Suchmaschinenforschung im Kontext einer zukünftigen Webwissenschaft[1]

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

1. Einleitung

Die Diskussion um eine eigenständige Webwissenschaft wurde vor allem durch einen Beitrag von Tim Berners-Lee und Kollegen (Berners-Lee et al. 2006a); ausführlicher in Berners-Lee et al. (2006b) angestoßen. Ziel dieses Beitrags ist die Darstellung der Themenfelder der Suchmaschinenforschung und ihrer Einordnung in den Kontext einer umfassenderen Webwissenschaft.

Suchmaschinen beziehen ihre Bedeutung aus der Tatsache, dass sie bei der Suche im Web nicht nur das bevorzugte Werkzeug darstellen, sondern sich ihre Stellung mittlerweile so weit ausgebaut hat, dass sie nahezu das einzige vom Nutzer verwendete Werkzeug zur Suche im Netz darstellen. Die Suche selbst bezieht ihre Bedeutung durch die Selbstverständlichkeit, die sie mittlerweile erlangt hat. Man kann sagen: Wer im Netz ist, der sucht auch. So ist es nicht verwunderlich, dass die Suche nach dem Schreiben von E-Mails der meistgenutzte Dienst im Web ist (Van Eimeren and Frees 2007).

Aus diesen Gründen muss die Frage geklärt werden, ob Suchmaschinen hinsichtlich Technik und Benutzerfreundlichkeit geeignet sind, qualitativ hochwertige Inhalte und auch hochspezialisierte Informationen zu finden. Dabei ist zu fragen, welche aktuellen und zukünftigen Entwicklungen für die Suchqualität von Bedeutung sind, wenn Suchmaschinen die entscheidende Schnittstelle zum unüberschaubaren Webangebots bilden.

Dieser Aufsatz bemüht sich darum, zu den jeweiligen Forschungsgebieten und –fragen nicht nur die maßgeblichen Forschungsaufsätze zu zitieren, sondern auch Überblickswerke, die es der Leserin/dem Leser möglich machen sollen, sich schnell in ein Themengebiet von besonderem Interesse einzuarbeiten.

2. Suchlösungen für das Web

Die Technik der Suchmaschinen ist nicht der einzige Ansatz, um eine Suche im Web zu ermöglichen. Unter Suchmaschinen wird hier die algorithmische Suchlösung verstanden, das heißt, Dokumente werden automatisch mittels Crawlern erfasst, automatisch erschlossen und auf eine Nutzeranfrage hin werden passende Dokumente – nach der angenommenen Relevanz sortiert – zurückgegeben.

Weitere Ansätze der Suche im Web sind die Kataloge (auch: Verzeichnisse), Suchlösungen zur Faktenextraktion, Frage-Antwort-Dienste und die Soziale Suche (social search).

In Katalogen werden Websites manuell erfasst, mit einer kurzen Beschreibung versehen und in ein Klassifikationssystem einsortiert. In manchen Fällen erfolgt eine zusätzliche Erschließung mittels Schlagwörtern. Der Ansatz der Kataloge gilt – zumindest für die nicht fachspezifische Erschließung des gesamten Web – als gescheitert. Die Gründe liegen zum einen in der schieren Masse der Inhalte (und den damit verbundenen Kosten für eine manuelle Erschließung), auf der anderen Seite in den

[1] Erscheint in: Scherfer, Konrad: Theorie und Praxis des Webs. Grundüberlegungen einer zukünftigen Webwissenschaft . - Münster: 2008

