchtipps-fur-die-internet-suche-mit-google.html (Letzter Zugriff 21.04.2008).

Speziell auf einen bestimmten Gegenstandsbereich fokussierte oder dokumenttypbezogene Spezialsuchdienste sind in ihrer Domäne wesentlich mächtiger, da sie in der Lage sind, spezifische, auf ihren jeweiligen Kontext bezogene Funktionalitäten bereitzustellen. Diese weisen teilweise weit über die bislang in den Abschnitten 2-4 beschriebenen Methoden hinaus und versuchen damit, die Schwächen von Universalsuchdiensten zu kompensieren bzw. eröffnen zusätzliche Suchoptionen. Wichtige Spezialsuchdienste und -Ansätze stellen u.a.

- News & Blogsuchmaschinen
- Wissenschaftssuchmaschinen
- Online-Datenbanken
- (Multi-)Mediasuchdienste
- Einsatz von Visualisierungskonzepten

dar. Spezialsuchdienste und Visualisierungsansätze werden in diesem Buch in eigenen Kapiteln dargestellt, auf die der Leser an dieser Stelle verwiesen sei. Dennoch ist hier festzuhalten, dass aus Nutzerperspektive Spezialsuchdienste oftmals wesentlich besser geeignet sind, Informationsbedürfnisse zu befriedigen, als Universalsuchdienste.

Dies ergibt sich zunächst aus dem jeweilig eingeschränkten Objekt- bzw. Domänenbezug. Dieser gestattet zugleich einen höheren Standardisierungs- und Strukturierungsgrad der Objekte der jeweiligen Domäne. Dadurch ist es wiederum möglich, eine deutlich reichhaltigere und (einheitlich) strukturiertere Erschließung vorzunehmen und/oder für Suchmaschinen nicht zugreifbare Wissensbestände im Internet erst zugreifbar zu machen und somit die Suchoptionen von Recherchierenden in hohem Maße zu erhöhen und zu verbessern. Dieses Potenzial wird aber auf Nutzerseite oft nicht realisiert. Zunächst beschränkt sich eine Vielzahl der Internetnutzer von vorneherein auf die großen Standardsuchmasken der populären Websuchdienste Google & Co [39]. Spezialsuchdienste haben so nur eine geringe Chance, wahrgenommen zu werden. Das gilt selbst dann, wenn derartige Spezialsuchdienste, wie es bei den meisten genannten Beispielen der Fall ist, in die großen Suchportale von Google, Yahoo, MSN eingebunden sind. Dieses Problem wurde vor einigen Jahren mit dem Begriff „Tab Blindness“ beschrieben. „Tab Blindness“ steht als Begriff dafür, dass Suchdienstenutzer in der überwiegenden Zahl der Fälle nicht in der Lage sind, Optionen zur Suchraumeingrenzung wahrzunehmen oder zu verstehen [40]. Dieser Mangel oder Unwille an Anwendungskompetenz bedeutet letztlich, dass es nicht genügt, dem Nutzer für jeweils unterschiedliche Kontexte effektive Suchwerkzeuge bereitzustellen, sondern dass es darüber hinaus notwendig bzw. sinnvoll ist, ihn auch bei der Auswahl der jeweils geeignetsten Suchdienste zu unterstützen. Eine Forderung, der, wie im Folgenden zu sehen ist, Google und Co. schrittweise besser nachzukommen bzw. gerecht zu werden suchen.

5. Metasuchdienste & Metasuchmaschinen

Der Begriff Metasuchdienste umfasst jede Art von Suchdiensten, die auf andere Suchdienste verweisen oder diese nutzen. In ihrer einfachsten Form entsprechen

Metasuchdienste den in Kap. 2 angesprochenen Verzeichnissen und Auswahlhilfen von Suchdiensten. Des Weiteren können auch Suchdienste wie beispielsweise Sputtr.com, die „All-In-One“-Eingabemasken zu anderen Suchdiensten anbieten, als Metasuchdienste verstanden werden.
Der Begriff Metasuchmaschinen ist enger gefasst. Metasuchmaschinen verfügen über keinen eigenen Index, sondern leiten Anfragen an andere Suchdienste weiter und führen die Treffer in einer Trefferliste zusammen. Dabei werden Duplikate i.d.R. eliminiert und eine fusionierte Relevanzbewertung durchgeführt. Damit ist die Qualität der Ergebnisse direkt abhängig von der Qualität der zugrunde liegenden Suchdienste. Abbildung 12 illustriert die Funktionsweise von Metasuchmaschinen.

