Qualitätsmessung bei Suchmaschinen

System- und nutzerbezogene Evaluationsmaße

erscheint in: Informatik Spektrum 30(2007)3

Prof. Dr. Dirk Lewandowski
Hochschule für Angewandte Wissenschaften Hamburg, Fabultät DMI, Department Information, Berliner Tor 5, 20099 Hamburg
E-Mail: dirk.lewandowski@bui.haw-hamburg.de

Dr. Nadine Höchstötter
Institut für Entscheidungstheorie und Unternehmensforschung, Universität Karlsruhe (TH), Kaiserstr. 12, 76128 Karlsruhe
E-Mail: nsh@topicflux.de

Zusammenfassung: Suchmaschinen stellen mittlerweile das wichtigste Medium dar, um Inhalte im Internet zu finden. Es gab deswegen in der Vergangenheit verschiedene Studien, welche die Qualität von Suchmaschinen untersuchten. Bei der Bewertung der Qualität wurden bisher jedoch nur wenige und oft sehr einseitige Aspekte benutzt. Die Qualität einer Suchmaschine wird meist mit der Bewertung der gefundenen Dokumente oder ihrer Inhalte gleichgesetzt. In dieser Arbeit wird der Fächer der messbaren Qualitätsaspekte aufgespannt, um zu zeigen, dass vielfältigere Maße nötig sind. Es wird nicht nur die Technik von Suchmaschinen bewertet, sondern es werden auch die Nutzer der Suchmaschinen einbezogen, um eine umfassendere Sicht auf die Qualitätsmessung zu bekommen.
Abstract: Search engines are the most important medium to find content on the web. In the past, several studies were conducted, which examined the quality of search engines. But, this evaluation of the quality so far only include a few and often very one-sided aspects. The quality of a search engine is usually equated with its ability to retrieve relevant results. In this work, we will present a full spectrum of measurable aspects of quality, in order to show that more varied measures are necessary, which do not only evaluate the technology of search engines, but also involve the users of search engines in order to get a more comprehensive view on the quality measurement of search engines.

Keywords: World Wide Web, Suchmaschinen, Qualitätsmessung, Nutzerverhalten, Benutzerfreundlichkeit, Retrievaltests

Suchmaschinen sind mittlerweile die wichtigsten Instrumente für die Informationsbeschaffung. Doch noch ist weitgehend unklar, wie gut sich die verschiedenen Suchmaschinen tatsächlich für die Recherche eignen.

Einleitung

Suchmaschinen haben einen hohen Stellenwert bei der alltäglichen Informationssuche in privaten wie beruflichen Bereichen erlangt und sind nach der E-Mail-Nutzung der meistgenutzte Dienst des Internet. Bis jetzt wurden in der Literatur verschiedene Qualitätsaspekte, wie die Qualität und Aktualität der Ergebnisse von Suchmaschinen diskutiert, wobei die Bewertung der Qualität in Bezug auf die Nutzer der Suchmaschinen weitgehend vernachlässigt wurde. Die Frage hierbei ist, wie gut Suchmaschinen tatsächlich aus der Sicht der Nutzer beziehungsweise Entwickler oder Anbieter von Online-Inhalten arbeiten und welche Evaluationsmaße erhoben werden müssen, um ein Gesamtbild auf die Qualität zu bekommen.
Vaughan [1] stellt hierfür spezielle Retrievalmaße vor. Aber die Qualität einer Suchmaschine basiert nicht nur auf der Performance beim Auffinden von Informationen oder Dokumenten. Die Qualität des Index, welcher einer Suchmaschine zugrunde liegt, und die Anwendbarkeit von speziellen Suchfunktionen, die den Suchenden unterstützen sollen, sind ebenso wichtig.

Aber ein verbesserter Ranking-Algorithmus oder verschiedene angebotene Serviceleistungen erhöhen nicht automatisch die Zufriedenheit eines Nutzers. Die Messung der Qualität ist für die Betreiber von Suchmaschinen wichtig, um Marktanteile zu sichern und eine gegebene Wechselbereitschaft von Kunden abzuwehren. Die Zufriedenheit eines Nutzers ist eine wichtige Voraussetzung, um eine Bindung an ein Produkt oder eine Dienstleistung herbeizuführen.
Bei der Bewertung von Informationssystemen findet mittlerweile ein Wechsel von einer rein technischen Sichtweise zu einer nutzerorientierten Sichtweise statt (vgl. [2]). Deswegen wird hier ein integrativer Ansatz gewählt, um die Qualität von Suchmaschinen zu messen, in der vor allem auch der Nutzer in den Vordergrund gerückt wird. Diese integrierte Betrachtungsweise von Technik und Nutzer wird zeigen, dass durchaus noch Möglichkeiten der Verbesserung vorhanden sind.
Die Qualität von gefundenen Informationen wird normalerweise durch die Bewertung eines Dokuments oder einer Menge von Dokumenten gemessen. Dabei müssen jedoch zwei Herangehensweisen unterschieden werden. Auf der einen Seite ist es wichtig, wie der Index aufgebaut wird und welche Dokumente dafür ausgewählt werden. Auf der anderen Seite spielt es eine große Rolle, wie Dokumente von einem Information-Retrieval-System wieder gefunden werden. Letzteres ist leicht auf den Web-Kontext übertragbar, während der erste Punkt nicht so einfach durchzuführen ist.
Ein Ansatz dabei ist, ein möglichst vollständiges Abbild des Web zur Verfügung zu stellen. Der Ansatz, nur bestimmte Dokumente zur Verfügung zu stellen (vertikale Suche), wird momentan seltener von den bekannten Suchmaschinen durchgeführt. Der Aufbau des Index ist ein wichtiges Maß bei der Bewertung der Qualität von Suchmaschinen.
Bei der Diskussion der Qualität von Suchergebnissen muss in Erinnerung gerufen werden, wie Suchmaschinen deren Relevanz berechnen. Ein Hauptfokus des Rankings liegt auf der Berechnung der Popularität einer Webseite, die durch die eingehenden Links und die Struktur des Webgraphen gemessen wird [3,4]. In den Rankingprozess werden jedoch keine Bewertungen von Webseiten durch Personen eingearbeitet. Der Linkpopularität wird diese Bewertung impliziert; link-basierte Maße sind unabhängig von der Suchanfrage [5]. Wenn ein Suchmaschinen-Nutzer eine Suchanfrage stellt, wird zusätzlich gemessen, wie diese Anfrage zu den Inhalten verschiedener Webseiten passt, um die besten vorzuschlagen. Hierbei spielt auch die Häufigkeit eines Suchwortes eine Rolle. Basierend auf beiden Maßen wird ein Ranking-Algorithmus entwickelt, der eine Suchergebnisliste generiert. Weitere Details zur Funktionsweise von Suchmaschinen sind in [5] zu finden.
Um das Ranking adäquat zu berechnen, ist es ebenfalls wichtig, Kenntnisse über das Nutzerverhalten zu haben. Um dieses Verhalten zu erheben, bieten sich vier Methoden an (siehe Tabelle 1). Zum einen können Studien durchgeführt werden, indem Suchmaschinen-Nutzer in einer Umfrage direkt zu ihrem Suchverhalten befragt werden. Der Vorteil dieser Methode liegt darin, dass auch demographische Merkmale verfügbar sind, durch welche die Nutzer charakterisiert werden können und welche bei der Definition von Nutzertypen helfen können. Ein großer Nachteil besteht darin, dass sich die Befragten oft in ein besseres Licht zu stellen versuchen, indem sie sich professioneller in ihrer Suchmaschinen-Nutzung darstellen. rch kann eine Verzerrung der Ergebnisse auftreten.
Eine weitere Möglichkeit ist die Beobachtung von Personen bei der Suche in einem Laborexperiment. Der Vorteil liegt in der detaillierten Beschreibung der einzelnen Schritte. Die Probanden können zusätzlich nach dem Grund ihres Verhaltens befragt werden. Meist werden bei diesen Laborexperimenten auch Fragen zur Demographie gestellt. Ein großer Nachteil besteht im Aufwand und den damit verbundenen Kosten eines solchen Laborexperiments, oft sind dadurch auch nur wenige Probanden bei den Versuchen beteiligt.
Eine objektivere Sicht auf das Suchverhalten bieten Logfiles von Suchmaschinen. Damit können große Stichproben genommen werden, zusätzlich kann das Verhalten über lange Zeiträume erhoben werden. Es sind damit auch zeitabhängige Auswertungen von Suchanfragen möglich. Nachteilig ist das Fehlen jeglicher demographischer Merkmale. Einzelne Suchsitzungen (Sessions) können ebenfalls betrachtet werden, das heisst die Interaktion des Suchenden, die mit dem Aufruf der Suchmaschinenseite gestartet und mit der Schließung des Fensters beendet wird.
Unter die objektive Betrachtung ist auch die Beobachtung von Livetickern einzuordnen. In Livetickern werden aktuell an die Suchmaschine gestellte Suchanfragen gezeigt (beispielsweise http://www.lycos.de/suche/livesuche.html). Hier tritt zusätzlich das Problem auf, dass Suchanfragen nicht mehr zu einer eindeutigen IP-Adresse zugeordnet werden können. Auswertungen auf Session-Basis sind deshalb nicht möglich. Bei manchen Livetickern werden zudem sexuelle oder andere Inhalte nicht angezeigt, wodurch ebenfalls eine Verfälschung der Ergebnisse auftreten kann.

