Web Information Retrieval

Dirk Lewandowski, Düsseldorf

Web Information Retrieval hat sich als gesonderter Forschungsbereich herausgebildet. Neben den im klassischen Information Retrieval behandelten Fragen ergeben sich durch die Eigenheiten des Web neue und zusätzliche Forschungsfragen. Die Unterschiede zwischen Information Retrieval und Web Information Retrieval werden diskutiert. Der zweite Teil des Aufsatzes gibt einen Überblick über die Forschungsliteratur der letzten zwei Jahre.

Web information retrieval
Web information retrieval is a research area of its own. While some questions are the same as in „classic" information retrieval, there are new questions concerning characteristics of the Web, which are discussed. The second part of the article gives an overview of the research literature published within the last two years.

1 Einleitung

Dieser Aufsatz gibt einen Überblick über den Stand der Forschung im Bereich Web Information Retrieval. Im ersten Teil werden die besonderen Probleme, die sich in diesem Bereich ergeben, anhand einer Gegenüberstellung mit dem „klassischen" Information Retrieval erläutert. Der weitere Text diskutiert die wichtigste in den letzten Jahren erschienene Literatur zum Thema, wobei ein Schwerpunkt auf die – so vorhanden – deutschsprachige Literatur gelegt wird. Der Schwerpunkt liegt auf Literatur aus den Jahren 2003 und 2004. Zum einen zeigt sich in dem betrachteten Forschungsfeld eine schnelle Entwicklung, so dass viele ältere Untersuchungen nur noch einen historischen bzw. methodischen Wert haben; andererseits existieren umfassende ältere Reviewartikel (s. v.a. *Rasmussen* 2003).

Schon bei der Durchsicht der Literatur wird allerdings deutlich, dass zu einigen Themenfeldern keine oder nur wenig deutschsprachige Literatur vorhanden ist. Leider ist dies aber nicht nur darauf zurückzuführen, dass die Autoren aus den deutschsprachigen Ländern ihre Ergebnisse in englischer Sprache publizieren. Vielmehr wird deutlich, dass in diesen Ländern nur wenig Forschung im Suchmaschinen-Bereich stattfindet. Insbesondere zu sprachspezifischen Problemen von Web-Suchmaschinen fehlen Untersuchungen.

Ein weiteres Problem der Forschung im Suchmaschinen-Bereich liegt in der Tatsache begründet, dass diese zu einem großen Teil innerhalb von Unternehmen stattfindet, welche sich scheuen, die Ergebnisse in großem Umfang zu publizieren, da sie fürchten, die Konkurrenz könnte von solchen Veröffentlichungen profitieren. So finden sich etwa auch Vergleichszahlen über einzelne Suchmaschinen oft nur innerhalb von Vorträgen oder Präsentationen von Firmenvertretern (z.B. *Singhal* 2004; Dean 2004).

Das Hauptaugenmerk dieses Artikels liegt auf der Frage, inwieweit Suchmaschinen in der Lage sind, die im Web vorhanden Inhalte zu indexieren, mit welchen Methoden sie dies tun und ob bzw. wie sie ihre Ziele erreichen. Ausgenommen bleiben damit explizit Fragen der Effizienz bei der Erschließung des Web und der Skalierbarkeit von Suchmaschinen. Anders formuliert: Diese Übersicht orientiert sich an klassisch informationswissenschaftlichen Fragen und spart die eher im Bereich der Informatik diskutierten Fragen weitgehend aus.

2 Quellen

Die Forschungsergebnisse zu Suchmaschinen finden sich in unterschiedlichsten Zeitschriften und in den Tagungsbänden unterschiedlicher Konferenzen statt. Bei den informationswissenschaftlichen Zeitschriften sind besonders *Information Processing and Management, Journal of the American Society for Information Science and Technology, Journal of Documentation* und *Online Information Review* hervorzuheben. Bei den Konferenzen sind neben den World-Wide-Web-Konferenzen vor allem die Veranstaltungen der Special Interest Group on Information Retrieval (SIGIR) der ACM zu nennen. Eine weitere wichtige Quelle für neuere Entwicklungen in diesem Bereich sind Patente oder Patentanmeldungen.[1] Im deutschsprachigen Bereich finden sich Forschungsergebnisse bisher nur vereinzelt und verstreut.

Als wichtigste Nachrichtenquelle für Entwicklungen im Suchmaschinen-Bereich ist Search Engine Watch (www.searchenginewatch.com) zu nennen; Nachrichten in deutscher Sprache (allerdings in weit geringerem Umfang und Tiefe) bietet @-Web (www.at-web.de); lesenswert sind auch der Weblog von Klaus *Schallhorn* (www.kso.co.uk/de/blog) sowie die Beiträge der Suchfibel-Diskussionsliste (www.suchfibel.de/aktuell/mailingliste.htm).