Inkonsistenzen der verwendeten Klassifikationen (vgl. Stock and Stock 2000), die eine Orientierung erschweren.
Im Bereich der Faktenextraktion (s. Berners-Lee et al. 2006b, S. 18) wird versucht, auf Faktenfragen direkte Antworten zu generieren. Der wichtigste Gedanke bei diesem Ansatz ist, dass der Nutzer in vielen Fällen keine Menge von Dokumenten als Ergebnis auf seine Suchanfrage erwartet, sondern eine einfache Antwort, zum Beispiel eine Zahl.
Frage-Antwort-Dienste versuchen ebenso, direkte Antworten auf Fragen zu geben, allerdings basieren sie auf einer Nutzergemeinschaft, die die gestellten Fragen individuell beantwortet. Während also bei der Faktenextraktion die bereits bestehende Informationsbasis des WWW ausgenutzt wird, um Antworten zu generieren, bauen Frage-Antwort-Dienste eine eigene Wissensbasis auf. Der Schwerpunkt dieser Dienste liegt jedoch (zumindest bisher) nicht auf einem bestehenden Archiv von bereits gestellten Fragen, sondern auf neuen Fragen, die von der Gemeinschaft beantwortet werden. Im Gegensatz zu allen anderen Suchdiensten liegt also bei den Frage-Antwort-Diensten in den meisten Fällen eine nennenswerte Zeit zwischen dem Stellen einer Anfrage und dem Ergebnis.
Im Kontext Web 2.0 haben sich Soziale Suchdienste herausgebildet (Graefe, Maaß, and Heß 2007). Diesen ist gemein, dass sie nicht als Suchdienste im eigentlichen Sinn angelegt sind, sondern primär Werkzeuge zur Verwaltung von Lesezeichen (bookmarks) darstellen. Im Gegensatz zu der bekannten Lesezeichenverwaltung im Browser werden die Lesezeichen jedoch mit anderen Nutzern geteilt und es werden *tags* (freie Stich- oder Schlagwörter) zu diesen vergeben. Die Suche erfolgt in den tags, eine Verbindung mit einer Volltextsuche (wie von den algorithmischen Suchmaschinen bekannt) erfolgt zumindest bisher nicht. Für die Suche ausnutzen lässt sich die Häufigkeitsverteilung innerhalb der vergebenen tags, so dass auch hier – trotz der geringen Textmenge zu den einzelnen Bookmarks – ein sinnvolles Relevance Ranking möglich wird.
Der Schwerpunkt dieses Artikels liegt jedoch auf den algorithmischen Suchmaschinen, die die Suche im Web (zumindest bisher) stark dominieren. Hier ist zu unterscheiden zwischen Universalsuchmaschinen (keine bewusste Einschränkung nach Region oder Thema), Spezialsuchmaschinen und Archivsuchmaschinen (Archivierung des Web) (Lewandowski 2005, S. 24). Der Schwerpunkt dieses Texts liegt auf den Universalsuchmaschinen, die als Vorreiter für die Spezialsuchmaschinen von besonderer Bedeutung sind. Bei den Archivsuchmaschinen handelt es sich um eine noch wenig entwickelte Form, die jedoch insbesondere im Zuge von auf das Web ausgedehnten Sammelaufträgen der Nationalbibliotheken an Bedeutung gewinnen dürften.

3. Felder der Suchmaschinenforschung

Machill, Beiler und Zenker (2007) teilen die Suchmaschinenforschung in vier Themenfelder ein: Suchmaschinenökonomie, Suchmaschinen und Journalismus, Technik und Qualität von Suchmaschinen, Nutzerverhalten und –kompetenz. Auffällig dabei ist vor allem, dass der journalistische Bereich eigens als Themenfeld auftaucht, andererseits aber die Bereiche Technik und Qualität zusammengefasst werden. Ersteres lässt sich leicht aus der Herkunft der Autoren aus der Journalistik/Kommunikationswissenschaft erklären, die Trennung von Technik und Qualität erscheint jedoch problematisch. Denn während die Forschung zur Suchmaschinentechnik durchaus breit angelegt ist und stark vorangetrieben wird, ist der Bereich der Qualitätsmessung nur unzureichend ausgebaut, vor allem in Hinblick auf ein Monitoring der bekannten, allgemein genutzten Suchmaschinen (was gerade aufgrund ihrer Dominanz als Informationsbeschaffungswerkzeuge für jedermann zu wünschen wäre). Die Trennung von Technik und Qualitätsmessung erscheint auch insofern vonnöten, da beide ja gerade nicht Hand in Hand gehen, sondern die Qualitätsmessung in gewisser Weise konträr zu den technischen Entwicklungen steht. Es sollte ja eben nicht derjenige ein System evaluieren, das er selbst erstellt hat.

Insofern wird hier eine andere Systematisierung vorgeschlagen, das aus den folgenden Bereichen besteht:

- Information-Retrieval-Technologie
- Qualität von Suchmaschinen
- Information Research
- Nutzerverhalten und Benutzerführung
- Suchmaschinenökonomie

Im weiteren Text sollen diese Bereiche mit den bereits erreichten Erfolgen, aber auch mit den Desideraten der Forschung vorgestellt werden.

3.1. Information-Retrieval-Technologie

Die Entwicklung von Suchmaschinen ist abhängig von der Entwicklung der zugrunde liegenden Verfahren, insbesondere der Verfahren des Information Retrieval (Ferber 2003; Stock 2007). Deren Anwendung auf das Web wird im sog. Web Information Retrieval (Lewandowski 2005a) behandelt. Besondere Anwendung des Web Information Retrieval ist unter anderem das Crawling, also die automatische Erfassung der Inhalte des Web.