Tabelle 1 Methoden zur Datenerhebung des Nutzerverhaltens [6]

Methode	**Vorteile**	**Nachteile**
Nutzerbefragung	Demographie, umfangreiche Fragen und Gruppierungen möglich	Beeinflussung, Befragte wollen sich besser darstellen
Laboruntersuchung	Sehr ausführliche Tests, Nutzer können ihr Verhalten direkt erklären oder kommentieren	Zeit – und kostenintensiv, wenige Teilnehmer, Beeinflussung durch Labor-Charakter

Live Ticker	Beobachtung, keine Beeinflussung	Keine IP, keine Auswertung von Sessions, keine Demographie
Logfile-Analyse	Beobachtung, keine Beeinflussung, IP vorhanden, Auswertung auf Session-Basis möglich	Keine Demographie

Im Folgenden wird die Qualitätsmessung von Suchmaschinen anhand von Evaluationsmaßen und aktuellen Studien vorgestellt. Empirische Erhebungen, die durchgeführt wurden, dienen dabei neben der Darstellung der eher technischen Betrachtungsweise auch der Darstellung der Sichtweise der Nutzer beziehungsweise deren Verhalten bei der Suche und beim Umgang mit Suchmaschinen. Durch das Suchverhalten von Internetnutzern können wiederum Probleme bei der geeigneten Berechnung der Ranking-Algorithmen entstehen, da beispielsweise Suchanfragen aus wenigen Wörtern zusammengesetzt werden oder nur die ersten Ergebnisse angeschaut werden. Das Suchverhalten wird also nicht nur direkt auf die Bewertung der Qualität von Suchmaschinen Einfluss nehmen, sondern auch eine Anpassung der Ranking-Algorithmen mit sich bringen.

Evaluationsmaße für Suchmaschinen

Die Bewertung der Qualität von Suchmaschinen geht über die Einteilung hinaus, ob gefundene Webdokumente relevant oder nicht relevant sind. Die Relevanz von gefunden Dokumenten spielt zwar die wichtigste Rolle, aber nicht die einzige.
In [7,8] wird das SERVQUAL (Service und Quality) Modell [9] auf Suchmaschinen angewendet. Hierbei wird die Qualität eines Services anhand verschiedener Kriterien abgefragt und mit den Erwartungen der Nutzer verglichen. Jedes Mal, wenn die gewünschte Qualität nicht den Erwartungen entspricht, entsteht eine Kluft, die ein Gefühl der Unzufriedenheit aufkommen lässt. Das Problem bei der Anwendung dieses Modells ist die ausschließliche Betrachtung des Nutzers, ohne auf technische Details zu achten.
Bei der klassischen Herangehensweise, um die Güte von Information-Retrieval-Systemen zu messen, kommen objektive Maße wie Precision und Recall zum Einsatz. Die Precision gibt dabei an, inwieweit ein Information-Retrieval-System relevante Resultate erzielt. Precision ist das Verhältnis der relevanten Dokumente zu allen vom System ausgegebenen Dokumenten. Der Recall hingegen gibt das Verhältnis der relevanten Dokumente, die gefunden wurden, zu allen relevanten Dokumenten wieder, die es zu einer Suchanfrage gibt.
Einen Überblick über weitere klassische Retrievalmaße gibt Korfhage [10]. Neuere Maße sind z.B. bei Greisdorf und Spink [11] oder bei Su [12] zu finden.
In der Geschichte der Suchmaschinen wurde jedoch schon früh erkannt, dass aufgrund der Eigenheiten der Web-Suche Web-spezifische Evaluationsmaße entwickelt werden müssen. Hierbei werden in Experimenten die Bewertungen beziehungsweise das Klick-Verhalten von Suchmaschinennutzern miteinbezogen. Beispielsweise wird die Korrelation zwischen dem Ranking von Web-Dokumenten durch Suchmaschinen und dem durch Suchende berechnet [1]. Diese Maße werden aber meist nur in Experimenten verwendet und kaum in realen Evaluationen eingesetzt.
Zieht man sowohl den system- als auch den nutzerzentrierten Ansatz in Betracht, kann der Ansatz von Lewandowski [13] um die Dimension der Usability erweitert werden. Es werden so vier grundsätzlich Evaluationsbereiche vorgestellt, die sich aus verschiedenen Evaluationsmaßen zusammensetzen:

- Qualität des Index: Hierunter fallen Evaluationsmaße wie die Größe und Vollständigkeit des Index, deren Aktualität sowie länderspezifische Unterschiede zwischen Indizes.
- Qualität der Suchresultate: Dies ist der Teil, bei dem gerade klassische Retrievalmaße angewendet werden, womit die Retrievaleffektivität der Systeme gemessen wird. Bei der Betrachtung von Suchmaschinen ist die Einzigartigkeit von Suchergebnissen beim Vergleich verschiedener Suchmaschinen interessant. Hierbei ist anzumerken, dass ein Großteil der Nutzer von sich sagen, dass sie in der Regel ohne größere Probleme in den Suchmaschinen das finden, was sie suchen (vgl. [14,15]). Das Problem dabei ist, dass Nutzer nicht wirklich die meist zahlreichen Suchergebnisse miteinander vergleichen können. Man muss davon ausgehen, dass Suchmaschinennutzer denken, dass sie finden, was sie suchen, da sie nicht wissen, wie das Angebot insgesamt aussieht.
- Qualität der Suchfunktionen: Bei einer Suchmaschine sollten spezielle Suchfunktionen angeboten werden, die den Nutzer bei der Suche unterstützen (beispielsweise Angabe der gewünschten Sprache oder eines Dokumenttyps) und ihm damit erlauben, die Qualität der Ergebnisse zu beeinflussen, indem beispielsweise ungewünschte Resultate ausgeschlossen werden.
- Nutzerfreundlichkeit von Suchmaschinen (Usability): Es gibt verschiedene Berührungspunkte von Nutzern mit Suchmaschinen, die widerspiegeln, wie gut Suchende mit Suchmaschinen umgehen können. Es werden dazu das Design des Interface betrachtet, die Akzeptanz spezieller Suchfunktionen und Operatoren, die Verarbeitung der Anfragen und die Benutzerführung. Alles in allem muss darauf geachtet werden, dass die

Nutzung von Suchmaschinen intuitiv und einfach zu bewerkstelligen ist, da Internetnutzer oft nicht die Funktionsweise und alle Möglichkeiten von Suchmaschinen kennen und deswegen Suchmaschinen nicht optimal einsetzen können.

Die vorgestellten Evaluationsmaße geben eine ganzheitliche Sicht auf die Qualität von Suchmaschinen und bewerten diese nicht nur nach klassischen Information-Retrieval-Methoden. Der Umgang und die Bedienbarkeit von Suchmaschinen werden dabei miteinbezogen. Dujmovic und Bai [16] gehen beispielsweise auf die Nutzbarkeit ein.

Empirische Ergebnisse

Im Folgenden werden Ergebnisse aus verschiedenen Studien zu den oben aufgeführten Punkten zusammengetragen. In diesen werden viele Teilaspekte der Qualität von Suchmaschinen behandelt, was fehlt, ist jedoch eine Untersuchung, die ein Gesamtbild auf die Qualität von Suchmaschinen gibt und zusätzlich das Suchverhalten von Suchmaschinen-Nutzern einbezieht. In dieser Hinsicht ist der in diesem Artikel gegebene Überblick als eine Vorarbeit zu einer solchen umfassenden Untersuchung zu sehen.

Qualität des Index

Der Index ist die Basis einer Suchmaschine. Die Art und Weise, wie er aufgebaut und gepflegt wird, hat einen entscheidenden Einfluss auf die Qualität einer Suchmaschine. Die Größe des Index (und damit die Abdeckung des Web) ist dementsprechendwichtig.
Doch ist nicht unbedingt die Suchmaschine die beste, die alle verfügbaren Seiten im Index hat (abgesehen davon sind alle Suchmaschinen weit von diesem Ziel entfernt). Außerdem ergeben sich Unterschiede, wenn man den Index verschiedener Länder betrachtet (z.B. Vergleich eines deutschen Index mit einem chinesischen). Da sehr oft aktuelle Inhalte zu Nachrichten und Events nachgefragt werden [17], ist die Aktualität ebenfalls ein wichtiges Merkmal, an dem die Qualität des Index festgestellt werden kann. Aber nicht nur wegen der Suche nach Nachrichten sollte ein Index aktuell sein. Es ist verständlich, dass eine Datenbank immer möglichst aktuell gehalten werden sollte.

Größe und Vollständigkeit des Index

Zur Erhebung der Größe des Web und der Abdeckung durch eine Suchmaschine gibt es drei Herangehensweisen:

- Selbstauskunft der Suchmaschinenbetreiber: Suchmaschinen geben manchmal die Menge der indizierten Seiten an, um zu zeigen, dass der Index weiter gewachsen ist, oder um nachweislich den größten Web-Index für sich reklamieren zu können.
- Maß der Überschneidung: Mit der Darstellung der Überschneidung der Indizes verschiedener Suchmaschinen kann die Zahl der insgesamt indizierten Seiten gemessen werden. Ein Problem besteht hierbei darin, dass alle Seiten, die in keinem der untersuchten Indizes vorhanden sind, nicht mit in die Berechnung eingehen.
- Zufallsstichproben: Es werden zufällig IP-Adressen herausgegriffen, die auf Verfügbarkeit getestet werden. Im nächsten Schritt wird geprüft, inwieweit diese Webseiten auch in Suchmaschinen zu finden sind.

Ein (auch zeitlicher) Vergleich der Indexzahlen auf Basis der Betreiberangaben ist bei SearchEngineWatch.com zu finden [18]. Die Abdeckung wird nicht abgeschätzt, aber es ist zumindest möglich, die Indexgrößen der angegebenen Suchmaschinen zu vergleichen. Allerdings muss man hier der korrekten Angabe durch Suchmaschinen vertrauen.
Zwei der wichtigsten Studien, welche die Größe des Web und dessen Abdeckung durch Suchmaschinen untersuchen, wurden von Bharat und Broder [19] beziehungsweise Lawrence und Giles [20,21] durchgeführt. Die Ergebnisse bei Bharat und Broder [19] zeigen, dass Suchmaschinen im Jahr 1997 eine mittlere Web-Abdeckung von 62 Prozent hatten, während die Überschneidung zwischen den einzelnen Suchmaschinen nur bei 1,4 Prozent lag. Die Erhebung wurde 1997 durchgeführt, die totale Anzahl der Seiten im Web wurde auf 200 Mio. geschätzt.
In der Studie von (Lawrence und Giles, 1998) wurde die Größe des Web durch die Überschneidung der zwei größten Suchmaschinen geschätzt. Sie kamen dabei auf 320 Mio. Seiten. Die Abdeckung der Suchmaschinen reichte von 3 bis zu 34 Prozent.

In einer aktuellen Studie von Gulli und Signori [22] wurde gezeigt, dass das indizierbare Web bei mindestens 11,5 Milliarden Webseiten liegt. Die Abdeckung des Web durch die Suchmaschinen Google, Yahoo, MSN und Ask liegt nach dieser Untersuchung zwischen 57 und 76 Prozent.
Alle Studien, welche die Abdeckung des Web durch Suchmaschinen betrachten, zielen nur auf das indizierbare Web ab (Surface Web), welches aber nur einen Teil des Web darstellt. Der Rest besteht aus dem so genannten *Invisible* oder *Deep Web*. In diesen Teil fallen Dokumente, die entweder durch Suchmaschinen nicht aufgenommen werden können oder die nicht aufgenommen werden sollen [23]. Unter den letzten Punkt fallen alle Spam-Seiten, die von den Suchmaschinen willentlich nicht indexiert werden. Oft wertvolle Inhalte, die eigentlich für die Suchmaschinen interessant wären, liegen in Datenbanken verborgen, auf die die Suchmaschinen-Crawler nicht zugreifen können. Dieser Bereich macht den wesentlichen Teil des Invisible Web aus.
Bergman [24] schätzt die Größe des Deep Web auf das 550-fache des Surface Web. Lewandowski und Mayr [25] fanden allerdings heraus, dass diese Zahl zu hoch geschätzt wurde. Das Invisible Web ist jedoch weiterhin weitgehend unerforscht.