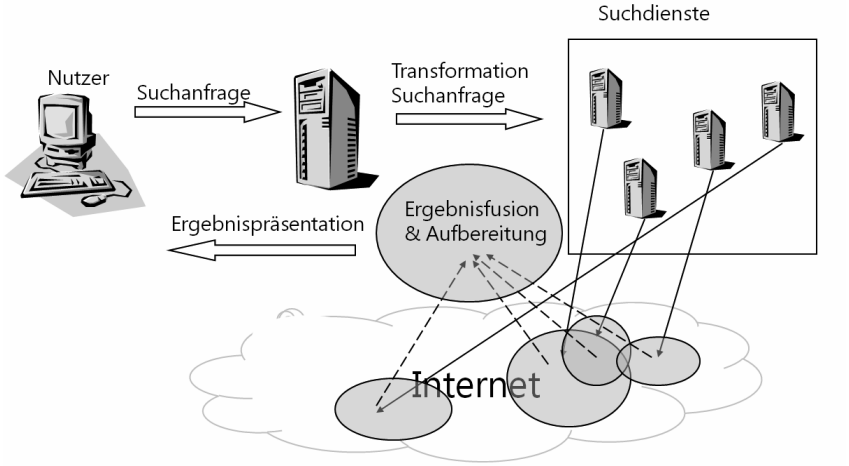

Abbildung 12. Funktionsweise von Metasuchmaschinen

Metasuchmaschinen wird durch die parallele Abfrage mehrerer anderer Suchdienste eine vielschichtigere Ergebnismenge und eine höhere Abdeckung zugeschrieben[66]. Da jedoch die abgefragten Suchdienste meist nur eine begrenzte Zahl ihrer jeweiligen Ergebnisse an Metasuchdienste weiterleiten, zeigt sich diese theoretisch höhere Abdeckung in der Praxis oft nicht. Nutzer sollten deshalb Metasuchdienste vor allem dann verwenden, wenn die einzelnen abgefragten Suchdienste nur jeweils wenige Treffer liefern. Ein weiterer konzeptioneller Nachteil ist, dass sich spezifische Suchoptionen der abgefragten Dienste nur teilweise nutzen lassen, da Metasuchmaschinen meist nur den kleinsten gemeinsamen Nenner der Suchoptionen der abgefragten Suchdienste anbieten.

Metasuchdienste und Metasuchmaschinen sind insbesondere deshalb für Recherchierende interessant, weil sie oftmals innovative Technologien, z.B. bezüglich der Suchdiensteauswahl, Suchanfragenformulierung, Ergebnissortierung, Ergebnisdarstellung oder auch von Social Search-Ansätzen, bereitstellen, die von den größeren Anbietern, die über eine eigene Suchinfrastruktur verfügen, nicht oder nur zögerlich angeboten werden[67].

[66] Vgl. http://metager.de/suma.html (Letzter Zugriff 21.04.2008).
[67] Für eine umfassende Darstellung dieser Dienste bleibt hier kein Platz. Der interessierte Leser findet aber beispielsweise unter der URL http://www.altsearchengines.com/ Informationen zu einer Vielzahl innovativer Retrievalsysteme und -ansätze, zumeist von Metasuchdiensten realisiert.

Aftervote[68] und Sproose[69] sind beispielsweise Social Search-Metasuchmaschinen, die es, ähnlich Wikia Search, registrierten Nutzern u.a. ermöglichen, Suchergebnisse zu bewerten und zu kommentieren.

Searchcrystal.com [70] ist eine Metasuchmaschine, welche über ein grafisches Display die Überlappung der abgefragten Suchdienste bzw. der Ergebnisse visualisiert. Ergebnisse, die von mehreren Suchdiensten gefunden werden, werden als besonders hochwertig betrachtet und z.B. im „Cluster display“ in das Zentrum der Ergebnisdarstellung gerückt, wie Abbildung 13 zeigt.