Ein Hauptaugenmerk der Information-Retrieval-Verfahren für das Web wird auf die Entwicklung von Rankingfaktoren für die Anordnung der Trefferliste gelegt. Dabei sind vier große Gruppen zu unterscheiden:

- Textstatistische Verfahren (Vergleich Anfrage – Dokument)
- popularitätsmessende Verfahren (linktopologische Verfahren wie PageRank, s. (Lewandowski 2005a, S. 117ff.)
- aktualitätsmessende Verfahren (Acharya et al. 2005)
- sprach- und geolokalisierende Verfahren

Für die Zukunft wird der verstärkte Einsatz semantischer Verfahren erwartet, wenn auch noch unklar ist, wie sich die entwickelten Verfahren des „Semantic Web“ tatsächlich für den Kontext des gesamten, in seiner Struktur und seinen Inhalten divergenten World Wide Web nutzen lassen. Als besonders spannend ist hier die Zusammenführung semantischer Verfahren mit den kollaborativen Ansätzen des Web 2.0 zu sehen. Die durch kollaborative Verfahren gewonnenen Daten (zum Beispiel tags, die für URLs in Social Bookmarking Diensten vergeben werden) lassen sich für das Ranking der Suchtreffer ausnutzen. Inwieweit eine solche Einbindung zu einer Qualitätsverbesserung der Ergebnisse führt, ist aber noch unklar.

3.2. Qualität von Suchmaschinen

Zur Qualität von Suchmaschinen existiert umfangreiche Literatur (ein Überblick findet sich in Lewandowski und Höchstötter 2007), allerdings handelt es sich leider eher um eine Ansammlung von Untersuchungen zu Einzelaspekten als um eine systematische Erschließung des Themenfeldes. Lewandowski und Höchstötter (2008) stellen einen Rahmen für eine solche systematische Untersuchung der Suchmaschinenqualität vor, der sich in vier Bereiche gliedert, auf die im Folgenden eingegangen werden soll.

3.2.1. Qualität des Index

Die Qualität des Index einer Suchmaschine bestimmt sich zuerst durch seine Größe. Suchmaschinen, die einen größeren Teil des Web abdecken, können dem Nutzer Dokumente bieten, die bei anderen Suchmaschinen nicht vorhanden sind. Allerdings werden bei den wenigsten Anfragen tatsächlich besonders „rare“ Dokumente benötigt. Die Indexgrößen der bekannten Suchmaschinen liegen im

zweistelligen Milliardenbereich (vgl. Fox 2007; Mayer 2005), die Zahlen werden von den Suchmaschinenbetreibern aber in der Regel nicht mehr veröffentlicht.
Trotz der enormen Menge der im Web verfügbaren Informationen decken Suchmaschinen diese nicht vollständig ab. Gegen eine vollständige Abdeckung sprechen strukturelle, technische und monetäre Gründe. In den Bereich der strukturellen Gründe fällt auch das sog. Invisible Web (Sherman and Price 2001), das zu einem wesentlichen Teil aus Datenbanken besteht, die zwar über das Web abgefragt werden können, jedoch nicht direkt im Web – und damit nicht für die Suchmaschinen – verfügbar sind.
Während die Größe des Invisible Web in der Vergangenheit wesentlich zu hoch angesetzt wurde (Bergman 2001; zur Berechnungsproblematik s. Lewandowski und Mayr 2006), kann man nun davon ausgehen, dass die Größe des Invisible Web (gemessen anhand der Zahl der Dokumente) im zweistelligen bis niedrig dreistelligen Milliardenbereich liegt.
Zumindest in indirektem Zusammenhang mit der Größe der Indizes steht die Problematik der Aktualität. Das Web befindet sich in stetem Fluss; neue Seiten werden hinzugefügt, alte gelöscht oder aktualisiert. Keine Suchmaschine ist in der Lage, alle ihr bekannten Dokumente stets aktuell zu halten. Daher sind (automatisierte) Entscheidungen zu treffen, welche Dokumente in welchem Rhythmus zu aktualisieren sind. Dass diese Entscheidungen zu unsystematisch sind und daher ein großer Teil selbst häufig aktualisierter Dokumente von den populären Suchmaschinen nicht zeitnah erfasst wird, zeigen die Untersuchungen von Lewandowski, Wahlig, und Meyer-Bautor (2006) und Lewandowski (2008a).
Die genannten Beschränkungen der einzelnen Suchmaschinen führen zu relativ geringen Überschneidungen zwischen den großen Suchmaschinen (Gulli und Signorini 2005).

