Dirk Lewandowski

Mit welchen Kennzahlen lässt sich die Qualität von Suchmaschinen messen?

erscheint in:
Marcel Machill / Markus Beiler (Hrsg.): Die Macht der Suchmaschinen / The Power of Search Engines . Köln: Herbert von Halem Verlag, 2007
ISBN 3-938258-33-0

Einleitung

Suchmaschinen bilden den bedeutendsten Zugang zu den im World Wide Web verfügbaren Informationen und haben andere Zugänge zu diesem Informationsbestand (wie etwa Internet-Verzeichnisse) weitgehend verdrängt. Der Suchmaschinen-Markt ist stark konzentriert (Lewandowski 2005: 21ff.; zur Verteilung der Suchanfragen siehe Webhits 2006); nur wenige Anbieter mit eigener Technologie bieten selbst Endnutzer-Lösungen an und lizenzieren ihre Technologie an die bekannten Web-Portale wie AOL oder T-Online.

Die im vorliegenden Beitrag dargestellten Kennzahlen zur Messung der Qualität von Suchmaschinen basieren einerseits auf Erkenntnissen über den *State of the Art* der Suchmaschinen-Technologie (Lewandowski 2005), andererseits stehen sie im Kontext eines umfassenderen Modells der Qualitätsmessung für Web-Suchmaschinen (Lewandowski/Schmidt-Mänz 2007).

Die besondere Bedeutung von Erkenntnissen über die Qualität der bestehenden Suchmaschinen ergibt sich einerseits aus ihrer Bedeutung für die Weiterentwicklung der Suchmaschinentechnologie. Qualitätsuntersuchungen geben Hinweise auf die Schwachstellen der Suchmaschinen im Allgemeinen und die ihrer Ranking-Algorithmen im Besonderen. Letztere sind als zentral für die Ergebnisqualität anzusehen und bilden den »Kern« der technologischen Leistungsfähigkeit einer jeden Suchmaschine.

Des Weiteren sind Qualitätsuntersuchungen aber auch als Grundlage für zurzeit laufende, wichtige Diskussionen vonnöten. Hier ist insbesondere die Diskussion um »alternative« Suchmaschinentechnologie (vorangetrieben vor allem vom »Gemeinnützigen Verein zur Förderung der Suchmaschinen-Technologie und des freien Wissenszugangs (SuMa-eV)« und durch ein Papier der Grünen Bundestagsfraktion (Neymanns 2005) und die Diskussion um eine europäische Alternative zu den großen Suchmaschinen-Anbietern wie Google zu nennen.

Die Diskussion ist von einigen immer wiederkehrenden Fragen bestimmt, die zu einem großen Teil höchstens spekulativ beantwortet werden können, da schlicht die grundlegenden Daten fehlen, um zu einer fundierten Antwort zu gelangen.

Beispielhaft sollen hier zwei wichtige, jedoch ungeklärte Fragen genannt werden:

- Werden von den Suchmaschinen (aufgrund willentlicher Entscheidungen oder struktureller Gegebenheiten des Web-Graphen) kommerzielle Ergebnisse bevorzugt?
- Liefern alle Suchmaschinen ähnliche Ergebnisse oder unterscheiden sich diese deutlich?

Diese Fragen können nur beantwortet werden, wenn die Treffer der Suchmaschinen genau untersucht werden; aus den Ergebnissen dieser Untersuchungen könnte dann entsprechend ein Handlungsbedarf abgeleitet werden.

Die Qualität der bestehenden Suchmaschinen ist also als ausschlaggebend für die Bewertung des Handlungsbedarfs anzusehen. Denkbar sind zwei extreme Szenarien:

- Einerseits könnten alle (größeren) bestehenden Suchmaschinen eine ebenbürtige Qualität bieten. Dann wäre die Frage nach den Monopolstrukturen auf dem Suchmaschinenmarkt nur noch eine nach einer gleichmäßigeren Verteilung der Suchanfragen.
- Die Qualität der bestehenden Suchmaschinen ist stark unterschiedlich. Tatsächlich ist die nach Marktanteilen führende Suchmaschine Google auch in Bezug auf die Qualität die beste. Dies wird zwar schon von den Nutzern so wahrgenommen (Machill/Neuberger/Schweiger/Wirth 2003: 187), eine Überprüfung der objektiven Ergebnisqualität steht aber noch weitgehend aus.

Vermutlich wird die reale Situation zwischen diesen beiden hypothetischen Szenarien liegen. Allerdings ist es erstaunlich, dass gerade angesichts der beschriebenen Situation nur wenige Untersuchungen zur Qualität der Suchmaschinen vorliegen, sowohl international als auch (besonders eklatant) im deutschsprachigen Raum. Den vorliegenden Untersuchungen ist außerdem gemein, dass sie nur Teilaspekte der Qualitätsmessung behandeln und damit kein umfassendes Bild der Suchmaschinen-Qualität bieten können.