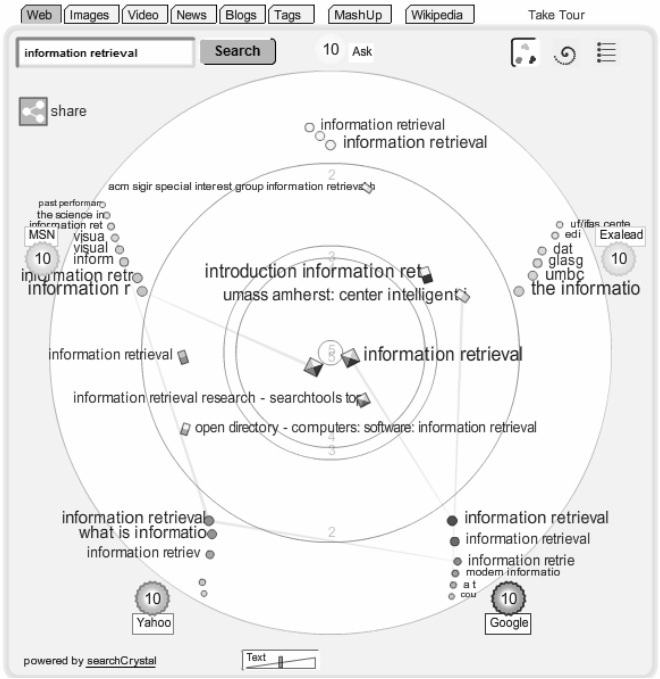

Abbildung 13. Searchcrystal.com Ergebnisvisualisierung

Des Weiteren ist es möglich, die Skala der Ergebnisdarstellung, die Text- oder Bildgröße interaktiv zu manipulieren. Die volle Funktionalität (Speichern von Suchanfragen, Boole´sche oder Ranking-Filter) wird erst nach einer Registrierung und Anmeldung bereitgestellt.

Sortfix [71] ist ein Metasuchdienst, der es gestattet, Suchanfragen über vorgeschlagene Terme, die per drag & drop in "Add to Search"- und "Remove"-Boxen gezogen werden können, zu verändern. Abbildung 14 zeigt die Benutzeroberfläche von Sortfix.

[68] URL http://www.aftervote.com/ (Letzter Zugriff 22.04.2008).
[69] URL http://www.sproose.com/static/tour/tour0.html (Letzter Zugriff 22.04.2008).
[70] URL http://www.searchcrystal.com/ (Letzter Zugriff 22.04.2008).
[71] URL http://sortfix.com/ (Letzter Zugriff 22.04.2008).

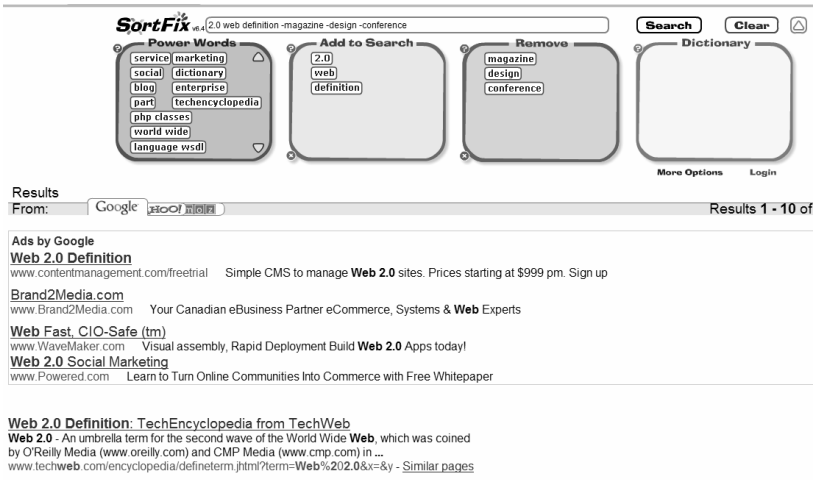

Abbildung 14. Suchanfragenmodifikation in Sortfix

Eine relativ neue Variante vertikaler Metasuchmaschinen stellen Dienste wie Pipl[72] und Yasni[73] dar. Diese Suchdienste aggregieren Informationen zu Personen und fragen dazu neben den populären Suchdiensten u.a. auch Soziale Netzwerke ab. Der/die interessierte Leser/in sei an dieser Stelle darauf hingewiesen, dass sich diese Dienste u.a. auch gut dazu eignen, zu überprüfen, welche persönlichen Informationen im Web für andere öffentlich und zugänglich sind.

6. Suchdienstemarkt

In den Kapiteln 2-4 wurde die konzeptuelle Funktionsweise der wesentlichen elementaren Suchdienstetypen dargestellt. Dabei wurden einerseits die zentralen Unterschiede sowie die Vor- und Nachteile manueller bzw. roboterbasierter Universalsuchdienste skizziert und zugleich deren Grenzen, insbesondere hinsichtlich der Bereitstellung von Informationen aus dem sogenannten Deep Web, sichtbar. Diesbezüglich stellen für Recherchierende die in Kap. 5 vorgestellten Spezialsuchdienste wichtige Ergänzungen dar.