Länder-Bias

Für die unterschiedlichen Länder ist es natürlich wünschenswert, dass ein möglichst hoher Anteil des „nationalen“ Web in den Indizes der Suchmaschinen enthalten ist. Eine alleinige gute Abdeckung des entsprechenden Sprachraums ist hier nicht ausreichend, wenn man beispielsweise an spanischsprachige Seiten aus Südamerika denkt, die für Nutzer in Spanien von geringerer Bedeutung sind als Seiten, die tatsächlich in Spanien erstellt wurden.
Vaughan und Thelwall [26] untersuchten die Unterschiede zwischen drei großen Suchmaschinen (Google, AlltheWeb und AltaVista) für verschiedene Länder. Die Länder, die in Betracht gezogen werden, sind die USA, China und Singapur und Taiwan. Diese Länder wurden auf Grund ihrer sprachlichen und ihrer kulturellen Unterschiede ausgesucht. Die Ergebnisse zeigten Unterschiede zwischen den Ländern und den Suchmaschinen. Die US-Seiten zeigten mit 80 bis 87 Prozent die beste Abdeckung, während diese für China bei 52 bis 70 Prozent liegt. Singapur liegt zwischen 41 und 56 Prozent. Taiwan erreichte eine Abdeckung von 4 bis 75 Prozent. Betrachtet man zusätzlich die Tiefe der Indexierung (also welcher Anteil der auf einem Server vorhandenen Seiten jeweils von den Suchmaschinen indexiert werden), ergeben sich im Mittel für die US-Seiten eine Indexierungstiefe von 89 Prozent, während es in China nur 22 Prozent sind.
Diese Zahlen geben einen Eindruck der Unterschiede von Indizes verschiedener Länder. Ein Vergleich europäischer Länder ist hier anzustreben, nachdem die Diskussion einer europäischen Suchmaschine verstärkt auftritt und das Thema gerade in Europa aufgrund der Sprachenvielfalt von großer Bedeutung ist.

Aktualität des Index

Die Aktualität ist aus zweierlei Gründen wichtig. Zum einen sollte ein Suchmaschinenindex aktuell sein, da häufig nach neuen Inhalten gesucht wird, und zum anderen spielt die Aktualität für das Ranking von Webseiten ebenfalls eine Rolle [27,28].
Eine Studie von Ntoulas et al. [29] ergab, dass ein großer Anteil aller Webseiten regelmäßig erneuert wird. Die Autoren geben an, dass jede Woche ungefähr 320 Mio. neue Webseiten entstehen. Ungefähr 20 Prozent der Webseiten, die heute zu finden sind, werden in einem Jahr nicht mehr vorhanden sein. Innerhalb des gleichen Zeitraums wird die Hälfte der Inhalte und ca. 80 Prozent der Links eine Änderung erfahren haben. Diese Hochrechnungen zeigen die Wichtigkeit der Aktualität des Index.
Notess [30] untersuchte das durchschnittliche Alter der Indizes von verschiedenen Suchmaschinen. Sie zeigten, dass zu künstlich erzeugten Suchanfragen bei den großen Suchmaschinen wie beispielsweise Google und MSN aktuelle Seiten und solche angezeigt wurden, die einen Tag alt waren. Kleinere Suchmaschinen wie Gigablast, Teoma und Wisenut enthielten Webseiten die bis zu 40 Tage alt waren. In diesem Fall kann nicht mehr von einer Aktualität des Index gesprochen werden. Im Durchschnitt gesehen schnitten alle Suchmaschinen schlechter ab. Selbst bei den großen Suchmaschinen wie Google lag das Durchschnittsalter bei einem Monat. Hier ist aber anzumerken, dass nicht jede Webseite täglich erneuert wird.
In der Untersuchung von Lewandowski, Wahlig und Meyer-Bautor [31] wurden zur Untersuchung der Aktualität des Index 38 deutsche Webseiten ausgesucht, die täglich erneuert werden. Hier wurden die Suchmaschinen Google, Yahoo und MSN untersucht. Das Ergebnis ist, dass Google die aktuellste Suchmaschine ist, da hier der größte Anteil der untersuchten Seiten täglich erneuert wird. MSN frischt den Index regelmäßig vollständig auf, während bei Yahoo keine geeignete Erneuerungsstrategie erkennbar ist.

Qualität der Suchresultate

Retrievaleffektivität

Zur Messung der Retrievaleffektivität von Information-Retrieval-Systemen im Allgemeinen hat die Informationswissenschaft in den letzten etwa 50 Jahren ein umfangreiches Instrumentarium entwickelt. Dieses wird seit einigen Jahren auch auf den Bereich der Web-Suchmaschinen angewendet beziehungsweise entsprechend weiterentwickelt, um für diesen Bereich geeignet einsetzbar zu sein.
Zunächst einmal lassen sich Ergebnisse aus Information-Retrieval-Systemen in relevante und nicht relevante Ergebnisse unterteilen. Bei einer solchen zweiwertigen Unterscheidung ergeben sich vier unterschiedliche Trefferbereiche (*hits, misses, noise, rejects*), wie sie in Tabelle 2 dargestellt werden. Sowohl die relevanten als auch die nicht relevanten Treffer können jeweils danach unterteilt werden, ob sie auf eine Suchanfrage hin vom System ausgegeben werden oder nicht.
Idealerweise würde ein Suchergebnis aus allen relevanten Treffern, die im Datenbestand enthalten sind, bestehen. Es enthielte also einerseits *nur* relevante Treffer, andererseits *alle* relevanten Treffer (sofern sie im System enthalten sind). Ein solches System erreichte sowohl hinsichtlich der Präzision der Suchergebnisse als auch hinsichtlich ihrer Vollständigkeit das Maximum.