3 Web Information Retrieval vs. „klassisches" Information Retrieval

Mit den Unterschieden zwischen klassischen Information Retrieval und dem Web-Retrieval haben sich bereits viele Autoren beschäftigt (vgl. u.a. *Huang* 2000; *Chowdhury* 1999; *Brooks* 2003; *Chu* 2003, 128-139; *Ferber* 2003, 285-292; *Savoy* 2002). Die Unterschiede sind in Tabelle 1 zusammengefasst, wobei sie in vier Klassen unterteilt wurden. Dies sind Unterschiede hinsichtlich des zugrunde liegenden Dokumentenkorpus, der Inhalte, der Nutzer und hinsichtlich der Eigenarten des IR-Systems selbst.

Von besonderer Bedeutung bei der Erschließung des Web ist, dass die genaue Dokumentenmenge des WWW nicht bekannt ist und auch nicht ermittelt werden kann und dass Hyperlink-Strukturen einer gewissen Form existieren, die die vollständige Erfassung erschweren (*Broder* et al. 2000). Diese Probleme liegen bei der Erschließung von Dokumenten in klassischen Datenbanken nicht vor. Hier ist die zu erfassende Datenmenge aufgrund der schon bei der Planung der Datenbank gemachten Einschränkung der Dokumentenmenge bekannt. Probleme des Auffindens von neuen Dokumenten bestehen nicht in der gleichen Form.

In Bezug auf die Sprache der zu erschließenden Dokumente besteht im Web das Problem, dass Dokumente potenziell in allen Sprachen vorkommen können. Da von Seiten der Suchmaschinen kein einheitliches Indexierungsvokabular vorliegt, sondern auf die Volltexterschließung gesetzt wird, können die Dokumente auch jeweils nur bei Eingabe der Suchbegriffe in der Sprache der zu findenden Dokumente gefunden werden. Im Bereich der Online-

1 Eine regelmäßige Übersicht neuer US-Patente und US-Patentanmeldungen im Bereich Information Retrieval bietet die News-Seite Resourceshelf (www.resourceshelf.com).

Datenbanken sind in einer Datenbank entweder nur Dokumente in einer Sprache enthalten, oder aber die in unterschiedlichen Sprachen verfassten Dokumente werden mittels eines einheitlichen Vokabulars in der Zielsprache der Datenbank erschlossen. Als weitere Hilfsmittel werden Klassifikationssysteme und multilinguale Thesauri eingesetzt.

Ein weiteres mit der Vielfalt des Web verbundenes Problem taucht in Form unterschiedlicher Medienarten bzw. Dateiformate auf. Das Web ist nicht auf Textdokumente beschränkt, sondern enthält beispielsweise viele Multimedia-Informationen. Die Erschließung dieser Informationen muss aufgrund der mangelnden Textmenge grundsätzlich anders erfolgen als die der Textdokumente. Probleme bei Web-Dokumenten bereitet auch die stark differierende Länge der Dokumente und deren eventuell bestehende Granularität (*Ferber* 2003, 287). Im Web finden sich aus nur wenigen Wörtern bestehende Dokumente ebenso wie komplette Bücher, die als einzelnes Dokument verfügbar gemacht wurden. Teils werden längere Dokumente jedoch auch in Teile zerlegt, um den Zugriff zu verbessern.

Eines der größten Probleme der Web-Indexierung ist die mangelnde Strukturierung der Dokumente. Zwar sind in HTML-Dokumenten durchaus Strukturen vorhanden, diese werden von den Autoren jedoch nicht bewusst ausgenutzt und sind daher eher implizit vorhanden, wodurch die Erschließung erschwert wird.

Während in Online-Datenbanken jedes Dokument nur einmal abgelegt wird und es klare Kriterien für die Aufnahme von Dokumenten in die Datenbank gibt (*Xie* 2004), findet sich im Web aufgrund der dezentralen Struktur eine hohe Anzahl an Dubletten. Einerseits werden komplette Server gespiegelt *(mirror hosts)*, andererseits werden die gleichen Texte in unterschiedliche Angebote integriert. Für die Suchmaschinen ist die Eliminierung jeglicher Dubletten von besonderer Bedeutung, weil sie die gerankten Trefferlisten verstopfen können. Weiterhin besteht das Problem unterschiedlicher Versionen des gleichen Texts. Während in Datenbanken in der Regel nur eine, nämlich die endgültige Fassung eines Dokuments abgelegt wird, existieren von vielen Dokumenten im Web unterschiedliche Versionen, die nicht leicht durch automatische Verfahren als solche erkannt werden können.