Hinsichtlich der realen Ausprägung des Suchdienstemarktes lässt sich die hier vorgenommene Typologie, im Sinne einer Unterscheidung auch von Suchdienstleistern, seit einigen Jahren immer weniger aufrechterhalten. Vielmehr stellen gerade die populären Suchdienstleister Yahoo (ursprünglich ein Webkatalog) und Google (gestartet als Suchmaschine) mittlerweile Suchportale dar, die eine Vielzahl unterschiedlichster Suchdienstetypen unter einem Dach vereinen. Das Schlagwort „Universal Search“ kennzeichnet dabei einen Entwicklungstrend, der dadurch gekennzeichnet ist, dass diese Suchportale, allen voran Google, zunehmend dazu übergehen, neben den roboterbasierten Suchmaschinentreffern – die nach wie vor die primären Treffer liefern bzw. das technische Rückgrat der populären Suchdienste

[72] URL http://pipl.com/ (Letzter Zugriff 22.04.2008).
[73] URL http://www.yasni.de (Letzter Zugriff 22.04.2008).

bilden – weitere Ergebnisse aus einer Vielzahl von unterschiedlichen Spezialsuchdiensten in die Standardsuche, d.h. die Ergebnisseiten automatisch mit einzubinden [30]. Abbildung 15 veranschaulicht die Integration von Spezialsuchdiensten in die Standardsuchergebnisse bei Google und Yahoo.

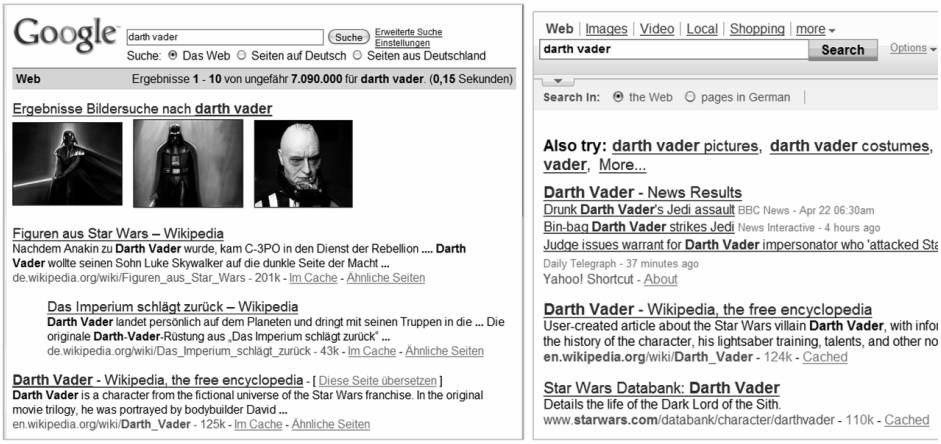

Abbildung 15. Einbindung von Treffern aus Spezialsuchdiensten

Damit folgen die populären Websuchdienste der Idee, in Abhängigkeit von der Art des Informationsbedürfnisses passende Dokumenttypen als Ergebnis auszugeben und dem Nutzer die Auswahl bzw. Anwahl vertikaler Suchdienste soweit als möglich abzunehmen. Aus der Sicht der Recherchierenden ist diese Entwicklung zu begrüßen. Aber inwieweit diese Ansätze geeignet sind, das Problem der „Tab Blindness" zu kompensieren, bleibt offen. Festzuhalten ist, „Universal Search" ist ein wichtiger Schritt, um unterschiedliche Datenquellen und Ergebnisdokumenttypen in Suchergebnisseiten zu integrieren[74]. In Kombination mit den in 4.2.4 genannten Ansätzen, die dahin zielen, menschliches Wissen und automatische Verfahren miteinander zu kombinieren, deutet sich eine Entwicklung hin zu deutlich komplexeren Retrievalsystemen bzw. Suchdiensten an.

Betrachtet man die Ergebnislisten nahezu aller Internetsuchdienste, so zeigt sich, dass die Treffer der bislang aufgeführten Suchdienstetypen nur einen Teil der Suchergebnisse darstellen, die von den Suchdiensten ausgeliefert werden. Gemeint ist die Suchmaschinenwerbung, die i.d.R. als „Anzeigen" oder „Sponsoren-Links" ausgezeichnet mit eingeblendet wird. Abbildung 16 zeigt die „organischen" und werbebasierten „Ad Words"-Suchergebnisse bei Google.