Tabelle 2 Mögliche Retrievalergebnisse [32], S. 188

	Relevant	Not Relevant	Total
Retrieved	a (hits)	b (noise)	a+b (all retrieved)
Not retrieved	c (misses)	d (rejects)	c+d (all nonretrieved)
Total	a+c (all relevant)	b+d (all nonrelevant)	a+b+c+d (total in the system)

Entsprechend dieser Idealvorstellung wurden die mittlerweile „klassischen“ Retrievalmaße entwickelt. Das bedeutendste (und in Retrievaltests meist verwendete) Maß ist die *Precision*. Diese gibt den Anteil der relevanten ausgegebenen Treffer an der Gesamtheit der ausgegebenen Treffer an. Dieses Maß ist relativ leicht zu bestimmen: Alle ausgegebenen Treffer werden einem oder mehreren Juroren zur Beurteilung vorgelegt und anschließend wird ausgezählt, wie hoch der Anteil der relevanten beziehungsweise nicht relevanten Treffer ist. Bei großen Treffermengen (wie sie bei Suchmaschinen die Regel sind) wird die Precision nur bis zu einem Cut-Off-Wert gemessen; meist werden die ersten 20 ausgegebenen Ergebnisse bewertet.
So einleuchtend die Messung der Trefferqualität mittels Precision ist, so ergeben sich doch relativ schnell Probleme; etwa, wenn man anstatt zweiwertiger Relevanzurteile eine Skala verwenden möchte, die eine zusätzliche Qualitätsabstufung möglich macht.
Das zweite klassische Retrievalmaß ist der *Recall*. Dieser wird als der Anteil der relevanten ausgegebenen Treffer an der Gesamtzahl der insgesamt vorhandenen relevanten Treffer bestimmt. Die Gesamtzahl der relevanten Treffer bezieht sich dabei auf die zugrunde liegende Datenbasis, also im klassischen Fall auf die gesamte Datenbank und im Web-Kontext auf alle im Web vorhandenen relevanten Seiten. Damit wird deutlich, dass sich der Recall nur sehr schwer messen lässt. Im Web-Kontext ist seine Messung nicht möglich beziehungsweise kann nur über Hilfsmethoden (wie etwa Pooling) erfolgen.
Weitere etablierte Retrievalmaße sind *Fallout* (der Anteil der ausgegebenen, aber nicht relevanten Treffer an der Gesamtzahl der nicht relevanten Treffer im Datenbestand) sowie *Generality* (der Anteil der relevanten Dokumente im zugrunde liegenden Datenbestand). Andere Maße wie *Expected Search Length*, *Satisfaction/Frustration* und *Success* haben den Eingang in die informationswissenschaftlichen Lehrbücher gefunden, werden jedoch in der Praxis nur selten verwendet.
Die klassischen Retrievalmaße sind nur bedingt für die Bewertung der Treffer von Web-Suchmaschinen geeignet. Dies soll beispielhaft anhand zweier neuerer Precision-Studien aus den Jahren 2004 und 2006 gezeigt werden.
Griesbaum [33] vergleicht drei deutsche Suchmaschinen (Google.de, Lycos.de und AltaVista.de) hinsichtlich der Relevanz ihrer Ergebnisse, wobei drei Trefferbewertungen möglich sind. Treffer können entweder direkt relevant oder nicht relevant sein, oder aber selbst nicht relevant sein, jedoch auf ein relevantes Dokument verweisen. Diese letzte Kategorie soll der Hypertext-Struktur des Web Rechnung tragen, indem die in ihr enthaltenen Treffer für die Auswertung den relevanten Treffern hinzugefügt werden, da sie als geeignet angesehen werden, ein Informationsbedürfnis zu befriedigen (d.h. zum Ziel zu führen). Die Konsequenz aus diesem Vorgehen ist eine höhere Precision, als sie bei der „klassischen“ Unterteilung in relevant/nicht relevant der Fall gewesen wäre.
In der Untersuchung werden 50 Suchanfragen verwendet, für die jeweils die ersten 20 Treffer jeder Suchmaschine ausgewertet werden. Die Ergebnisse zeigen, dass die Unterschiede zwischen den Suchmaschinen relativ gering sind. Google erreicht eine durchschnittliche Precision von 0,65, Lycos von 0,60 und AltaVista von 0,56.

Obwohl die Ergebnisse der Untersuchung inzwischen als überholt gelten müssen, zeigen sie doch deutlich, dass sich zumindest die großen Suchmaschinen hinsichtlich der erreichten Precision ähneln.
Bestätigt wird dieser Befund durch eine neuere Untersuchung aus dem Jahr 2006 von Véronis [34]: Auch bei Bewertung der Relevanz auf einer Skala (von 0 bis 5) liegen die führenden Suchmaschinen nahe beieinander. Die Studie vergleicht sechs Suchmaschinen, darunter die Marktführer Google, Yahoo und MSN, anhand von Anfragen aus 14 Themenbereichen. Die Auswertung zeigt keine signifikanten Unterschiede zwischen Google und Yahoo (beide erreichen im Durchschnitt 2,3), auch MSN liegt mit einem Durchschnitt von 2,0 nur wenig hinter den beiden.
Die Ähnlichkeit der Precision-Werte zumindest bei den international führenden Suchmaschinen zeigt, dass diese ähnlich gut für die Beantwortung von Standard-Suchanfragen mit informationsorientierter Ausrichtung geeignet sind.
Die nur geringen Unterschiede bei den Precisionwerten sollten allerdings nicht dahin gehend gedeutet werden, dass schlicht keine signifikanten Unterschiede mehr zwischen den einzelnen Suchmaschinen bestehen. Vielmehr ist zu fragen, ob Precision das geeignete (alleinige) Maß ist, um die Trefferqualität von Suchmaschinen zu bewerten.

Einzigartigkeit der Suchresultate

Betrachtet man zwei verschiedene Suchmaschinen, die aber auf dem gleichen Index basieren, so treten dennoch unter Umständen verschiedene Suchergebnisse auf. Die Menge der gefundenen Ergebnisse ist zu groß, als dass der Nutzer eine Auswahl treffen könnte. Es wird deswegen mit Hilfe von Ranking-Algorithmen eine sortierte Liste geliefert, die von Suchmaschine zu Suchmaschine abweicht. Der Faktor „Ranking“ ist für den Nutzer sehr wichtig [34,35]; auch Suchmaschinen, die auf dem gleichen Index basieren, sind aufgrund der unterschiedlichen Anordnung der Ergebnisse für den Nutzer oft nicht unterscheidbar (vgl. [5], S. 21-23).
Es wurden verschiedene Studien durchgeführt, die Überschneidungen von Suchergebnissen in den Ergebnislisten untersuchen [36,37,38]. In den genannten Quellen wurden nur sehr wenige Überschneidungen festgestellt, die Suchmaschinenlandschaft und die Anzahl der eigenständigen Indizes, auf die Suchmaschinen zugreifen, haben sich jedoch mittlerweile geändert.
Spink et al. [39] fanden in ihrer Studie heraus, das 84,9 Prozent von den ersten 10 Ergebnissen nur bei einer Suchmaschine zu finden ist, 11,4 Prozent bei zwei und 2,6 Prozent bei drei und nur 1,1 Prozent bei allen vier untersuchten Suchmaschinen (Ask Jeeves, Google, Yahoo, MSN). Die Autoren schließen daraus, dass der Gebrauch einer einzigen Suchmaschine nicht ausreicht. Es ist aber anzumerken, dass von jeder Suchmaschine nur die ersten zehn Resultate ausgewertet wurden. Bei einer Betrachtung beispielsweise der ersten drei Ergebnisseiten jeder Suchmaschine ergäben sich wiederum andere Zahlen. Weitere Untersuchungen zu den Überschneidungen von vorgeschlagenen Webseiten und eine detaillierte Angabe der Ergebnisse sollten durchgeführt werden.
Véronis [34] fand ebenfalls wenige Überschneidungen unter den ersten 10 Resultaten von 2,9 Prozent bis 25,1 Prozent zwischen jeweils zwei Suchmaschinen. Interessant ist, dass die größte Überschneidung zwischen Google und Yahoo zu finden war.