Ein besonderes Problem der Dokumentenkollektion betrifft die Zuverlässigkeit der zu erschließenden Dokumente. Während im klassischen Information Retrieval nie das Problem bestand, jedes zu erfassende Dokument auf seine Qualität hin kontrollieren zu müssen, ist dies für die Aufnahme in einen Suchmaschinen-Index essenziell. Nur Dokumente, die tatsächlich für den Benutzer zur Lösung eines Informationsproblems sinnvoll sind, sollen in den Datenbestand aufgenommen werden (*Chu* 2003, 128); alle Suchmaschinen bestimmen inzwischen einen Wert für die *Autorität* jedes Dokuments.

Auch in Bezug auf die Nutzer gibt es wesentliche Unterschiede zwischen den Online-Datenbanken und dem Web (s.u.). Vergleicht man die Art der gestellten Anfragen von Web-Nutzern und den Nutzern von Online-Datenbanken, so lässt sich klar feststellen, dass die Datenbank-Nutzer mit den Abfragesprachen und komplexen Suchmöglichkeiten dieser Systeme umgehen können und genau formulierte Suchanfragen verwenden. Weiterhin enthalten die von Suchmaschinen-Nutzern gestellten Anfragen zu einem hohen Anteil fehlerhafte Eingaben. Dazu kommt, dass das Nutzerinteresse bei Online-Datenbanken aufgrund der Homogenität der Inhalte klar fokussiert ist; an Suchmaschinen werden hingegen Anfragen unterschiedlichster Ausrichtung gestellt (vgl. *Broder* 2002).

Als letzte Klasse der Unterschiede zwischen den beiden Typen von IR-Systemen sind schließlich die Eigenarten der jeweiligen Systeme zu nennen. Dieser Bereich ist allerdings am ehesten Veränderungen unterworfen, da sich die Funktionen der Systeme relativ leicht ändern bzw. verbessern lassen. Allerdings haben sich bei den Suchmaschinen bestimmte Standards in Bezug auf die Funktionalitäten herausgebildet, die sich wesentlich von denen bei den Online-Datenbanken unterscheiden.

Wie schon beschrieben, sind die Suchanfragen bei Web-Suchmaschinen weit weniger komplex als die in Online-Datenbanken. Während frühe Suchmaschinen wie AltaVista noch versuchten, die komplexen Abfragemöglichkeiten der klassischen IR-Systeme nachzubilden, verzichten neuere Suchmaschinen weitgehend auf diese, da solche Funktionen von den Nutzern nur in sehr geringem Umfang angenommen werden. Suchmaschinen bieten also keine den klassischen IR-Systemen vergleichbare Suchmöglichkeiten. Dies gilt sowohl für die Standard-Abfragemöglichkeiten wie boolesche Suche, Abstandsoperatoren und Trunkierung als auch für speziellere Abfragemöglichkeiten wie gewichtetes Retrieval oder Fuzzy-Suche (*Chu* 2003, 130f.; siehe auch Stock 2000; *Lewandowski* 2004a).

Bei den Interfaces der Web-Suchmaschinen ist als großer Vorteil hervorzuheben, dass sich die Interfaces stark ähneln und ein Wechsel von einem zum anderen System daher in der Regel problemlos möglich ist. Online-Datenbanken verfügen in der Regel über weit komplexere Interfaces, die oft auch gezielt auf die speziellen Inhalte der jeweiligen Datenbank ausgerichtet sind. Es existieren allerdings auch Interfaces von kommerziellen Hosts, die auf die Suche in sehr großen Datenbeständen ausgerichtet sind. Hier erfolgt die Suche jedoch in mehreren Schritten, so dass die Treffermenge schon in der Vorbereitung der eigentlichen Suche eingeschränkt werden kann. Vor allem geschieht dies durch eine gezielte Auswahl der zu durchsuchenden Quellen. Auch die Möglichkeiten der Modifikation einer bereits gestellten Suchanfrage sind bei den Suchmaschinen außerordentlich beschränkt. In der Regel wird nur die Option angeboten, nochmals in den bereits gefundenen Ergebnissen zu suchen.

Eine Eigenheit der Suchmaschinen ist der automatische Vorschlag von weiteren Suchbegriffen, um die Suche einzuschränken, zu erweitern oder zu verändern (*Chu* 2003, 134f.).