[74] Vgl. auch die Aussage von Mayer, M.: "While today's releases are big steps in making the world's information more easily accessible, these are just the beginning steps toward the universal search vision. Stay tuned!" URL http://googleblog.blogspot.com/2007/05/universal-search-best-answer-is-still.html (Letzter Zugriff 23.04.2008).

Abbildung 16. Pay per Click: Screen-Estate bei Google

Eine Typologie der Suchdienste wäre ohne die Darstellung dieser sogenannten Pay per Click-Suchdienste unvollständig. Zumal diese Dienste die finanzielle Grundlage bzw. Ertragskomponente nahezu aller, für den Nutzer kostenfreien, Suchdienste im Web darstellen. Dokumentbeschaffung und Indexierung ähneln bei Pay per Click-Suchdiensten der von Webkatalogen. Im Rahmen vorgegebener Rechts- und Qualitätsrichtlinien können Informationsanbieter Suchanfragen buchen und die Indexierungsangaben der von ihnen eingeblendeten Links, i.d.R. Titel, Beschreibungstext und URL, weitgehend frei bestimmen. Die eingeblendeten Links werden nach einer Kombination aus Gebotshöhe, Klickrate und weiteren Qualitätsfaktoren [75] spezifiziert. D.h. die Ergebnisse werden primär nach Zahlungsbereitschaft und Klickpopularität sortiert. Dies lässt sich auch qualitativ begründen, wenn man davon ausgeht, dass die Zahlungsbereitschaft des Informationsanbieters mit der Relevanz seines Angebots für den Informationsnachfrager korrespondiert.

Aus retrievaltechnischer Perspektive sind Pay per Click-Dienste vor allem deshalb interessant, weil sie Werbetreibenden seit Jahren eine Vielzahl von Steuerungsoptionen bereitstellen, die bei den Websuchmaschinen derselben Anbieter nicht zu finden sind. So ist es bei Google Ad Words gemäß der Option „Weitgehend passende Keywords" möglich, bei Suchanfragen auch Pluralformen und Synonyme zu berücksichtigen, d.h. die Werbung auch dann einzublenden, wenn Nutzer ähnliche, aber nicht gleiche Suchanfragen eingeben[76].

Historisch betrachtet entwickelten sich Pay per Click-Dienste seit Ende der 1990er Jahre, als der damalige Pionier Goto.com erstmals erfolgreich einen Suchdienst startete, der Suchanfragen vermarktete und Suchergebnisse nach Zahlungsbereitschaft sortierte [2]. Mittlerweile beliefern die Großen Pay per Click-Dienste von Google und Yahoo eine Vielzahl weiterer Suchdienste mit „Anzeigen" bzw. Sponsoren-Links.

In einer Gesamtperspektive betrachtet, verdeutlicht die bisherige Darstellung, dass es zu kurz greift, Google, Yahoo und andere populäre Suchdienste alleine als

[75] URL http://adwords.google.com/support/bin/answer.py?hl=de&answer=21388 (Letzter Zugriff 23.04.2008).

[76] URL http://www.google.com/intl/de/adwords/learningcenter/19135.html (Letzter Zugriff 23.04.2008).

roboterbasierte Websuchmaschinen zu begreifen. Vielmehr stellen sie ein komplexes Gerüst unterschiedlicher Suchdienstetypen dar. Abbildung 17 veranschaulicht die konzeptionelle Struktur der „Big Player“ im Suchdienstemarkt.