3.2.2. Qualität der Suchresultate

Die Qualität der Suchresultate kann als der Kernbereich der Qualitätsmessung bei Suchmaschinen angesehen werden, wenn auch die weiteren in diesem Artikel genannten Faktoren eine nicht zu unterschätzende Rolle spielen. Zur Messung der Qualität der Resultate werden Verfahren der Relevanzmessung aus dem klassischen Information Retrieval auf die Besonderheiten des Web Information Retrieval (Lewandowski 2005a, Lewandowski 2005b) adaptiert. Die Tests zeigen, dass zwar hinsichtlich der Trefferqualität verschiedene Gruppen von Suchmaschinen zu unterscheiden sind, allerdings keine Suchmaschine für jede Anfrage die besten Treffer liefert. Neuere Untersuchungen (Griesbaum 2004; Lewandowski 2008c; Véronis 2006) zeigen, dass sich die Suchmaschinen annähern und ihre Ergebnisse nicht wesentlich verbessern können. Allerdings ist hier auch zu beachten, dass das Erreichen bestimmter Relevanzgrenzen auch mit der Methodik der Messungen zusammenhängen kann (Lewandowski 2007).
Bei einem Vergleich der Relevanz der Suchergebnisse verschiedener Suchmaschinen sind auch stets die Überschneidungen der Suchmaschinen auf den vorderen Trefferplätzen zu beachten. So ist es möglich, dass zwei Suchmaschinen auf den ersten Plätzen vollkommen unterschiedliche Treffer liefern, diese jedoch alle relevant zur gestellten Suchanfrage sind. Untersuchungen zeigen, dass die Überschneidungen auf der ersten Trefferseite (d.i. 10 Treffer) gering sind (Spink et al. 2006); allerdings ist der übliche Vergleich der Überschneidungen auf Basis eines reinen URL-Vergleichs methodisch problematisch, was weitere Untersuchungen dringend wünschenswert macht.

3.2.3. Qualität der Suchfunktionen

Neben der einfachen Suche mit einem einzigen Suchschlitz, in den in der Regel nur wenige Wörter pro Anfrage eingegeben werden, bieten Suchmaschinen meist eine erweiterte Suche, die eine gezielte Recherche ermöglichen soll. Neben dem Umfang der Suchfunktionen (Lewandowski 2004a) ist auch deren Zuverlässigkeit von Bedeutung. Erste Untersuchungen potentiell problembehafteter Funktionen

(Datumsbeschränkung, Spracheinschränkungen) zeigen ein schlechtes Abschneiden einzelner oder gar aller großer Suchmaschinen (Lewandowski 2004b, Lewandowski 2008b).

3.2.4. Benutzerfreundlichkeit und Benutzerführung

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 (Hotchkiss, Garrison, und Jensen 2004). Die meisten Personen betrachten das Ergebnisfenster nur sehr schnell und oberflächlich (Spink und Jansen 2004).

Die Links, die zu Werbezwecken gekauft werden, sind oft nicht klar von den sog. 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 (Schmidt-Maenz und Bomhardt 2005). Zusätzlich ist es wichtig, nur wenige Ergebnisse zu zeigen, da Suchende nicht bereit sind zu scrollen (Hotchkiss, Garrison, und Jensen 2004). Das sichtbare Fenster besteht aber gerade wieder aus mehreren Werbelinks.

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 (Schmidt-Maenz und Bomhardt 2005). Zudem ist anzunehmen, das Suchende nicht wissen, ob sie auf organische oder auf gekaufte Links klicken. In der Untersuchung von Machill et al. (2003) geben die befragten Personen an, dass sie unzufrieden mit angezeigten Ergebnissen sind, bei denen nicht deutlich wird, ob sie für Marketingzwecke gekauft wurden.

Suchmaschinennutzer wissen im Allgemeinen nicht, wie Suchmaschinen funktionieren (Machill et al. 2003; Schmidt-Maenz und Bomhardt 2005). 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.

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.

Hinsichtlich der Benutzerführung und der Benutzerfreundlichkeit liegen die Aufgaben einer Suchmaschinenforschung einerseits in der Verbesserung der Nutzerführung durch die Gestaltung von intuitiv bedienbaren Interfaces bzw. benutzerleitenden Verfahren, auf der anderen Seite in der Beobachtung der gängigen Suchmaschinen und der daraus resultierenden Aufklärung über Benutzer(um)leitung auf beispielsweise werbliche Treffer.

3.3. Information Research

Im Bereich Information Research geht es auf der einen Seite darum, dem Nutzer die möglichst optimalen Recherchewerkzeuge an die Hand zu geben (siehe oben), auf der anderen Seite geht es um die Schulung der Nutzer in Recherchekompetenz, d.h. um die Ausbildung zu Rechercheprofis. Die Suchmaschinen suggerieren zwar, dass jeder Nutzer in ihnen ohne großen Aufwand nicht nur recherchieren, sondern tatsächlich zu den bestmöglichen Ergebnissen gelangen kann, in der Praxis zeigt sich jedoch schnell, dass durchschnittliche (und auch fortgeschrittene) Nutzer schon an relativ einfachen Suchaufgaben scheitern.