Alle Suchmaschinen setzten bei der Sortierung der Trefferlisten auf Ranking-Mechanismen. Auch in klassischen IR-Systemen werden teils Ranking-Verfahren eingesetzt. Diese verwenden in der Regel Faktoren wie *term frequency, term proximity, term location* und *inverse document frequency*, bei WWW-Suchmaschinen kommen weitere Faktoren hinzu: linktopologische Verfahren, Verfahren auf Basis der Auswertung von Seitenbesuchen (Klicks) und Mischverfahren aus klassischem Ranking und linktopologischen Verfahren.

Aus den dargestellten Merkmalen ergeben sich für das Web-Information-Retrieval eigene Forschungsthemen, die auf dem klassischen Information Retrieval basieren. Diese werden im folgenden Text anhand eines Literaturüberblicks dargestellt. Eine Übersicht der auf algorithmischer Ebene bestehenden Probleme bieten *Henzinger, Motwani* u. *Silverstein* (2002). Diese sind: Spam, die Qualität der Inhalte, Web-Konventionen, gespiegelte Hosts und die nur schwache Struktur der im Web vorhandenen Daten.

Die Forschung im Bereich Web Information Retrieval findet zu einem großen Teil bei den Anbietern von Suchmaschinen selbst statt. Aufgrund der Konkurrenzsituation zwischen den Anbietern und der Gefahr der Manipulation der Trefferlisten durch Search Engine Optimizers (SEOs) werden die Ergebnisse entsprechender Forschungen bzw. Implementierungsbeispiele oft nicht veröffentlicht und können nur im Nachhinein durch die Analyse der implementierten Funktionen in den Suchmaschinen erschlossen werden. Dies erschwert die Erschließung des Forschungsfelds.

Tabelle 1: Unterschiede zwischen Web-IR und klassischem Information Retrieval

Unterscheidungsmerkmal	Web	Klassische Datenbanken
Merkmale des Dokumentenkorpus Sprachen	Dokumente liegen in einer Vielzahl von Sprachen vor; aufgrund der Volltexterschließung keine einheitliche Erschließung über Sprachgrenzen hinweg	Einzelne Sprache oder Dokumente liegen in vorher definierten Sprachen vor; Erschließung von Dokumenten verschiedener Sprachen mittels einer einheitlichen Indexierungssprache.
Medienarten	Dokumente in unterschiedlichen Formaten	Dokumente liegen in der Regel in nur einem Format vor.
Länge und Granularität der Dokumente	Länge der Dokumente variiert, große Dokumente werden oft aufgeteilt.	Länge der Dokumente variiert innerhalb eines gewissen Rahmens; pro Dokument eine Dokumentationseinheit.
Spam	Problem der von den Suchmaschinen unerwünschten Inhalte	Beim Aufbau der Datenbank wird vorab definiert, welche Dokumente erschlossen werden.
Hyperlink-Struktur	Dokumente sind miteinander verbunden.	Dokumente sind in der Regel nicht miteinander verknüpft; keine Notwendigkeit, aus Verlinkungsstrukturen auf die Qualität der Dokumente zu schließen.
Inhalte Datenmenge/Größe des Datenbestands	genaue Datenmenge nicht bestimmbar; keine vollständige Indexierung möglich	genaue Datenmenge aufgrund formaler Kriterien bestimmbar
Abdeckung des Datenbestands	Abdeckung der Zielmenge unklar	Abdeckung gemäß dem bei der Planung der Datenbank gesteckten Ziel in der Regel vollständig
Dubletten	Dokumente können mehrfach/vielfach vorhanden sein; teils auch in unterschiedlichen Versionen	Dublettenkontrolle bei der Erfassung der Dokumente. Versionskontrolle in der Regel nicht notwendig, da jeweils eine endgültige Fassung existiert und diese in die Datenbank eingestellt wird
Nutzer unterschiedliche Interessen Art der Anfragen	aufgrund heterogener Informationsbedürfnisse der Nutzer sehr unterschiedlich	genaue Zielgruppe mit klarem Informationsbedürfnis
Ill-formed queries	geringe Kenntnis der Nutzer über angebotene Suchfunktionen/Recherche allgemein	Nutzer sind mit der jeweiligen Abfragesprache vertraut
IR-System Interface	einfache, intuitiv bedienbare Interfaces für Laien-Nutzer	oft komplexe Interfaces; Einarbeitung notwendig
Ranking	Relevance Ranking aufgrund der großen Treffermengen notwenig	Relevance Ranking aufgrund genau formulierter Suchanfragen und dadurch geringerer Treffermengen meist nicht nötig
Suchfunktionen	beschränkte Suchfunktionen	komplexe Abfragesprachen
Modifikation der Suche	in der Regel nur Möglichkeiten zur weiteren Einschränkung der Suchanfrage	umfangreiche Modifikationsmöglichkeiten
Strukturierung der indexierten Dokumente	schwache Strukturierung; Feldsuche nur bedingt für die Recherche geeignet	starke Strukturierung; Suche innerhalb einzelner Felder gut für die Recherche geeignet
Auswahl der Dokumente	abgesehen vom Ausschluss von Spam keine weitere Auswahlkriterien	Klare Auswahlkriterien werden schon bei der Planung der Datenbank bestimmt.