Qualitätsfaktoren

Wie im Modell bei Lewandowski/Schmidt-Mänz (2007) dargelegt, kann die Qualität der Suchmaschinen nur durch eine Kombination unterschiedlicher Faktoren gemessen werden, die im Modell in vier Bereiche gefasst werden:

- Index-Qualität (*index quality*): Hier werden die den Suchmaschinen zugrunde liegenden Datenbestände untersucht. Faktoren sind unter anderem die Größe der Datenbestände, die Vollständigkeit der Erfassung bestimmter Bereiche des Web (etwa bestimmter Sprachbereiche (vgl. Vaughan/Thelwall 2004) und die Aktualität der Datenbestände (Lewandowski/Wahlig/Meyer-Bautor 2006).
- Qualität der Treffer (*quality of the results*): In diesem Bereich sind die »klassischen« Suchmaschinen-Tests angesiedelt. Neben der Messung der Zufriedenheit der Nutzer

(durch Befragungen) werden vor allem Tests zur Messung der Retrievaleffektivität eingesetzt.

- Recherche-Qualität (*quality of the search features*): Besonders für geübte Nutzer ist das Vorhandensein umfangreicher Suchfunktionen (»erweiterte Suche«) bzw. das Vorhandensein einer komplexen Abfragesprache wichtig. Der Umfang und die Qualität der von den Suchmaschinen angebotenen Funktionen unterscheiden sich deutlich (Lewandowski 2004a), teils bestehen Probleme mit der Funktionstüchtigkeit der angebotenen Funktionen (Lewandowski 2004b).
- Usability (*search engine usability*): Letztlich ist auch die Gestaltung der Benutzerführung von Bedeutung für die Qualität einer Suchmaschine.

Dieser Beitrag beschäftigt sich im Folgenden allein mit dem zweiten Bereich, also der Qualität der Treffer. Dieser bildet sicher den in der informationswissenschaftlichen Forschung am weitesten bearbeiteten Bereich. Manches Mal wurde allerdings übersehen, dass die Ergebnisse zur Trefferqualität nur im Zusammenspiel mit Ergebnissen aus den anderen Bereichen wirklich aussagekräftig sind. Während solche »Einzelbewertungen« im Kontext von klassischen Datenbanken noch angebracht sein mögen (siehe die bekannten Evaluierungsinitiativen wie TREC und CLEF), erscheint im Kontext der Web-Suchmaschinen eine kombinierte Evaluation, die am »Live«-System durchgeführt wird, als der einzig gangbare Weg.

Mit der allgemeinen Problematik der Relevanzmessung haben sich eine Vielzahl von Autoren auseinandergesetzt (Überblicksartikel u.a. Mizzaro 1997; Lancaster/Gale 2003; Jacsó 2005). Neben der Zusammenstellung der Testkollektionen sind vor allem die verschiedenen Retrievalmaße kritisch zu sehen; beides vor allem in Hinblick auf die durch besondere Probleme der Dokumentkollektion und des speziellen Nutzerverhaltens einzigartigen Web-Suchmaschinen.

Klassische Retrievalmaße

Die Frage bei der Evaluierung der Trefferqualität ist aber nun, mit welchen Maßen sich belastbare Aussagen treffen lassen. Zur Messung der Retrievaleffektivität von Information-Retrieval-Systemen allgemein 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 bzw. entsprechend weiterentwickelt, um für diesen Bereich geeignet 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 1 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 würde also einerseits *nur* relevante Treffer enthalten, andererseits *alle* relevanten Treffer (sofern sie im System enthalten sind) zurückgeben. Ein solches System würde also sowohl hinsichtlich der Präzision der Suchergebnisse als auch hinsichtlich deren Vollständigkeit das Maximum erreichen.

Tabelle 1: Mögliche Retrievalergebnisse (Chu 2003:188)

	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 einer vollständigen und vollständig relevanten Treffermenge 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 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 schon relativ schnell Probleme; etwa, wenn man anstatt zweiwertiger Relevanzurteile Skalen verwenden möchte.

Das zweite klassische Retrievalmaß ist der *Recall*. Dieser wird bestimmt als der Anteil der relevanten ausgegebenen Treffer an der Gesamtzahl der insgesamt vorhandenen relevanten Treffer. 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 bzw. kann nur über Hilfsmethoden erfolgen.[1]

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[2]). 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.