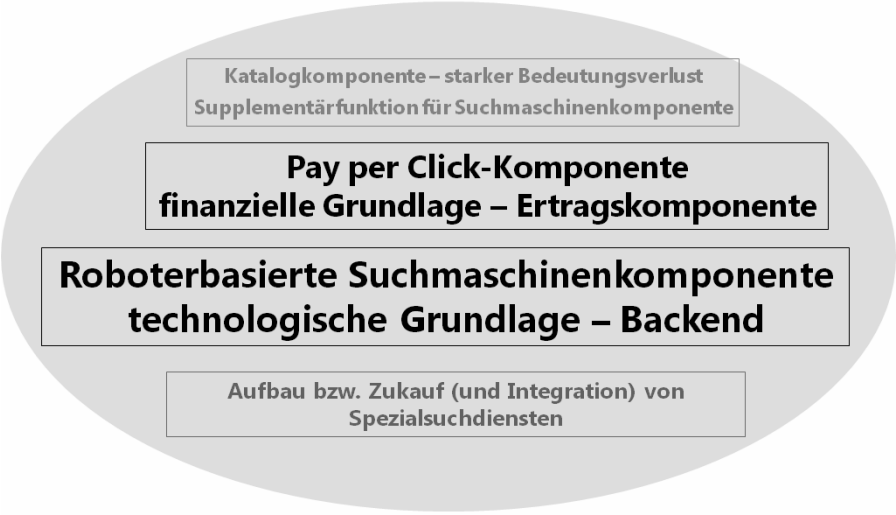

Abbildung 17. Struktur populärer Suchdienste

Abschließend bleibt zu konstatieren, dass auch die verschiedenen Suchdiensteanbieter selbst teilweise eng miteinander verflochten sind. Neben ihrer führenden Vermarkterrolle bezüglich Suchmaschinenwerbung bildet die Websuchmaschinentechnologie der zwei größten Suchdienste Google und Yahoo auch das technologische Rückgrat einer Vielzahl anderer Suchdienste[77].

7. Ergebnis & Ausblick

Der vorliegende Text gibt einen knappen Überblick über die wesentlichen methodischen Ansätze und Unterschiede elementarer Typen von Suchdiensten im Internet und die reale Ausprägung des Suchdienstemarkts. Dabei zeigt sich ein sehr vielfältiges und vielschichtiges Bild. Typisierungskategorien lassen sich vor allem hinsichtlich folgender Kriterien festlegen:

- Automatische vs. manuelle/intellektuelle Dokumentbeschaffung/Erschließung
- Universeller und spezialisierter Anwendungskontext
- Hoher vs. niedriger Strukturierungsgrad bezüglich der Erschließung und korrespondierend der Spezifikation der Treffermenge
- Spamresistenz.

Deutlich wird, dass jenseits von Katalogen und Websuchmaschinen eine Vielzahl weiterer Suchdienstetypen bereit stehen und auch, dass die populären Suchdienste weitaus mehr darstellen als Websuchmaschinen. Die derzeitigen Entwicklungsansätze zeigen auch, dass Web Information Retrieval mittlerweile weit über die lange Zeit fokussierten termbasierten Retrievalmodelle hinausreicht und derzeit als der Entwicklungsmotor in der Retrievalforschung begriffen werden kann.

[77] Vgl. URL http://www.bruceclay.com/searchenginerelationshipchart.htm (Letzter Zugriff 14.04.2008).

Web Information Retrieval wird zwar nach wie vor als primär technologisch ausgerichtete Forschungsdisziplin betrachtet [1], [2]. Allerdings verdeutlichen Publikationen aus unterschiedlichen wissenschaftlichen Disziplinen, so etwa der Titel „Auf dem Weg in die "Google-Gesellschaft"“ [42] wie auch die hohe Präsenz des Themas in den Massenmedien einerseits die Interdisziplinarität des Forschungsfelds [43] und andererseits die Relevanz des Themas für Millionen Endnutzer. Technologische Neuerungen (z.B. personalisierte Suche) werfen in Kombination mit Oligolipolisierungs- bzw. Monopolisierungstendenzen im Suchdienstemarkt ökonomische, soziale und politische Fragen auf, die in den Kern einer globalisierten Welt hineingreifen. Die letztgenannten Punkte gehen dabei zwar weit über den Fokus dieses Textes hinaus. Dennoch hoffen die Autoren mit dem hier gegebenen Überblick zu konzeptionellen Ansätzen der Suche im Internet einen Beitrag zu leisten, der weitergehende Diskussionen zu befruchten vermag.