4 Suchmaschinen-Markt

Den Veränderungen auf dem Suchmaschinen-Markt haben sich verschiedene Autoren gewidmet. *Griesbaum* (2003), *Lewandowski* (2003), *Karzauninkat* (2003) und *Griesbaum* u. *Bekavac* (2004) geben jeweils einen Überblick über die laufenden Entwicklungen und stellen Momentaufnahmen eines sich im Wandel befindlichen Markts dar. In diesen Darstellungen wird deutlich, dass sich der Markt auf wenige bedeutende Anbieter konzentriert. Hervorzuheben ist als Element des Artikels von Karzauninkat (2003) ein Schaubild der Verflechtungen der Suchdienste-Anbieter, welches auch in aktualisierter Form im Web verfügbar ist.[2]

5 Index-Qualität

Größe

Zur Größe der Suchmaschinen-Indizes gibt es einerseits Angaben der Betreiber, andererseits werden Untersuchungen durchgeführt, die die tatsächlichen Größen hochzurechnen versuchen. Die von den Suchmaschinen-Betreibern angegebenen Indexgrößen und deren Entwicklung im Lauf der Jahre werden in *Sullivan* (2003) dargestellt. Noch nicht berücksichtigt sind hier allerdings die neueren Entwicklungen wie der Start der Microsoft-eigenen MSN-Suchmaschine mit nach eigenen Angaben etwa fünf Milliarden indexierten Dokumenten und der Erweiterung des Google-Index auf etwa acht Milliarden Dokumente.

Von manchen wichtigen Suchmaschinen sind keine Angaben über die Indexgröße zu erhalten. So gibt Yahoo beispielsweise keine Zahlen bekannt; im deutschsprachigen Bereich ist hier die Suchmaschine Seekport zu nennen.

Da die von den Betreibern genannten Zahlen nicht durch eine unabhängige Stelle überprüft werden, wurden sie oft angezweifelt. Greg *Notess* (Notess 2003a, Notess 2003b) führte regelmäßig eine Untersuchung durch, die die tatsächliche Größe der Suchmaschinen-Datenbanken ermitteln sollte. Ausgehend von einer in einer Suchmaschine (All the Web) ermittelbaren tatsächlichen Anzahl indexierter Dokumente werden anhand von ausgewählten Suchanfragen die Ergebnismengen verglichen und auf ihrer Basis die angenäherte tat-

2 http://www.suchfibel.de/5technik/suchmaschinen_beziehungen.htm [19.11.2004]

sächliche Indexgröße berechnet. Die letzten Ergebnisse (vom Januar 2003) zeigen, dass die damals größten Suchmaschinen Google und AlltheWeb ihre Indexgrößen mit drei bzw. zwei Milliarden indexierten Dokumenten korrekt angaben, während AltaVista und Teoma deutlich mehr Dokumente indexiert hatten als angegeben. Die damals auf dem Inktomi-Index basierenden Suchmaschinen HotBot und MSN, die angaben, etwa drei Milliarden Dokumente indexiert zu haben, hatten jedoch de facto nur etwas mehr als eine Milliarde Dokumente in ihrem Datenbestand.

Die Untersuchung zeigt, dass die Angaben der Suchmaschinen-Betreiber nur bedingt als zuverlässig angesehen werden können. Dazu kommt, dass von manchen Anbietern keine Zahlen vorliegen. Insbesondere über die von den originär deutschen Suchmaschinen bzw. zu den von den internationalen Suchmaschinen indexierten deutschsprachigen Dokumenten liegen keine Untersuchungen vor; hier besteht noch ein erheblicher Forschungsbedarf.