Qualität spezieller Suchfunktionen

Spezielle Suchfunktionen und Operatoren können bei der Websuche entscheidend dazu beitragen, dass die Ergebnisse seitens der Suchmaschine besser werden. Eine Diskussion, welche speziellen Suchfunktionen bei Suchmaschinen angeboten werden sollten und zu welchem Grad diese genutzt werden, ist bei Lewandowski [40] zu finden. Einen tabellarischen Überblick der aktuellen Suchmöglichkeiten wird in [41] gegeben.
Das Problem bei der Anwendung erweiterter Suchfunktionen ist jedoch, dass sie oft in ihrer tatsächlichen Wirkungsweise unterschätzt werden. Manche Möglichkeiten, wie die Einschränkung der Top-Level-Domain sind nicht schwierig zu verstehen und funktionieren zuverlässig. Andere, wie die Anwendung von Operatoren (AND, OR) sind durchaus komplizierter und arbeiten dennoch nicht einwandfrei [40,42,43]. Weitere Funktionen wie die Einschränkung der Sprache, Angabe von ähnlichen Webseiten und Filter wurden noch nicht ausführlich untersucht. Die Funktion, sich das Erstellungsdatum einer Webseite anzeigen zu lassen, wird von Suchmaschinen wie beispielsweise Google, Yahoo und Teoma unzureichend erfüllt. Ein Drittel der Seiten, die untersucht wurden, hatten bereits eine explizit neuere Angabe des letzten Veränderungsdatums der Seite [28]. Eine neuere Studie [31] bestätigt ebenfalls die Annahme, dass Suchmaschinen Schwierigkeiten bei der Angabe des korrekten Änderungs- oder Erstelldatums einer Webseite haben

Nutzbarkeit von Suchmaschinen

In den folgenden Abschnitten werden typische Probleme von Suchenden bei der Nutzung von Suchmaschinen aufgezeigt. Durch die Darstellung des Suchverhaltens werden Mängel bei der Nutzbarkeit von Suchmaschinen deutlich.

Design des Interface

Die heutigen Interfaces von Suchmaschinen besitzen in den meisten Fällen nur eine Dimension, es wird jedem Nutzer die gleiche Funktionalität geboten, aber es gibt nachweislich verschiedene Suchtypen, welche unterschiedliche Bedürfnisse haben [44]. Die meisten Personen betrachten das Ergebnisfenster nur sehr schnell und oberflächlich [45].
Das Google-Design ist immer noch sehr schlicht gehalten, wohingegen Yahoo als Portal auftritt und eine überladene Startseite präsentiert [46]. Die Links, die zu Werbezwecken gekauft werden, sind oft nicht klar von den sogenannten organischen Ergebnissen, die aus dem Index geliefert werden, unterscheidbar. Dass es sich um Werbung handelt, ist oft in winzigen Lettern geschrieben und die Hintergrundfarbe ist in kaum wahrnehmbaren Pastelltönen abgesetzt. Das wird ein Grund sein, warum Suchende angeben, dass sie das Gefühl haben, öfter auf gekaufte Links zu klicken [15]. Zusätzlich ist es wichtig, nur wenige Ergebnisse zu zeigen, da Suchende nicht bereit sind zu scrollen [44]. Das sichtbare Fenster besteht aber gerade wieder aus mehreren Werbelinks.
Zudem sollte eine Suchmaschine spezielle Suchfunktionen anbieten, damit Nutzer ihre Suche in gewünschter Weise eingrenzen zu können, um damit direkt Einfluss auf die Qualität der Ergebnisse zu nehmen. Jede der von Fauldrath und Kunisch [47] untersuchten Suchmaschinen (mit Ausnahme von Excite) bieten diese Services.

Akzeptanz von speziellen Suchfunktionen und Operatoren

Die Beobachtung von drei Livetickern (Fireball, Lycos und Metaspinner) über den Zeitraum von 399 bis 403 Tagen zeigte eine sehr niedrige Akzeptanz von speziellen Suchfunktionen und Operatoren [17]. Eine detaillierte Aufstellung ist in Tabelle 3 zu finden.
Suchanfragen sind kurz und das hat sich in den letzten Jahren auch nicht geändert. Deutsche Suchanfragen sind sogar etwas kürzer als englische, da im Deutschen Zusammensetzungen von Substantiven zusammengeschrieben werden und nicht wie im Englischen als einzelne Wörter (wie zum Beispiel bei christmas tree - Weihnachtsbaum). Nahezu die Hälfte der Suchanfragen enthält nur einen Term. Betrachtet man Terme, die nahezu täglich vertreten sind, dann sind darunter Füllwörter und falsch gebrauchte Operatoren zu finden [17]. Das zeigt, dass die Masse nicht weiß, wie Suchmaschinen funktionieren. Füllwörter dienen beispielsweise nicht der Einschränkung von Suchanfragen, es sei denn sie werden bewusst in der Phrasensuche genutzt („sein oder nicht sein“).
Die Nutzung von Operatoren und Phrasensuchen („Karl der Große“), die in der Literatur als komplexe Suche bezeichnet werden, lag für die untersuchten Suchmaschinen unter drei Prozent. Die Ergebnisse früherer Studien konnten so von Schmidt-Mänz und Koch [17] nicht bestätigt werden. In früheren Studien, die auf kurzen Zeiträumen basieren wurde eine häufigere Verwendung von Operatoren festgestellt.

Tabelle 3 Ergebnisse aus der Beobachtung von drei Livetickern [17]

Suchmaschine	Jahr	Tage	Anzahl Suchanfragen	Ø Länge	1-elementige Anfragen	Komplexe Anfragen	Phrasen-suche	Such-funktion
Fireball	2004	399	132.833.007	1,8	50,1%	<3,0%	2,1%	65,8%
Lycos	2004	403	189.930.859	1,7	51,9%	<3,0%	2,4%	-
Metaspinner	2004	314	4.089.731	1,8	48,4%	<3,0%	2,5%	87,9%

Die Liveticker der Suchmaschinen Fireball und Metaspinner geben zusätzlich den benutzen Suchbereich an. Diese Einschränkung ist ebenfalls als spezielle Suchfunktion zu sehen, da damit die Suche eingeschränkt werden kann. Die Ergebnisse zeigen, dass 65,8 Prozent beziehungsweise 87,9 Prozent eine Suchfunktion benutzen, wobei es sich in diesem Fall um die Einschränkung auf deutschsprachige Webseiten handelt. Aber das ist gerade die Voreinstellung der Suchmaschinen. Es werden demnach keine weiteren Einschränkungen vorgenommen, woraus geschlossen werden muss, dass spezielle Suchfunktionen oder die sogenannte erweiterte Suche nur in seltenen Fällen genutzt wird. Eine Internetumfrage, an der ca. 7000 Personen teilnahmen, ergab auch, dass Suchmaschinennutzer „ihre“ Suchmaschine nicht an ihre Bedürfnisse anpassen [15]. 76,6 Prozent gaben an, dass sie keine Personalisierungsmöglichkeiten für „ihre“ Suchmaschine einsetzen.