Precision-Untersuchungen der Suchmaschinen

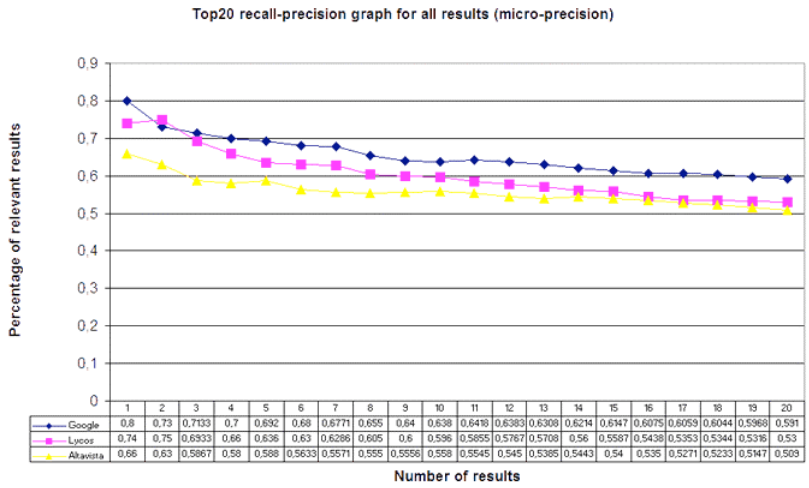
Dass die klassischen Retrievalmaße nur bedingt für die Bewertung der Treffer von Web-Suchmaschinen geeignet sind, soll beispielhaft anhand zweier neuerer Precision-Studien aus den Jahren 2004 und 2006 gezeigt werden.

Griesbaum (2004) 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 zwar selbst nicht relevant sein, jedoch auf ein relevantes Dokument verweisen. Diese letzte Kategorie soll der Hypertext-Struktur des Web Rechnung tragen und die in ihr enthaltenen Treffer werden für die Auswertung schließlich den relevanten Treffern zugerechnet, 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 *Mean Avergage Precision* von 0,65, Lycos 0,60 und AltaVista 0,56. Abbildung 1 zeigt den kompletten Recall-Precision-Graphen.

Während die Ergebnisse der Untersuchung inzwischen überholt sind[3], zeigen sie doch deutlich, dass sich zumindest die großen Suchmaschinen hinsichtlich der erreichten Precision annähern.

Abbildung 1: Precision-Graph aus der Untersuchung von Griesbaum (2004)