Vaughan und *Thelwall* (2004) untersuchen die Abdeckung von Websites in unterschiedlichen Ländern durch drei Universalsuchmaschinen. In die Untersuchung einbezogen werden die Suchmaschinen *Google, All the Web* und *AltaVista*. Die untersuchten Länder sind die Vereinigten Staaten, China, Singapur und Taiwan. Die Länder sind so gewählt, dass bei der Auswertung der Ergebnisse eine Unterscheidung getroffen werden kann, ob eine eventuell auftauchende Verzerrung aufgrund der Sprache oder aufgrund der Verlinkungsstruktur besteht. Für die USA und Singapur wurden englischsprachige Seiten ausgewertet, für China und Taiwan jeweils chinesischsprachige. Die Untersuchung ergab, dass die Abdeckung der Sites nach Ländern und Suchmaschinen erheblich differiert. Die beste Abdeckung ergab sich wie erwartet bei den US-Sites, sie lag hier zwischen 80 und 87 Prozent. Die Abdeckung der Sites aus China lag zwischen 52 und 70 Prozent, derjenigen aus Singapur zwischen 41 und 56 Prozent und der aus Taiwan zwischen vier und 75 Prozent, wobei hier AltaVista mit nur vier Prozent Abdeckung einen deutlichen Ausreißer gegenüber den anderen beiden Suchmaschinen darstellt. Auch bei der Tiefe der Indexierung der Sites zeigen sich deutliche Unterschiede. Während von den US-Sites durchschnittlich 89 Prozent der Seiten indexiert werden, sind dies bei den Sites aus China nur 22 Prozent und bei denen aus Taiwan sogar nur drei Prozent. Bisher unbekannt ist die Abdeckung der Seiten aus anderen Ländern.

Aktualität

Die Aktualität der Suchmaschinen-Datenbanken ist von Bedeutung, da sich eine große Anzahl von Webseiten oft verändert. Während natürlich auch Seiten vorhanden sind, die ins Netz gestellt werden und sich dann nicht mehr verändern, gibt es viele Seiten, die beispielsweise aktuelle Nachrichten anbieten und deshalb von den Suchmaschinen in kurzen Abständen besucht werden sollten. Das Problem wurde besonders bei wichtigen Nachrichten-Ereignissen deutlich (vgl. *Schiff* 2003). Die Suchmaschinen reagierten einerseits mit eigenen Nachrichtensuchmaschinen und andererseits mit häufigeren Aktualisierungen ihrer Haupt-Indizes.

Greg *Notess* führte in regelmäßigen Abständen eine Untersuchung zur Aktualität der Datenbestände der Suchmaschinen durch (zuletzt 2003: Notess 2003c; zur grundlegenden Problematik Notess 2001). Als Ergebnis ist festzuhalten, dass die meisten Suchmaschinen zwar in den letzten Tagen ihrem Index neue Dokumente hinzugefügt haben, der Großteil der indexierten Dokumente allerdings vor etwa einem Monat zuletzt indexiert wurde. Es finden sich aber auch Dokumente, die schon seit längerer Zeit nicht mehr besucht wurden.

6 Invisible Web

Nachdem im Jahr 2001 einerseits die grundlegende Monographie zur Invisible-Web-Thematik (*Sherman* u. *Price* 2001), andererseits eine vielbeachtete Untersuchung zur Größe des Invisible Web (*Bergman* 2001) erschienen war, brachten die folgenden Jahre weniger Literatur zu diesem Thema. Einige Aufsätze beschäftigten sich mit den genannten Quellen und diskutieren vor allem die Größenschätzung von Bergman (so z.B. *Stock* 2003). Allgemeiner Tenor ist, dass die Schätzung des Invisible Web mit einer angeblichen Größe des 400- bis 500-fachen des Surface Web abgelehnt wird. Die Größe wird weit geringer angenommen.

Um die Invisible-Web-Problematik zu lösen, werden zwei unterschiedliche Ansätze verfolgt. Einerseits sollen Invisible-Web-Quellen in Meta-Suchinterfaces integriert werden, die neben regulären Suchmaschinen eben auch Datenbanken abfragen, die für normale Suchmaschinen nicht erreichbar sind (*Hamilton* 2003). Allerdings können diese Lösungen die Problematik nicht im Gesamten erfassen, sondern beschränken sich immer nur auf die Integration weniger Quellen. Dazu kommen die von den Meta-Suchmaschinen bekannten Limitierungen.

Der zweite Ansatz versucht, Dokumente, die eigentlich *invisible* sind, in das Surface Web zu holen. Dazu werden Brückenseiten generiert, die in HTML-Form Inhalte aus

Datenbanken darstellen. Solche Ansätze werden vor allem von kommerziellen Anbietern verfolgt, die beispielsweise die Inhalte aus Buchkatalogen in die Trefferlisten der Suchmaschinen bringen wollen. Allerdings wird auch von öffentlicher Seite auf solche Methoden gesetzt (Seiffert 2003). Eine Diskussion der Problematik auch unter ethischen Gesichtspunkten bietet *Heinisch* (2003).