Literaturangaben

[1] Dirk Lewandowski. Web Information Retrieval. Technologien zur Informationssuche im Internet. Informationswissenschaft; 7, DGI, 2005.
[2] Bernard Bekavac. Metainformationsdienste des Internet. In Rainer Kuhlen, Thomas Seeger, Dietmar Strauch, (Hg), Grundlagen der praktischen Information und Dokumentation, (1): 399-407, Saur, 2004.
[3] Chris Sherman and Gary Price. The Invisible Web: Finding Hidden Internet Resources Search Engines Can't See. Cyberage Books, 2001.
[4] Mike Moran and Bill Hunt. Search Engine Marketing Inc. IBM Press, 2006.
[5] Phil Bradley. Human-powered Search Engines: An Overview and Roundup. Ariadne, (54), 2008.
[6] Tim O'Reilly. What Is Web 2.0. Design Patterns and Business Models for the Next Generation of Software. 2005.
[7] Karin Regulski. Aufwand und Nutzen beim Einsatz von Social-Bookmarking-Services als Nachweisinstrument für wissenschaftliche Forschungsartikel am Beispiel von BibSonomy. Bibliothek. Forschung und Praxis, 2, 177-184, 2007.
[8] Andreas Hotho and Robert Jäschke and Christoph Schmitz and Gerd Stumme. Information Retrieval in Folksonomies: Search and Ranking. Proceedings of the 3rd European Semantic Web Conference, (4011): 411-426, Springer, Budva, Montenegro, 2006.
[9] Gerhard Knorz. Informationsaufbereitung II: Indexieren. In Rainer Kuhlen and Thomas Seeger and Dietmar Strauch, editor(s), Grundlagen der praktischen Information und Dokumentation, (1): 179-188, Saur, 2004.
[10] Jakob Voß. Tagging, Folksonomy & Co – Renaissance of Manula Indexing. In Achim Oßwald and Maximilian Stempfhuber and Christian Wolff, editor(s), Neue Perspektiven im Kontext von Information und Wissen. Proceedings des 10. Internationalen Symposiums für Informationswissenschaft (ISI 2007), 243-254, UVK Verlagsgesellschaft, Köln, 2007.
[11] Paul Heymann and Georgia Koutrika and Hector Garcia-Molina. Can social bookmarking improve web search?. WSDM '08: Proceedings of the international conference on Web search and web data mining, 195-206, ACM, New York, NY, USA, 2008.
[12] Soumen Chakrabarti. Mining the Web - Discovering Knowledge from Hypertext Data. Morgan Kaufmann, 2003.
[13] Reginald Ferber. Information Retrieval, dpunkt.verlag, 2003.
[14] Wolfgang G. Stock. Information Retrieval. Oldenbourg, 2007.
[15] A. Broder and R. Kumar and F. Maghoul and P. Raghavan and S. Rajagopalan and R. Stata and A. Tomkins and J. Wiener. Graph structure in the Web. Computer Networks, (33)1: 309-320, 2000.
[16] Michael K. Bergman. The Deep Web: Surfacing Hidden Value, 2001.
[17] Dirk Lewandowski and Philipp Mayr. Exploring the Academic Invisible Web. Library Hi Tech, (24)4: 529-539, 2006.
[18] Alexandros Ntoulas and Junghoo Cho and Christopher Olston. What's new on the web?: the evolution of the web from a search engine perspective. International World Wide Web Conference archive. Proceedings of the 13th international conference on World Wide Web, 1-12, ACM, 2004.
[19] Dirk Lewandowski. A three-year study on the freshness of Web search engine databases. Journal of Information Science, (34), 2008.