Für den Profi-Rechercheur steht nicht allein die Relevanz der Ergebnisse (d.h. vereinfacht: die Genauigkeit der Übereinstimmung zwischen Suchanfrage und Ergebnis) im Vordergrund, sondern vor allem deren Validität. Die Rankingverfahren der Suchmaschinen bewerten die Qualität der Dokumente vor allem aufgrund der Verlinkung oder anderer Popularitätsfaktoren, wobei die Richtigkeit der Treffer nicht bewertet wird. Hier ist noch ein enormer Forschungsbedarf zu sehen: bislang ist vollkommen unklar, inwieweit die Treffer der Suchmaschinen valide Ergebnisse liefern. Genannt sei hier beispielhaft nur die Problematik der Wikipedia-Treffer: Diese erfüllen in der Regel alle Faktoren der Relevanzbewertung (v.a. textuelle Übereinstimmung, Popularität, Aktualisierungsfrequenz), über die Validität der Inhalte wird jedoch heftig diskutiert.

3.4. Nutzerverhalten

Das Verhalten der Suchmaschinennutzer ist ein relativ gut erforschter Bereich. Bei den Methoden ist hier zwischen Befragungen, Laboruntersuchungen, Logfile-Untersuchungen und der Auswertung von Live-Tickern zu unterscheiden (Höchstötter 2007):

- In Befragungen werden Nutzer mittels (Online-)Fragebögen nach ihrem Rechercheverhalten befragt. Da repräsentative Befragungen sehr aufwendig sind (die einzige so durchgeführte Befragung deutscher Nutzer ist die von Machill et al. 2003), wird auf Online-Erhebungen ausgewichen (Schmidt-Maenz und Bomhardt 2005), die hinsichtlich der Teilnehmerstruktur eine starke Verzerrung aufweisen.
- In Laboruntersuchungen können Nutzer genau beobachtet werden; der Nachteil liegt allerdings darin, dass ein erwartungskonformes Verhalten bei dieser Form der Untersuchung wahrscheinlich ist. Eine Laboruntersuchung zum Verhalten deutscher Nutzer findet sich in (Machill et al. 2003). Als zusätzliches Instrument in Laboruntersuchungen kann Eye-tracking eingesetzt werden, wodurch eine genaue Beobachtung des Blickverhaltens möglich wird (s. Granka, Joachims, und Gay 2004).
- Eine Möglichkeit, das Nutzerverhalten auszuwerten, ohne dass die Nutzer selbst etwas davon bemerken, ist die Logfile-Analyse. Dabei werden automatisch von den Suchmaschinen erhobene Protokolldaten verwendet, so dass potentiell sehr große Datenmengen ausgewertet werden können. Beispiele solcher Untersuchungen sind Beitzel et al. (2004) und Beitzel et al. (2007). Problematisch sind die oft für diesen Untersuchungstyp recht kleinen Datenmengen aufgrund der Beschränkungen der Suchmaschinenbetreiber, die diese Daten zur Verfügung stellen müssen); so werden oft nur die Daten eines Tages ausgewertet (etwa in den Untersuchungen von Spink et al. (s. Spink und Jansen 2004), was zu Verzerrungen führen kann). Eine Übersicht zu Methodik und Ergebnissen von Logfile-Untersuchungen findet sich in Jansen und Spink (2006).
- Liveticker-Untersuchungen (Schmidt-Mänz 2007) ähneln den Logfile-Analysen insofern, dass auch sie große Datenmengen auswerten, allerdings sind sie für die Datenerhebung nicht auf die Kooperation mit den Suchmaschinenanbietern angewiesen. Vielmehr greifen sie die Daten über Ticker ab, die die momentan an eine Suchmaschine gestellten Anfragen anzeigen. Dadurch können aber nur die Anfragen selbst ohne weitere Angaben erfasst werden.

Forschungsbedarf bei der Untersuchung des Nutzerverhaltens besteht vor allem in einer differenzierteren Betrachtung unterschiedlicher Nutzergruppen und der Berücksichtigung ihrer Wünsche bei der Entwicklung der Suchmaschinen. Auch die Unterschiede des Suchverhaltens zwischen Nutzern in unterschiedlichen Ländern bzw. Regionen ist noch weitgehend unerforscht. Als dritter Bereich ist die Auswertung des (über die Suchmaschinen hinausgehenden) Navigationsverhaltens zu sehen, welches Rückschlüsse auf die Interessen und Bedürfnisse der Nutzer zulässt und sich damit für eine Personalisierung der Suchergebnisse nutzen lässt (Riemer und Brüggemann 2007).

3.5. Suchmaschinenökonomie

Im Bereich der Suchmaschinenökonomie spielt einerseits die Analyse des Suchmaschinenmarkts mit seinen Konzentrations- und Monopoltendenzen eine Rolle, auf der anderen Seite geht es um die zugrunde liegenden Geschäftsmodelle, vor allem also um die Werbefinanzierung. Daran knüpft die Analyse des Feldes der Suchmaschinenoptimierung an, also der Branche, die als Dienstleister für Unternehmen deren Websites so optimiert, dass diese in den Rankings der Suchmaschinen auf den vorderen Plätzen erscheinen.