7 Ranking

Rankingverfahren sind traditionell der Bereich, zu dem es von Seiten der Suchmaschinen am wenigsten Veröffentlichungen gibt, da diese Verfahren den Kern der Suchmaschinen ausmachen und ihre individuelle Qualität bestimmen. Neue Rankingverfahren werden selten der Fachöffentlichkeit vorgestellt.
Suchmaschinen arbeiten mit drei verschiedenen Arten von Rankingverfahren:

- klassische Verfahren, die den Inhalt der Dokumente auswerten
- nutzungsstatistische Verfahren, die das Klick-Verhalten der Nutzer auswerten (z. B. *Culliss* 2000; *Dean* et al. 2002)
- linktopologische Verfahren, die die Struktur des Web zur Basis des Rankings machen (die beiden klassischen Verfahren sind *Page* et al. 1998 und *Kleinberg* 1999; vor allem zu PageRank liegen einige Verbesserungsvorschläge vor, z.B. *Richardson* u. *Domingos* 2004; *Thelwall* u. *Vaughan* 2004)

An Bedeutung gewinnen Verfahren, die die „Qualität" von Dokumenten messen. Dies kann durch die beiden letztgenannten Verfahren erreicht werden; allerdings sind weitere Kriterien nötig (vgl. den Beitrag von *Mandl* in diesem Heft, S. 13).

8 Retrievaltests/Evaluierung

Die Leistungsfähigkeit von Suchmaschinen wird, abgesehen von Vergleichen der Indexgrößen, durch Retrievaltests gemessen. Hier wird in der Regel die Precision einer bestimmten Anzahl von Treffern gemessen und auf dieser Basis werden dann unterschiedliche Suchmaschinen miteinander verglichen. Auffällig ist hier der hohe Anteil populärwissenschaftlicher sowie kommerzieller Untersuchungen (vgl. u. a. Veritest 2003).

Der für deutschsprachige Suchanfragen sicher wichtigste derartige Test ist der von *Griesbaum*, *Rittberger* u. *Bekavac* (2002), ergänzt durch eine Nachfolgeuntersuchung (*Griesbaum* 2004). In diesen Untersuchungen werden die Suchmaschinen AltaVista, Google und Lycos miteinander verglichen[3]. Die Basis sind 50 Suchanfragen aus dem (sozial-) wissenschaftlichen Bereich; die Bewertung der Treffer wird von unabhängigen Juroren durchgeführt. Als relevante Treffer werden sowohl direkt relevante Treffer gewertet als auch Seiten, die selbst nicht relevant sind, jedoch auf ein relevantes Dokument verweisen. Im Vergleich schneidet die Suchmaschine Google jeweils am besten ab; auffällig ist jedoch erstens, dass die Mean Average Precision in den Top 20 bei allen Suchmaschinen unter 0,6 liegt und zweitens, dass die Unterschiede zwischen den untersuchten Suchmaschinen relativ gering sind.

Einen Überblick über die Trends bei der Evaluierung von Information-Retrieval-Systemen allgemein mit besonderem Schwerpunkt auf Web- und Multimedia-Dokumenten findet sich in *Mandl* (2003).

Mit den Mängeln der klassischen Retrievaltests in ihrer Anwendung auf Web-Suchmaschinen beschäftigen sich *Egghe* (2004), *Vaughan* (2004) und *Spink* (2002). Spink bemängelt, dass in Retrievaltests das Nutzerverhalten keine Beachtung findet. Zu denken ist hier etwa an Suchmaschinen, die den Nutzer in seinem Rechercheprozess unterstützen und ihn so vielleicht besser zu seinem Ziel führen.

9 Nützlichkeit von Funktionen

Die Funktionalitäten und der Befehlsumfang von Suchmaschinen werden im Allgemeinen kritisch gesehen. Gegenüber den klassischen Retrieval-Systemen bieten die Suchmaschinen nur eingeschränkte Funktionen; für die Bedürfnisse des professionellen Rechercheurs sind sie weniger geeignet. Dennoch bieten sie einige der „klassischen" Funktionen sowie eigene Funktionen, die in anderen Retrieval-Systemen nicht vorhanden sind bzw. Besonderheiten der Suche im Web bedienen. Tabellarische Überblicke finden sich in *Notess* (2004) und *Lewandowski* (2004a)[4]. *Hock* (2004) beschäftigt sich mit den Möglichkeiten der feldbeschränkten Suche; *Gelernter* (2003) betont die Nützlichkeit von spezialisierten gegenüber den allgemeinen Suchmaschinen und zeigt anhand von Fallbeispielen deren Recherchemöglichkeiten und Quellenspektrum auf.