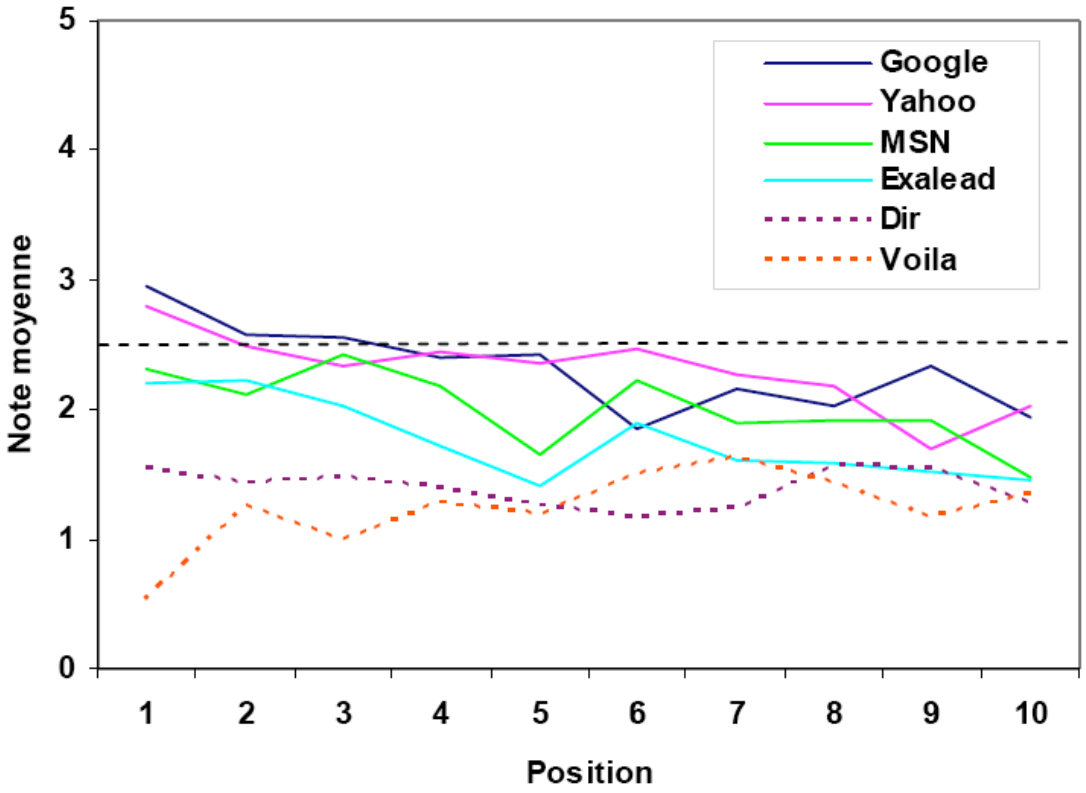


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Google	0,8	0,73	0,7133	0,7	0,692	0,68	0,6771	0,655	0,64	0,638	0,6418	0,6383	0,6308	0,6214	0,6147	0,6075	0,6059	0,6044	0,5968	0,591
Lycos	0,74	0,75	0,6933	0,66	0,636	0,63	0,6286	0,605	0,6	0,596	0,5855	0,5767	0,5708	0,56	0,5587	0,5438	0,5353	0,5344	0,5316	0,53
Altavista	0,66	0,63	0,5867	0,58	0,588	0,5633	0,5571	0,555	0,5556	0,558	0,5545	0,545	0,5385	0,5443	0,54	0,535	0,5271	0,5233	0,5147	0,509

Bestätigt wird dieser Befund durch eine neuere Untersuchung aus dem Jahr 2006 von Véronis: 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.

Abbildung 2: Precision-Graph aus der Untersuchung von Véronis (2006)



Die Annäherung der Precision-Werte zumindest bei den international führenden Suchmaschinen zeigt, dass diese ähnlich gut geeignet sind für die Beantwortung von Standard-Suchanfragen mit informationsorientierter Ausrichtung. In keiner der genannten Studien wurden zusätzlich Anfragen anderer Anfragetypen berücksichtigt.[4]
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 der Suchmaschinen zu bewerten.

Weitere Retrievalmaße

Die Beispiele der Retrievaltests haben gezeigt, dass die Notwendigkeit besteht, weitere Retrievalmaße für die Bewertung der Suchmaschinen einzusetzen. Hier sollen beispielhaft einige wichtige Retrievalmaße sowohl allgemeiner Natur als auch solche, die speziell für die Evaluierung von Web-Suchmaschinen entwickelt wurden, vorgestellt

werden. Ihnen allen gemeinsam ist, dass im Rahmen ihrer Entwicklung zwar kleinere Untersuchungen durchgeführt wurden, jedoch keine belastbaren Ergebnisse für größere Kollektionen vorliegen. Die Maße befinden sich also allesamt noch in einem experimentellen Stadium und weitere Tests im Rahmen der Suchmaschinen-Evaluierung sind zu wünschen.

Für den Kontext der Suchmaschinen-Evaluierung erscheinen vor allem die im Folgenden beschriebenen allgemeinen Retrievalmaße als vielversprechend:

- *Median Measure* (Greisdorf/Spink 2001): Hier wird berücksichtigt, wie viele Ergebnisse insgesamt vom System zurückgegeben werden. Im Gegensatz zu herkömmlichen Maßen kann dadurch zwischen Ergebnismengen mit den gleichen Precision-Werten unterschieden werden: So würden ein System, welches fünf relevanten und zehn irrelevante Dokumente ausgibt, und ein zweites, welches 50 relevante und 100 irrelevante Dokumente ausgibt, die gleiche Precision erreichen. Für den Nutzer besteht aber durchaus ein Unterschied zwischen den beiden Systemen, da sich für ihn in erster Linie die Frage stellen dürfte, durch wie viel Ballast er sich durcharbeiten muss, um zu den gewünschten relevanten Treffern zu gelangen.
- *Average Distance Measure* (Della Mea/Mizzaro 2004): Hierbei handelt es sich um einen mehr nutzerorientierten Ansatz, der die Unterschiede zwischen der Bewertung der Treffer durch das Information-Retrieval-System und der Bewertung durch den Nutzer misst. Dieses Maß erscheint für die Suchmaschinen-Evaluation vielversprechend, wurde im Web-Kontext bisher jedoch noch nicht angewendet.
- *Importance of completeness of search results; Importance of precision of the search to user* (Su 1998): In diesen beiden Maßen wird versucht, das typische Nutzerverhalten in den Evaluierungsprozess einzubinden. Dabei wird berücksichtigt, ob der Nutzer nur ein paar relevante Ergebnisse benötigt oder aber Wert auf eine vollständige Ergebnismenge legt, dafür aber eine geringere Precision in Kauf nimmt. Für die Evaluierung von Suchmaschinen sollte auf das typische Verhalten der Nutzer dieser Anwendungen (s.u.) zurückgegriffen werden.

Eine Vielzahl von Maßzahlen verwendet Su (1998; 2003; 2003) für die Bewertung der Qualität von Suchmaschinen durch die Nutzer, wobei sich die Bewertung der Relevanz auf vier Maßzahlen beschränkt. Die weiteren Maßzahlen sind hauptsächlich dem Bereich Nutzerzufriedenheit zuzuordnen und dürften stark von der untersuchten Nutzergruppe (in der präsentierten Untersuchung sind dies Studenten des Grundstudiums) abhängen. Das in den Untersuchungen entwickelte Evaluierungsmodell ist nicht suchmaschinenspezifisch; es handelt sich eher um ein allgemeines Modell für IR-Systeme, das in der Fallstudie auf Suchmaschinen angewendet wurde.