[20] Monika Henzinger and Steve Lawrence. Extracting knowledge from the World Wide Web. Proceedings of the National Academy of Sciences of the United States of America, 5186-5191, 2004.
[21] Nadine Schmidt-Mänz. Untersuchung des Suchverhaltens im Web: Interaktion von Internetnutzern mit Suchmaschinen. Kovač, Hamburg, 2007.
[22] SEOmoz.org, Google Search Engine Ranking FactorsV2, URL http: //www.seomoz.org/article/search-ranking-factors (Letzter Zugriff 18.04.2008).
[23] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. In Philip H. Enslow, Jr. and Allen Ellis, editor(s), Computer Networks and ISDN Systems. Proceedings of the Seventh International World Wide Web Conference, (30)1-7: 107-117, Elsevier, 1998.
[24] Eugene Garfield. Citation analysis as a tool in journal evaluation. Science, (178): 471-479, 1972.
[25] Thomas Mandl. Automatische Bewertung der Qualität von Web-Seiten. Erscheint 2008.
[26] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, (46)5: 604-632, 1999.
[27] Thomas Mandl. Implementation and evaluation of a quality-based search engine. Proceedings of the 17th ACM Conference on Hypertext and Hypermedia (HT '06) Odense, Denmark, August 22nd-25th., 73-84, ACM Press, 2006.
[28] David Pennock and Gary Flake and Steve Lawrence and Eric Glover and C. Lee Giles. Winners don't take all: Characterizing the competition for links on the web. Proceedings of the National Academy of Sciences, (99) 8: 5207-5211, 2002.
[29] Kai Riemer and Fabian Brüggemann. Personalisierung der Internetsuche. Lösungstechniken und Marktüberblick. Wirtschaftsinformatiik, (49)2: 116-126, 2007.
[30] Joachim Griesbaum. Entwicklungstrends im Web Information Retrieval: Neue Potentiale für die Webrecherche durch Personalisierung & Web 2.0-Technologien. In Marlies Ockenfeld, editor(s), nformation in Wissenschaft, Bildung und Wirtschaft. Proceedings der 29. Online-Tagung der DGI, 91-111, DGI, Frankfurt a.M., 2007.
[31] Barry Smyth and Evelyn Balfe. Anonymous personalization in collaborative web search. Information Retrieval, (9)2: 165-190, Springer, 2006.
[32] Andrei Broder. From query based Information Retrieval to context driven Information Supply, 2006.
[33] Dirk Lewandowski. Datumsbeschränkung bei WWW-Suchanfragen: Eine Untersuchung der Möglichkeiten der zeitlichen Einschränkung von Suchanfragen in den Suchmaschinen Google, Teoma und Yahoo. In Bernard Bekavac and Josef Herget and Marc Rittberger, (Hg.): Informationen zwischen Kultur und Marktwirtschaft. Proceedings des 9. Internationalen Symposiums für Informationswissenschaft 2004. Konstanz: UVK Verlagsgesellschaft mbH, 2004. S. 301 – 316
[34] Rebecca Blood. The Weblog Handbook: Practical Advice on Creating and Maintaining Your Blog. Perseus Books, Cambridge MA, 2002.
[35] Marcel Machill; and Dirk Lewandowski. Journalistische Aktualität im Internet: Ein Experiment mit den News-Suchfunktionen von Suchmaschinen. In Marcel Machill and Norbert Schneider, editor(s), Suchmaschinen: Herausforderungen für die Medienpolitik, 105-164, Vistas, Berlin, 2005.
[36] Dirk Lewandowski. Nachweis deutschsprachiger bibliotheks- und informationswissenschaftlicher Aufsätze in Google Scholar. IWP - Information: Wissenschaft und Praxis, (58)3: 165-168, 2007.
[37] Alexei Yavlinsky. Behold: a content based image search engine for the World Wide Web. 2006.
[38] Michael S. Lew and Nicu Sebe and Chabane Djeraba Lifl and Ramesh Jain. Content-based Multimedia Information Retrieval: State of the Art and Challenges. ACM Transactions on Multimedia Computing, Communications, and Applications, (2)1: 1-19, 2006.
[39] Joachim Griesbaum. Zur Rolle von Websuchdiensten und Fachinformation im Suchverhalten von Studierenden. Befunde einer explorativen Studie. In Harald Weigel, editor(s), Wa(h)re Information, 29. Österreichischer Bibliothekartag, Bregenz, 19.- 23. September 2006, 174-182, Neugebauer Verlag, Graz-Feldkirch, 2007.
[40] Joachim Griesbaum and Bernard Bekavac. Web-Suche im Umbruch? Entwicklungstendenzen bei Web-Suchdiensten. In Bernard Bekavac and Marc Rittberger and Josef Herget, editor(s), Information zwischen Kultur und Marktwirtschaft. Proceedings des 9. Internationalen Symposiums für Informationswissenschaft (ISI 2004), Chur, 6.-8.Oktober 2004, 283-299, UVK Verlagsgesellschaft mbH, Konstanz, 2004.
[41] Amy N. Langville and Carl D. Meyer. Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University Press, 2006.
[42] Rudi Schmiede. Auf dem Weg in die "Google-Gesellschaft". In Friedemann Mattern (Hg), Wie arbeiten die Suchmaschinen von morgen? Informationstechnische, politische und ökonomische Perspektiven, 127-133, Fraunhofer IRB Verlag, 2008.
[43] Marcel Machill and Markus Beiler and Martin Zenker. Suchmaschinenforschung: Überblick und Systematisierung eines. In Friedemann Mattern (Hg), Wie arbeiten die Suchmaschinen von morgen? Informationstechnische, politische und ökonomische Perspektiven, 23-58, Fraunhofer IRB Verlag, 2008.

[44] Stock, W. G.. Folksonomies and science communication: A mash-up of professional science databases and Web 2.0 services. Information Services & Use, 27, 97-103, (2007).
[45] Rolf Däßler. Informationsvisualisierung - Stand, Kritik und Perspektiven. In Methoden/Strategien der Visualisierung in Medien, Wissenschaft und Kunst. Trier: Wissenschaftlicher Verlag Trier.
[46] Bernard Bekavac and Josef Herget and Sonja Hierl and Sonja Öttl. Visualisierungskomponenten bei Web-basierten Suchmaschinen: Methoden, Kriterien und ein Marktüberblick. In IWP - Information Wissenschaft & Praxis 58 (2007) 3, S. 149-158.