Der Suchmaschinenmarkt zeichnet sich durch starke Konzentration aus. Zwar gibt es vordergründig eine Vielfalt bei den Anbietern, vor allem aber die großen Portale kaufen ihre Suchergebnisse bei einer der großen Suchmaschinen zu. Die aktuellen Marktanteile für Deutschland finden sich bei Webhits (2007); allerdings ist die Ermittlungsmethode kritikwürdig. Insofern ist eine valide, dauerhafte Beobachtung des deutschen Suchmarkts nach wie vor ein Desiderat der Forschung. Eine (leider schon etwas ältere) Analyse des deutschen Suchmarkts findet sich bei Karzauninkat (2003), die internationale Entwicklung (unter wirtschaftlich-historischen Gesichtspunkten) wird von Van Couvering (2008) dargestellt.

Der Verkauf von kontextbezogenen Textanzeigen stellt das Hauptgeschäftsfeld der Suchmaschinen dar. Da die Anzeigen ebenso wie die organischen Treffer nach Relevanz (in Verbindung mit dem für die Anzeige gebotenen Betrag) geordnet sind, ergeben sich oftmals werbliche Ergebnisse, die für den Nutzer relevant sein können (Jansen 2007). Inwieweit den Nutzern allerdings klar ist, wann sie bezahlte Einträge anklicken und wann es sich um reguläre Treffer handelt, ist bis auf einige kleinere Untersuchungen (Marable 2003) noch unbekannt. Diese Frage ist allerdings als Kernproblem der Textanzeigen in Suchmaschinen zu bewerten, da bei einer verbreiteten Nicht-Erkennung dieser Anzeigen als solche Regulierungsbedarf bestehen würde.

Die gezielte Platzierung von Websites in den Suchmaschinen (über Anzeigen oder über die reguläre Trefferliste) wird unter dem Begriff Suchmaschinenmarketing gefasst, während der Teilbereich Suchmaschinenoptimierung sich exklusiv mit der Optimierung der Seiten für die regulären Trefferlisten befasst. Diese Optimierung kann von der Verbesserung des Texts auf der Website bis hin zum Aufbau komplizierter Verlinkungsstrukturen reichen, die die Suchmaschinen dahingehend täuschen sollen, dass eine bestimmte Seite zu einem Thema besonders relevant ist. In der Forschung noch unberücksichtigt ist die Frage, inwieweit es den Suchmaschinenoptimierern schon gelungen ist, für einen nennenswerten Anteil der gestellten Suchanfragen die Trefferlisten hin auf kommerzielle Treffer zu optimieren und inwieweit solche kommerzialisierten Trefferlisten als Qualitätsproblem anzusehen sind. Auf jeden Fall besteht eine Divergenz zwischen der tatsächlichen Zusammenstellung der Trefferlisten und dem Glauben der Nutzer an die Objektivität der Suchmaschinen.

4. Fazit

In diesem Artikel konnte gezeigt werden, dass es sich bei der Suchmaschinenforschung keineswegs – wie oft angenommen – um ein rein technisches Forschungsgebiet handelt, in dem es allein darum geht, Algorithmen für eine optimale Suche zu entwickeln. Vielmehr handelt es sich um einen vielschichtigen Forschungsbereich, der sich aus Ansätzen unterschiedlicher Disziplinen zusammensetzt. Neben den Kernfächern Informatik und Informationswissenschaft sind hier vor allem die Geistes- und Gesellschaftswissenschaften zu nennen. Aber auch die Rechtswissenschaften (hier wegen ihrer wenig ausgeprägten themenspezifischen Beschäftigung mit Suchmaschinen nicht weiter aufgeführt) beschäftigen sich mittlerweile mit den durch Suchmaschinen aufkommenden Problemen.

Im Sinne einer interdisziplinären Webwissenschaft muss auch die Suchmaschinenforschung als einer ihrer Teilbereiche interdisziplinär ausgerichtet sein. Keine einzelne Disziplin ist in der Lage, das gesamte durch Suchmaschinen aufgeworfene Themenfeld abzudecken. Wünschenswert ist allerdings

eine verstärkte Zusammenarbeit der einzelnen Disziplinen, die bisher – oft ohne die Kenntnis der Forschungsergebnisse der jeweils anderen Fächer – weitgehend für sich alleine arbeiten.

Ziel dieses Aufsatzes ist es, Anregungen für die weitere Forschung zu geben und zu den einzelnen Facetten des Themas Anknüpfungspunkte aufzuzeigen. In diesem Sinne würde es der Verfasser begrüßen, wenn sich weitere thematisch orientierte Kooperationen ergäben und diesem wichtigen Bereich der Webwissenschaft dadurch auch mehr Aufmerksamkeit zuteil werden würde.