Web-spezifische Retrievalmaße

Der nur bedingten Eignung der allgemeinen Retrievalmaße für den Web-Kontext haben sich mehrere Autoren gestellt und eigene, auf diesen Kontext zugeschnittene Maße entwickelt.
Als vielversprechend für weitere Untersuchungen erscheinen insbesondere die folgenden Maße:

- *Salience* (Ding/Marchionini 1996): Bewertung der Precision einer einzelnen Suchmaschine in Relation zum Gesamtabschneiden aller Suchmaschinen. Hier wird berücksichtigt, dass sich das Niveau der Ergebnisse auch allgemein je nach der Anfrage unterscheiden kann, nicht nur zwischen den untersuchten Suchmaschinen.
- *CBC ratio* (MacCall/Cleveland 1999): Anteil der inhaltsorientierten Klicks (»content-bearing clicks«) am Gesamt der Klicks im gesamten Suchverlauf. Hier wird also gemessen, wie viele (bzw. wenige) Klicks notwendig sind, um an die relevanten Ergebnisse zu gelangen.
- *Quality of result ranking* (Vaughan 2004): Hierbei wird die Qualität des von der Suchmaschine durchgeführten Rankings dem Ranking durch menschliche Gutachter gegenübergestellt und die Übereinstimmung zwischen den beiden gemessen (ähnlich dem *Average Distance Measure* (Della Mea/Mizzaro 2004).
- *Ability to retrieve top ranked pages* (Vaughan 2004): Hierbei werden von unterschiedlichen Suchmaschinen jeweils die top gerankten Dokumente bis zu einem bestimmten Cut-off-Wert (z.B. zehn) zusammengeführt und menschlichen Gutachtern zur Bewertung vorgelegt. Dann werden die von den Menschen am besten bewerteten Dokumente ausgefiltert, wobei wieder ein Cut-Off festgelegt wird (bspw. 75 Prozent der Dokumente sollen in die Wertung eingehen). Letztlich wird für jede Suchmaschine berechnet, wie hoch der Anteil dieser Dokumente im Ergebnis ist.
- *Aktualitätsmaße* (Bar-Ilan 2004): Anteil der toten Links, Anteil neu aufgefundener Seiten, Anteil vollkommen neuer Seiten (die bisher keiner anderen Suchmaschine bekannt sind). Diese Maße können als Basis für die Entwicklung von »Einzigartigkeitsmaßen« dienen.

Abbildung des Nutzerverhaltens in Retrievaltests

Mit der Evaluierung von Suchmaschinen sollte letztlich immer gemessen werden, wie gut diese die Informationsbedürfnisse ihrer Nutzer befriedigen. Dass dazu weitergehende Untersuchungen als die alleinige Messung der Trefferqualität nötig sind, wurde bereits angesprochen. Aber auch bei einer Beschränkung auf die Bewertung der Trefferqualität selbst sind einige Punkte zu beachten, wenn die Evaluierung das Nutzerverhalten mit berücksichtigen soll.

Wie in diversen Untersuchungen (Hölscher 2002; Machill/Neuberger/Schweiger/Wirth 2003; Spink/Jansen 2004; Schmidt-Mänz/Koch 2006) festgestellt wurde, zeichnet sich das Nutzerverhalten vor allem durch die folgenden Tatsachen aus:

- Nur die Ergebnisse, die auf den vorderen Plätzen der Trefferlisten gezeigt werden, werden auch angesehen. Dies betrifft vor allem die ersten Treffer, die auf der ersten Bildschirmseite ohne Scrollen sichtbar sind, und im Weiteren die restlichen Treffer auf der ersten Ergebnisseite (mit in der Regel zehn Ergebnissen). Die Nutzer blättern selten weiter auf die folgenden Ergebnisseiten. Dieses Nutzerverhalten wird allerdings in den Evaluierungen nicht entsprechend abgebildet: Dort wird eher auf (relative) Vollständigkeit abgestellt, die aber vom Nutzer offensichtlich gar nicht gewünscht wird (bzw. aufgrund der schieren Masse an Dokumenten nicht zu bewältigen ist).
- Im Verlauf einer Such-Session werden nur wenige Dokumente angesehen. Auch hier wird wieder der Anspruch der relativen Vollständigkeit in Frage gestellt: Wenn Nutzer oft nur nach einem oder zwei passenden Dokumenten suchen und den Suchvorgang beenden, sobald diese angesehen wurden, ist die Anzahl der relevanten Ergebnisse (auch in der Top10) nicht mehr von absoluter Bedeutung. Wichtig wäre hier eher, dass möglichst alle Facetten eines Themas (s.u.) abgebildet werden, damit dem Nutzer auf jeden Fall (also auch bei mehrdeutigen Anfragen) die von ihm gewünschte Anzahl an relevanten Treffern ausgegeben wird.