Zufriedenheit mit den Trefferlisten

Der größte Störfaktor bei den Ergebnislisten von Suchmaschinen sind für die Nutzer Webseiten, die nur auf ein besseres Ranking hin optimiert wurden, und solche, die nicht zu den Suchanfragen passen, die gestellt wurden, da beispielsweise Schlagwortlisten generiert wurden (21,4 Prozent von 2014 Personen [15]).

Zudem ist anzunehmen, das Suchende nicht wissen, ob sie auf organische oder auf gekaufte Links klicken. In [14] geben die befragten Personen an, dass sie unzufrieden mit angezeigten Ergebnissen sind, bei denen nicht deutlich wird, ob sie für Marketingzwecke gekauft wurden.
Die Personen gaben ebenfalls an, dass Sie finden, was sie suchen. Die Qualität des Gefundenen ist jedoch unklar. 70,8 Prozent von 6722 Personen kehren sehr häufig sofort zur Suchmaschine zurück, wenn sie auf der aufgerufenen Seite nicht gleich finden, was sie suchen. Bei [45], S. 45 werden bei 13,9 Prozent aller untersuchten Such-Sessions in der Suchmaschine Alltheweb die aufgerufenen Dokumente jeweils weniger als 30 Sekunden betrachtet.
Diese Ergebnisse zeigen, dass Suchmaschinen-Nutzer durchaus Mängel an Suchmaschinen feststellen und diese auch artikulieren können.

Benutzerführung

Suchmaschinennutzer wissen im Allgemeinen nicht, wie Suchmaschinen funktionieren [14,15]. Personen, die jedoch über die Funktionsweise Bescheid wissen, nutzen mehr Operatoren und Phrasensuchen. Durch diese Tatsache wird deutlich, wie wichtig es ist, eine klare und intuitiv bedienbare Suchmaschine bereitzustellen. Fauldrath und Kunisch [47] stellen aber fest, dass nur 57 Prozent der Suchmaschinen eine Hilfe-Seite haben, die auch leicht zu finden ist. Meistens werden diese Seiten mit „Alles über...“ oder ähnlichem betitelt. Nur 71 Prozent der untersuchten Suchmaschinen geben einen Leitfaden zur Suche.
Ein weiterer Aspekt sind zusätzliche Informationen zu den angezeigten Resultaten. Jede Suchmaschine gibt den Titel des Dokuments, eine kurze Beschreibung und die URL wieder. Weitere interessante Informationen wären die Aktualität der Seite oder wann sie in den Index aufgenommen wurde. Interessant ist auch die Angabe von ähnlichen Suchtermen, um gegebenenfalls eine weitere Suche zu starten. 71 Prozent der betrachteten Suchmaschinen geben zeitliche Informationen an und 29 Prozent schlagen ähnliche Suchterme vor [47].

Fazit

Momentan ist die tatsächliche Qualität von Suchmaschinen unbekannt. Es wurden bisher nur Studien durchgeführt, die sich mit speziellen Problembereichen von Suchmaschinen beschäftigen. Mit den aufgeführten Evaluationsmaßen und dem kurzen Abriss bisher durchgeführter Studien ist es möglich, eine umfassende Studie durchzuführen, die Suchmaschinen hinsichtlich aller wichtigen Merkmale gegenüberstellt.
Bei dem Kampf um die Marktanteile auf dem Suchmaschinenmarkt wird die Diskussion und die Vergleichbarkeit der Qualität an Bedeutung gewinnen. Diese Angaben werden dabei helfen, die Transparenz auf dem Markt zu fördern. Durch den wachsenden Druck auf Suchmaschinen können diese mit Hilfe der beschriebenen Maße Alleinstellungsmerkmale hervorheben.
Insgesamt wurde gezeigt, dass tatsächlich eine Kluft zwischen den Bedürfnissen von Suchenden und den Angeboten von Suchmaschinen besteht. Suchmaschinen-Nutzer haben noch nicht gelernt, optimal mit Suchmaschinen umzugehen, wobei es vor allem auch an einem generellen Verständnis für deren Funktionsweise fehlt. Dadurch ist es Nutzern meist nicht möglich, eine effektive und effiziente Suche durchzuführen. Zusätzlich ist die Darstellung der Suchergebnisse für den Suchenden zu verwirrend. Klare Tipps müssen offeriert werden, damit Nutzern eine geeignete Nutzungsweise herangetragen wird. Momentan erwartet der Nutzer von „seiner“ Suchmaschine viel, ist aber nicht bereit, dafür auch etwas über die Technologie von Suchmaschinen zu lernen, um von sich aus bessere und präzisere Suchanfragen zu stellen. Es ist deswegen für die Suchmaschinenbetreiber schwierig, geeignet zu reagieren. Beispielsweise müssen die Ranking-Algorithmen an die als intuitiv zu bezeichnende Benutzung angeglichen werden. Insgesamt bieten aber die vorgestellten Evaluationsmaße eine Möglichkeit, neben den technischen Grundqualitäten, die eine Suchmaschine erfüllen sollte, Probleme bei der Anwendung durch den Nutzer aufzudecken.
Suchende sollten sich eingehender mit Suchmaschinen und deren Funktionsweise beschäftigen. Auf den Hilfe-Seiten von Suchmaschinen sind oft Tipps und Anmerkungen zu den bereitgestellten Suchfunktionen zu finden. Suchanfragen können kurz formuliert werden, sollten aber einschränkende Begriffe enthalten, worunter auch das Datei-Format fällt. Oft hilft es, sich die gewünschte Seite vorzustellen, um darauf eine Suchanfrage zu starten, die alle gewünschten Wörter enthält. Sollen Texte gesucht werden, schränkt die anschließende Eingabe des Datei-Formats (zum Beispiel „pdf“ oder „ps“) die Ergebnisliste ein, da weniger Treffer angezeigt werden. Wird nach Ausdrücken oder mehreren zusammengehörigen Wörtern gesucht, dann ist die Phrasensuche, also das Setzen von Anführungszeichen um diese Worte, die beste Lösung, um die Trefferliste einzuschränken.
In der Zukunft wird basierend auf den in diesem Artikel aufgezeigten Desideraten der Forschung eine umfassende Studie durchgeführt werden, deren Ergebnisse den Nutzern dabei helfen soll, die Suchmaschine zu finden, die am besten zu ihren Bedürfnissen passt.