Literatur

Acharya, Anurag / Cutts, Matt / Dean, Jeffrey / Haahr, Paul / Henzinger, Monika / Hoelzle, Urs / Lawrence, Steve / Pfleger, Karl / Sercinoglu, Olcan / Tong, Simon 2005: *Information retrieval based on historical data*. USA.

Beitzel, Steven M. / Jensen, Eric C. / Chowdhury, Abdur / Grossman, David / Frieder, Ophir 2004. "Hourly analysis of a very large topically categorized web query log." In: *Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* S. 321-328.

Beitzel, Steven M. / Jensen, Eric C. / Chowdhury, Abdur / Frieder, Ophir / Grossman, David 2007. "Temporal analysis of a very large topically categorized web query log." In: *Journal of the American Society for Information Science and Technology* 58, 2, S. 166-178.

Bergman, Michael K. 2001. "The deep web. Surfacing hidden value." In: *Journal of Electronic Publishing* 7, 1.

Berners-Lee, Tim / Hall, Wendy / Hendler, James A. / Shadbolt, Nigel / Weitzner, Daniel J. 2006a. "Creating a science of the web." In: *Science* 313, 5788, S. 769-771.

Berners-Lee, Tim / Hall, Wendy / Hendler, James A. / O'Hara, Kieron / Shadbolt, Nigel / Weitzner, Daniel J. 2006b. "A framework for web science." In: *Foundations and Trends in Web Science* 1, 1, S. 1-130.

Ferber, Reginald 2003: *Information Retrieval. Suchmodelle und Data-mining-Verfahren für Textsammlungen und das Web*. Heidelberg: d.punkt.

Fox, Vanessa 2007: *Live blogging. Microsoft searchification day 2007*. http://searchengineland.com/070927-000001.php (Abruf: 04.02.2008)

Graefe, Gernot / Maaß, Christian / Heß, Andreas 2007. "Alternative searching services. Seven theses on the importance of 'Social bookmarking'." In: *Conference on Social Semantic Web*. Hrsg. von Sören Auer, Chris Bizer, Claudia Müller und Anna V. Zhdanova. Leipzig: GI Gesellschaft für Informatik.

Granka, L. A. / Joachims, T. / Gay, G. 2004. "Eye-tracking analysis of user behavior in www search." In: *Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* S. 478-479.

Griesbaum, Joachim 2004. "Evaluation of three german search engines. Altavista.de, google.de and lycos.de." In: *Information Research* 9, 4.

Gulli, A. / Signorini, A. 2005. "The indexable web is more than 11.5 billion pages." In: *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, S. 902-903. Chiba, Japan.

Hotchkiss, G. / Garrison, M. / Jensen, St. 2006: *Search engine usage in North America. A research initiative by enquiro*. www.enquiro.com (Abruf: 16.03.2006)

Höchstötter, Nadine 2007. "Suchverhalten im Web. Erhebung, Analyse und Möglichkeiten." In: *Information Wissenschaft und Praxis* 58, 3, S. 135-140.

Jansen, Bernard J. 2007. "The comparatative effectiveness of sponsored and nonsponsored links for web e-commerce queries." In: *ACM Transactions on the Web* 1, 1, S. 1-25.

Jansen, Bernard J. / Spink, Amanda 2006. "How are we searching the world wide web? A comparison of nine search engine transaction logs." In: *Information Processing & Management* 42, 1, S. 248-263.

Karzauninkat, Stefan 2003. "Die Suchmaschinenlandschaft 2003. Wirtschaftliche und technische Entwicklungen." In: *Wegweiser im Netz*. Hrsg. von Marcel Machill und Carsten Welp. Gütersloh: Verlag Bertelsmann Stiftung, S. 509-538.

Lewandowski, Dirk 2004a. "Abfragesprachen und erweiterte Suchfunktionen von www-Suchmaschinen." In: *Information Wissenschaft und Praxis* 55, 2, S. 97-102.

Lewandowski, Dirk. 2004b. "Date-restricted queries in web search engines." In: *Online Information Review* 28, 6, S. 420-427.

Lewandowski, Dirk. 2005a. "*Web information retrieval. Technologien zur Informationssuche im Internet*. Frankfurt am Main: DGI.

Lewandowski, Dirk. 2005b. "Web searching, search engines and information retrieval." In: *Information Services and Use* 18, 3, S. 137-147.

Lewandowski, Dirk. 2007. "Mit welchen Kennzahlen lässt sich die Qualität von Suchmaschinen messen?" In: *Die Macht der Suchmaschinen / The power of search engines*. Hrsg. von Marcel Machill und Markus Beiler. Köln: von Halem, S. 243-258.