Es stellt sich also die Frage, wie die Retrievalmaße an dieses besondere Nutzerverhalten angepasst werden können. Zusätzlich besteht ein generelles Evaluierungsproblem, welches nicht spezifisch für die Suchmaschinen ist, sondern für alle Information-Retrieval-Systeme gilt: Die Recherche ist als interaktiver Prozess anzusehen und nicht als ein Vorgang, der nur aus *einer* Suchanfrage und *einer* Ergebnispräsentation besteht. Bisherige Evaluierungen sind meist entweder rein systemorientiert (»Wie gut ist die objektive Qualität der Treffer?«) oder aber rein nutzerorientiert (»Welche Suchmaschine empfindet der Nutzer als die beste?«; z.B. Xie/Wang/Goh 1998; Wang/Xie/Goh 1999). Die Diskussion um diese Evaluierungsansätze stellt sie leider meist einander gegenüber, während allerdings eine Kombination aus beiden Ansätzen notwendig ist (vgl. (Saracevic 1995; Lewandowski/Schmidt-Mänz 2007).

Zusammensetzung der Trefferlisten

Die Suchmaschinen-Nutzer stellen meist kurze Suchanfragen, die aus nur einem oder wenigen Wörtern bestehen (Spink/Jansen 2004; Schmidt-Mänz/Koch 2006). Gleichzeitig erwarten sie aber, dass diese Anfragen von den Suchmaschinen befriedigend beantwortet werden können. Eine Reaktion der Suchmaschinen darauf ist der Versuch, bei mehrdeutigen Anfragen (und darum handelt es sich wohl bei den

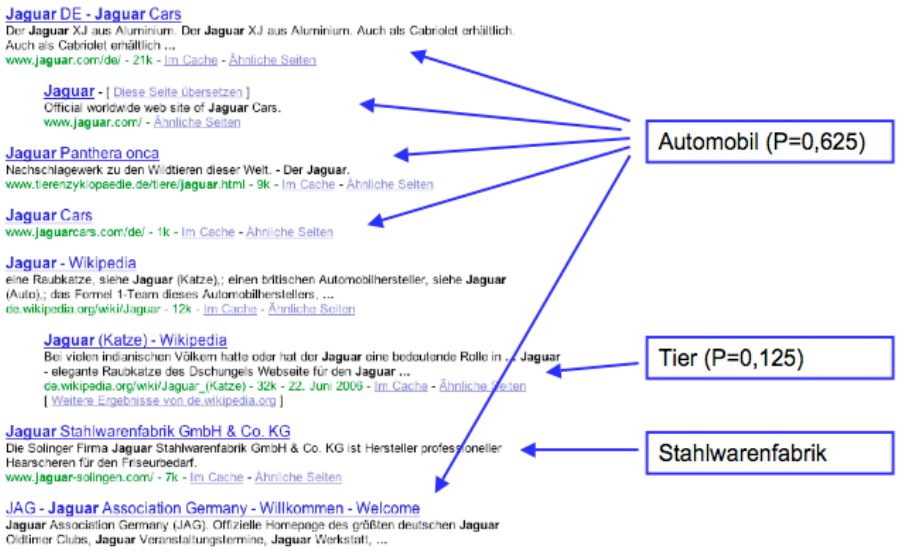
meisten Ein- und Zweiwortanfragen) eine »durchmischte« Trefferliste auszugeben, die mehrere Facetten der Suchbegriffe enthält.

Abbildung 3 zeigt einen Ausschnitt einer Trefferliste für die Suchanfrage »Jaguar«. Von den acht hier auswertbaren Treffern[5] beschäftigen sich fünf mit der Automarke, einer mit dem Tier, einer verweist auf eine Stahlwarenfabrik diesen Namens und einer verweist auf einen Film. Aus der Suchanfrage selbst ist nicht zu entnehmen, welche Bedeutung der Nutzer gemeint hat.

Wendet man nun auf diesen Fall das klassische Retrievalmaß Precision an, so fällt diese je nach der Bedeutung, nach der bewertet wird, unterschiedlich aus. Meinte der Nutzer die Automarke, ergibt sich eine Precision von 0,625, meinte er das Tier, die Stahlwarenfabrik oder den Film, liegt die Precision jeweils bei nur 0,125. In keinem Fall wird der höchste Wert erreicht; dafür wird für jede der genannten Bedeutungen zumindest ein relevantes Ergebnis zurückgegeben. Würde sich die Suchmaschine auf eine der Bedeutungen beschränken, könnte hier zwar die maximale Precision erreicht werden, für alle anderen Bedeutungen läge sie jedoch bei Null.