Literatur

1. Vaughan, L.: New Measurements for Search Engine Evaluation Proposed and Tested. Information Processing & Management 40, 677-691 (2004)
2. Ingwersen, P., Järvelin, K.: The Turn: Integration of Information Seeking and Retrieval in Context. Dordrecht: Springer 2005
3. Kleinberg, J. M.: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM 46, 604-632 (1999)
4. The PageRank Citation Ranking: Bringing order to the Web. http://dbpubs.stanford.edu:8090/pub/1999-66, 24.7.2006
5. Lewandowski, D.: Web Information Retrieval: Technologien zur Informationssuche im Internet. Frankfurt am Main: DGI 2005
6. Lewandowski, D., Höchstötter, N.: Web Searching: A Quality Measurement Perspective. In: Spink, A., Zimmer, M. (Hrsg.): Web Searching: Interdisciplinary Perspectives. Berlin, Heidelberg: Springer 2007 [im Druck]
7. Wang, H., Xie, M., Goh, T. N.: Service Quality of Internet Search Engines. Journal of Information Science 25, 499-507 (1999)
8. Xie, M., Wang, H., Goh, T. N. Quality Dimensions of Internet Search Engines. Journal of Information Science 24, 365-372 (1998)
9. Parasuraman, A., Zeithaml, V.A., Berry, L.L. SERVQUAL: A Multiple-item Scale for Measuring Consumer Perceptions of Service Quality. Journal of Retailing 64, 12-40 (1988)
10. Korfhage, R. R.: Information Storage and Retrieval. New York: Wiley 1997
11. Greisdorf, H., Spink, A.: Median Measure: an Approach to IR Systems Evaluation. Information Processing & Management 37, 843-857 (2001)
12. Su, L. T.: Value of Search Results as a Whole as the Best Single Measure of Information Retrieval Performance. Information Processing & Management 34, 557-579 (1998)
13. Lewandowski, D.: Zur Bewertung der Qualität von Suchmaschinen. In: Eberspächer, J., Holtel, St. (Hrsg.): Suchen und Finden im Internet. Berlin, Heidelberg: Springer 2006
14. Machill, M., Neuberger, Ch., Schweiger, W., Wirth, W.: Qualität und Nutzung von Suchmaschinen. In: Machill, M., Welp, C. (Hrsg.): Wegweiser im Netz. Gütersloh: Verlag Bertelsmann Stiftung 2003
15. Schmidt-Maenz, N., Bomhardt, Ch.: Wie Suchen Onliner im Internet?. Science Factory 2/2005, Absatzwirtschaft, 5-8 (2005)
16. Dujmovic, J., Bai, H.: Evaluation and Comparison of Search Engines Using the LSP Method. Computer Science and Information Systems 3, 31-56 (2006)
17. Schmidt-Maenz, N., Koch, M.: A General Classification of (Search) Queries and Terms. Proceedings of International Conference on Information Technologies: Next Generations, Las Vegas, Nevada, USA, 375-381 (2006)
18. Search Engine Sizes. http://searchenginewatch.com/showPage.html?page=2156481 24.7.2006
19. Bharat, K., Broder, A.: A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines. Computer Networks and ISDN Systems 30, 379-388 (1998)
20. Lawrence, St., and Giles, C.L.: Searching the World Wide Web. Science 280, 98-100 (1998)
21. Lawrence, St., Giles, C.L: Accessibility of Information on the web. Nature 400, 107-109 (1999)
22. Gulli, A., Signorini, A.: The Indexable Web is more than 11.5 Billion Pages. Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, Chiba, Japan, 902-903 (2005)
23. Sherman, Ch., Price, G.: The Invisible Web: Uncovering Information Sources Search Engines Can't See. Medford, NJ: Information Today 2001
24. Bergman, M. K.: The Deep Web: Surfacing Hidden Value. Journal of Electronic Publishing 7 (2001)
25. Lewandowski, D., Mayr, Ph.: Exploring the Academic Invisible Web. Library Hi Tech 24 (2006)
26. Vaughan, L., Thelwall, M.: Search Engine Coverage Bias: Evidence and Possible Causes. Information Processing & Management 40, 693-707 (2004)
27. Acharya, A., Cutts, M., Dean, J., Haahr, P., Henzinger, M., Hoelzle, U., Lawrence, St., Pfleger, K., Sercinoglu, O., Tong, S.: Information Retrieval based on Historical Data. Google Patent, USA 2005
28. Lewandowski, D.: Date-restricted Queries in Web Search Engines. Online Information Review 28, 420-427 (2004)
29. Ntoulas, A., Cho, J., Olston, Ch.: What's New on the Web? The Evolution of the Web from a Search Engine Perspective. Thirteenth WWW Conference, New York, USA 2004
30. Notess, G.: Search Engine Statistics: Freshness Showdown. URL: http://www.searchengineshowdown.com/stats/freshness.shtml
31. Lewandowski, D., Wahlig, H., Meyer-Bautor, G.: The Freshness of Web Search Engine Databases. Journal of Information Science 32, 133-150 (2006)

32. Chu, H.: Information Representation and Retrieval in the Digital Age. Medford, NJ: Information Today 2003
33. Griesbaum, J.: Evaluation of three German Search Engines: Altavista.de, Google.de and Lycos.de. Information Research 9 (2004)
34. Véronis, J.: A Comparative Study of six Search Engines. URL: http://www.up.univ-mrs.fr/veronis/pdf/2006-comparative-study.pdf
35. Spink, A., Park, M., Jansen, B. J., Pedersen, J.: Multitasking during Web Search Sessions. Information Processing & Management 42, 264-275 (2006)
36. Chignell, M. H., Gwizdka, J., Bodner, R. C.: Discriminating Meta-Search: a Framework for Evaluation. Information Processing and Management 35, 337-362 (1999)
37. Gordon, M., Pathak, P.: Finding Information on the World Wide Web: the Retrieval Effectiveness of Search Engines. Information Processing & Management 35, 141-180 (1999)
38. Schwartz, C.: Web Search Engines. Journal of the American Society for Information Science and Technology 49, 973-982 (1998)
39. Spink, A., Jansen, B. J., Blakely, Ch., Koshman, Sh.: A Study of Results Overlap and Uniqueness among Major Web Search Engines. Information Processing & Management 42, 1379-1391 (2006)
40. Lewandowski, D.: Abfragesprachen und erweiterte Suchfunktionen von WWW-Suchmaschinen. Information: Wissenschaft und Praxis 55, 97-102 (2004)
41. Notess, G.: Search Engine Features Chart. URL: http://www.searchengineshowdown.com/features/, 5.2.2007
42. Notess, G.: Boolean Searching on Google. URL: http://www.searchengineshowdown.com/features/google/googleboolean.html, 5.2.2007
43. Eastman, C., Jansen, B.: Coverage, Relevance, and Ranking: The Impact of Query Operators on Web Search Engine Results. ACM Transactions on Information Systems 21, 383-411 (2003)
44. Search Engine Usage in North America, A Research Initiative by Enquiro. www.enquiro.com, 16.3.2006
45. Spink, A., Jansen, B. J.: Web Search: Public Searching of the Web. Dordrecht, Boston, London: Kluwer Academic Publishers 2004
46. Geoghegan, T.: Search Wars - Which is Best? news.bbc.co.uk/2/hi/uk_news/magazine/4003193.stm, 2.2.2007
47. Fauldrath, J., Kunisch, A.: Kooperative Evaluation der Usability von Suchmaschineninterfaces. Information: Wissenschaft und Praxis 56, 21-28 (2005)