Lewandowski, Dirk. 2008a. “A three-year study on the freshness of web search engine databases.” In: *Journal of Information Science* 34 (im Druck).

Lewandowski, Dirk. 2008b. Problems with the use of web search engines to find results in foreign languages (erscheint).

Lewandowski, Dirk. 2008c. “The retrieval effectiveness of web search engines. Considering results descriptions.” In: *Journal of Documentation* 64 (im Druck).

Lewandowski, Dirk / Höchstötter, Nadine 2007. “Qualitätsmessung bei Suchmaschinen. System- und nutzerbezogene Evaluationsmaße.“ In: *Informatik Spektrum* 30, 3, S. 159-169.

Lewandowski, Dirk / Höchstötter, Nadine 2008. “Web searching: A quality measurement perspective.” In: *Web searching. Multidisciplinary perspectives*. Hrsg. von Amanda Spink und Michael Zimmer. Dordrecht: Springer.

Lewandowski, Dirk / Mayr, Philipp 2006. “Exploring the academic invisible web.” In: *Library Hi Tech* 24, 4, S. 529-539.

Lewandowski, Dirk / Wahlig, Henry / Meyer-Bautor, Gunnar 2006. “The freshness of web search engine databases.” In: *Journal of Information Science* 32, 2, S. 133-150.

Machill, Marcel / Beiler, Markus / Zenker, Martin 2007. “Suchmaschinenforschung. Überblick und Systematisierung eines interdisziplinären Forschungsfeldes.“ In: *Die Macht der Suchmaschinen / The power of search engines*. Hrsg. von Marcel Machill und Markus Beiler. Köln: von Halem, S. 7-43.

Machill, Marcel / Neuberger, Christoph / Schweiger, Wolfgang / Wirth, Werner 2003. “Wegweiser im Netz. Qualität und Nutzung von Suchmaschinen.“ In: *Wegweiser im Netz*. Hrsg. von Marcel Machill und Carsten Welp. Gütersloh: Bertelsmann Stiftung.

Marable, Leslie ?: *False oracles. Consumer reaction to learning the truth about how search engines work*. http://www.consumerwebwatch.org/dynamic/search-report-false-oracles.cfm (Abruf: 4.2.2008)

Mayer, Tim ?: *Our blog is growing up - and so has our index*. http://www.ysearchblog.com/archives/000172.html (Abruf: 4.2.2008)

Riemer, Kai / Brüggemann, Fabian 2007. “Personalisierung der Internetsuche. Lösungstechniken und Marktüberblick.“ In: *Wirtschaftsinformatik* 49, 2, S. 116-126.

Schmidt-Mänz, Nadine 2007: *Untersuchung des Suchverhaltens im Web. Interaktion von Internetnutzern mit Suchmaschinen*. Hamburg: Verlag Dr. Kovac.

Schmidt-Maenz, Nadine / Bomhardt, Christian 2005. “Wie suchen Onliner im Internet?“ In: *Science Factory/Absatzwirtschaft* 2, S. 5-8.

Sherman, Chris / Price, Gary 2001: *The invisible web. Uncovering information sources search engines can't see*. Medford, NJ: Information Today.

Spink, Amanda / Jansen, Bernard J. 2004. “Web search. Public searching of the web.” In: *Information science and knowledge management*. Hrsg. von J. Owen. Dordrecht u.a.: Kluwer Academic Publishers.

Spink, Amanda / Jansen, Bernard J. / Blakely, Chris / Koshman, Sherry 2006. “A study of results overlap and uniqueness among major web search engines.” In: *Information Processing & Management* 42, 5, S. 1379-1391.

Stock, Mechthild / Stock, Wolfgang G. 2000. “Klassifikation und terminologische Kontrolle. Yahoo!, Open Directory und Oingo im Vergleich.“ In: *Password* 12, S. 26-33.

Stock, Wolfgang G. 2007: *Information Retrieval. Informationen suchen und finden*. München: R. Oldenbourg.

Van Couvering, Elizabeth 2008. “The history of the internet search engine. Navigational media and traffic commodity.” In: *Web searching. Multidisciplinary perspectives*. Hrsg. von Amanda Spink und Michael Zimmer. Berlin, Heidelberg: Springer, S. 177-206.

Van Eimeren, Birgit / Frees, Beate 2007. “Internetnutzung zwischen Pragmatismus und Youtube-Euphorie. ARD/ZDF-Online-Studie 2007.“ In: *Media Perspektiven* 38, 8, S. 362-378.

Véronis, Jean 2006: *A comparative study of six search engines*. http://www.up.univ-mrs.fr/veronis/pdf/2006-comparative-study.pdf (Abruf: 4.2.2008)

Webhits 2007: *Webhits web-barometer*. http://www.webhits.de/deutsch/index.shtml?webstats.html (Abruf: 4.2.2008)