Bisher ist vollkommen unklar, wie mit diesem Problem in der Evaluierung verfahren werden soll. Es handelt sich nicht um ein Problem, welches nur bei einzelnen Retrievalmaßen besteht, sondern welches bei allen auftritt. Einen Ausweg könnte in der Verwendung der (für den Web-Kontext modifizierten) Maße *Importance of completeness of search results* und *Importance of precision of the search to user* (s.o.) liegen. Hier besteht allerdings noch erheblicher Forschungsbedarf.

Abbildung 3: Darstellung mehrerer Facetten in einer Trefferliste



Fazit und Ausblick

In diesem Beitrag wurde gezeigt, dass sich die Qualität von Web-Suchmaschinen einerseits nur durch eine kombinierte Qualitätsmessung bestimmen lässt, und dass sich die Qualität der Treffer nicht (allein) mit den gebräuchlichen Retrievalmaßen bestimmen lässt. Einige neu entwickelte Retrievalmaße erscheinen vielversprechend für die Anwendung auf Suchmaschinen, wurden jedoch bisher nicht in größeren Tests eingesetzt.

Neben der Weiterentwicklung der beschriebenen Maße bzw. der Neuentwicklung solcher Maße – wo nötig – ist daher der Aufbau einer Testkollektion geplant, anhand derer Retrievalmaße für Suchmaschinen getestet werden können. Auf der Basis einer Vielzahl von Anfragen soll eine Datenbank mit menschlichen Relevanzurteilen sowohl über die Trefferbeschreibungen als auch die Treffer selbst bei den unterschiedlichen Suchmaschinen aufgebaut werden. Der große Umfang und die aufwendige Begutachtung der Treffer unter verschiedenen Gesichtspunkten erfordern eine sorgfältige Planung und eine hohe Anzahl an Teilnehmern bei der Relevanzbewertung. Zurzeit werden die nötigen Voraussetzungen geschaffen, so dass die Datenbank (und damit die Möglichkeit für weitere Auswertungen) hoffentlich noch im Lauf des Jahres 2007 zur Verfügung stehen. Die auf Basis dieser Daten dann zu gewinnenden Erkenntnisse über die Trefferqualität der Suchmaschinen sollen im Rahmen eines umfassenden Untersuchungsprogramms zur Qualität von Web-Suchmaschinen eingebunden werden.

Literatur

Bar-Ilan, J.: Search Engine Ability to Cope With the Changing Web. In: Levene, M; A. Poulovassilis (Hrsg.): *Web Dynamics: Adapting to Change in Content, Size, Topology and Use*. Heidelberg [Springer Verlag] 2004, S. 195-215

Broder, A.: A taxonomy of web search. In: *SIGIR Forum,* 2, 2002: 3-10

Chu, H.: *Information Representation and Retrieval in the Digital Age*. Medford, NJ [Information Today] 2003

Della Mea, V.; S. Mizzaro: Measuring Retrieval Effectiveness: A New Proposal and a First Experimental Validation. In: *Journal of the American Society for Information Science and Technology,* 6, 2004, S. 530-543

Ding, W.; G. Marchionini: A comparative study of web search service performance. In: Hardin, S. (Hrsg.): *Proceedings of the 59th American Society for Information Science Annual Meeting*. Medford, NJ [Information Today] 1996, S. 136-142

Greisdorf, H.; A. Spink: Median measure: an approach to IR systems evaluation. In: *Information Processing & Management,* 6, 2001, S. 843-857

Griesbaum, J.: Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de. In: *Information Research,* 4, 2004

Hölscher, C.: *Die Rolle des Wissens im Internet. Gezielt suchen und kompetent auswählen*. Stuttgart [Klett-Cotta] 2002

Jacsó, P.: Relevance in the eye of the search software. In: *Online Information Review*, 6, 2005, S. 676-682

Lancaster, F. W.; V. Gale: Pertinence and Relevance. In: Drake, M. A. (Hrsg.): *Encyclopedia of Library and Information Science*. New York [Dekker] 2003, S. 2307-2316

Lewandowski, D.: Abfragesprachen und erweiterte Suchfunktionen von WWW-Suchmaschinen. In: *Information Wissenschaft und Praxis,* 2, 2004, S. 97-102

Lewandowski, D.: Date-restricted queries in web search engines. In: *Online Information Review,* 6, 2004, S. 420-427

Lewandowski, D.: *Web Information Retrieval: Technologien zur Informationssuche im Internet*. Frankfurt/M. [Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis] 2005

Lewandowski, D.; N. Schmidt-Mänz: Web Searching: A Quality Measurement Perspective. In: Spink, A; M. Zimmer (Hrsg.): *Web Searching: Interdisciplinary Perspectives*. Heidelberg [Springer] 2007 (im Druck)

Lewandowski, D.; H. Wahlig; G. Meyer-Bautor: The Freshness of Web search engine databases. In: *Journal of Information Science,* 2, 2006, S. 133-150

MacCall, S. L.; A. D. Cleveland: A Relevance-based Quantitative Measure for Internet Information Retrieval Evaluation. In: Woods, L. (Hrsg.): *Proceedings of the American Society for Information Science Annual Meeting*. Medford, NJ [Information Today] 1999, S. 763-768

Machill, M.; C. Neuberger; W. Schweiger; W. Wirth: Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen. In: Machill, M; C. Welp (Hrsg.): *Wegweiser im Netz*. Gütersloh [Bertelsmann Stiftung] 2003, S. 13-490

Mizzaro, S.: Relevance: The Whole History. In: *Journal of the American Society for Information Science*, 9, 1997, S. 810-832

Neymanns, H.: *Suchmaschinen: Das Tor zum Netz*. Online: http://www.gruene-bundestag.de/cms/publikationen/dokbin/63/63265. 2005

Saracevic, T: Evaluation of evaluation in information retrieval. In: Fox, E.A. (Hrsg.): *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY [ACM Press] 1995, S. 138-146

Schmidt-Mänz, N.; M. Koch: *A General Classification of (Search) Queries and Terms.* 3rd International Conference on Information Technologies: Next Generations, Las Vegas, NV, USA, 2006

Spink, A.; B. J. Jansen: *Web Search: Public Searching of the Web*. Dordrecht [Kluwer Academic Publishers] 2004

Su, L. T.: Value of Search Results as a Whole as the Best Single Measure of Information Retrieval Performance. In: *Information Processing & Management*, 5, 1998, S. 557-579

Su, L. T. : A comprehensive and systematic model of user evaluation of web search engines: I. theory and background. In: *Journal of the American Society for Information Science and Technology*, 13, 2003, S. 1175-1192

Su, L. T.: A comprehensive and systematic model of user evaluation of web search engines: II: an evaluation by undergraduates. In: *Journal of the American Society for Information Science and Technology*, 13, 2003, S. 1193-1223

Vaughan, L.: New Measurements for Search Engine Evaluation Proposed and Tested. In: *Information Processing & Management*, 4, 2004, S. 677-691

Vaughan, L.; M. Thelwall: Search Engine Coverage Bias: Evidence and Possible Causes. In: *Information Processing & Management*, 4, 2004, S. 693-707

Véronis, J.: *A comparative study of six search engines*. http://www.up.univ-mrs.fr/veronis/pdf/2006-comparative-study.pdf. 2006

Wang, H.; M. Xie; T. N. Goh: Service quality of internet search engines. In: *Journal of Information Science*, 6, 1999, S. 499-507

Webhits: *Webhits Web-Barometer*. http://www.webhits.de/deutsch/index.shtml?webstats.html. 2006

Xie, M.; H. Wang; T. N. Goh: Quality dimensions of Internet search engines. *Journal of Information Science*, 5, 1998, S. 365-372

[1] Hier wird mit einem Verfahren namens *Pooling* gearbeitet: Die Ergebnisse aller untersuchten Suchmaschinen werden zu einem Pool zusammengefasst und diese Menge wird im Folgenden als das vollständige Ergebnisset, welches von jedem einzelnen System hätte erreicht werden können, angesehen. Der Recall bemisst sich als der Anteil der von einem System ausgegebenen Treffer an der Gesamtzahl der Treffer im Pool.

[2] Generality trifft damit eine Aussage über den Datenbestand selbst und nicht über die ausgegebenen Treffer. Es ist aber einleuchtend, dass es für ein Information-Retrieval-System die Ausgabe von relevanten Treffern ungleich schwieriger ist, wenn im Datenbestand von Vornherein nur wenige relevante Treffer vorhanden sind (und damit der Anteil des Ballasts sehr hoch liegt). Generality versucht, diesem Umstand Rechnung zu tragen.

[3] Dies ist ein generelles Problem bei der Untersuchung von Suchmaschinen bzw. bei der Untersuchung von »Live«-Systemen allgemein. Die Ergebnisse sind schon zum Zeitpunkt der Veröffentlichung veraltet und stellen nur eine Momentaufnahme dar. Nichtsdestotrotz können solche Untersuchungen Trends aufzeigen und insbesondere bei mehrmaliger Wiederholung in gewissen Zeitabständen zu belastbaren Aussagen führen.

[4] Bei den an Suchmaschinen gestellten Anfragen wird allgemein zwischen informationsorientierten, navigationsorientierten und transaktionsorientierten Anfragen unterschieden (Unterteilung nach Broder (2002).

[5] Der fünfte Treffer stellt eine Verweisseite auf unterschiedliche Bedeutungen des Begriffs dar.