Mit der Nützlichkeit von Operatoren bei der Suche in Web-Suchmaschinen befassen sich *Eastman* u. *Jansen* (2003). Sie kommen zu dem Schluss, dass der Einsatz von Operatoren die Suchergebnisse nicht signifikant verbessert. Allerdings mag dies zu einem wesentliche Teil den in der Untersuchung verwendeten Suchanfragen geschuldet sein, die einerseits thematisch wenig komplex waren und andererseits jeweils nur einen Operator (diesen allerdings teils mehrmals) einsetzten. Anfragen, in denen verschiedene Operatoren kombiniert eingesetzt wurden, wurden nicht gestellt.

Die Frage nach der Funktionstüchtigkeit der von den Suchmaschinen angebotenen Suchfunktionen wird nur selten gestellt. So ist beispielsweise unklar, wie gut die von allen Suchmaschinen angebotene Beschränkung auf eine oder mehrere Sprachen tatsächlich funktioniert. Beobachtungen zeigen, dass durchaus Dokumente der falschen oder keiner Sprache zugeordnet werden. Systematische Untersuchungen hierzu stehen jedoch noch aus.

Eine Untersuchung von *Lewandowski* (2004b) hinterfragt die Funktionsfähigkeit der von den Suchmaschinen angebotenen Datumsbeschränkung und untersucht exemplarisch die Suchmaschinen Google, Teoma und Yahoo!. Das Ergebnis ist, dass die Suchmaschinen diese Funktion nur unzureichend beherrschen. Selbst bei der Suchmaschine, die im Vergleich am besten abschneidet, werden noch etwa vier von zehn Treffern dem falschen Zeitraum zugeordnet.

Ntoulas et al. (2004) untersuchen die Datumsproblematik und empfehlen, Suchmaschinen nicht nur die Veränderungsfrequenz berücksichtigen zu lassen, sondern auch den Grad der Veränderung. Bisher werten Suchmaschinen auch kleine Veränderungen an einem Dokument als Aktualisierung desselben.

10 Nutzer

Im Bereich der Nutzerstudien liegen aktuell zwei umfangreiche Werke vor. Für den US-amerikanischen Raum sind dies *Spink* u. *Jansen* (2004)[5], deutsche Nutzer werden in *Machill* et al. (2003) untersucht. Eine Untersuchung des Nutzerverhaltens bei Suchen im Bereich des privaten Interesses bietet *Rieh* (2004).

Spink u. Jansen verwenden für ihre Untersuchungen Logfiles der Suchmaschinen Excite, All the Web und AltaVista. Es wird jeweils das komplette Logfile eines Tages ausgewertet. Die Daten wurden zwischen 1997 und 2002 erhoben; damit liegt die erste Untersuchung des Nutzerverhaltens über einen längeren Zeitraum vor. Die Kernergebnisse sind: Die Anfragen der Suchmaschinen-Nutzer sind kurz, die Länge der Anfragen nimmt nur langsam zu. Weiterhin sind die Anfragen wenig komplex und enthalten nur zu einem ge-

3 in der ersten Untersuchung auch Fireball
4 Eine Aktualisierung der dort abgedruckten Tabelle mit Berücksichtigung der seitdem neu hinzugekommenen Suchmaschinen findet sich unter www.durchdenken.de/lewandowski/doc/tabelle.php [19.11.2004]
5 s. auch Besprechung in diesem Heft, S X. Einen Überblick der wesentlichen Ergebnisse bietet auch Spink (2003).

ringen Teil Operatoren. Thematisch hat sich ein Wandel vollzogen weg von Anfragen vor allem aus den Bereichen Technologie und Sex hin zu E-Commerce-Anfragen. Spink u. Jansen bieten weitere Untersuchungsergebnisse für einzelne Themenbereiche. Diese sind E-Commerce, medizinische Themen, Sex und Multimedia.

In der Untersuchung von *Machill* et al. (2003) werden Nutzer einerseits in einer repräsentativen Umfrage befragt, andererseits wird ihr Verhalten in einem Laborexperiment beobachtet. Auch hier wird deutlich, dass der typische Suchmaschinennutzer keine klare Recherchestrategie verfolgt, sondern einfache Anfragen stellt und sich dann aus der Trefferliste passende Treffer herauspickt. Von den umfangreichen Möglichkeiten der Suchmaschinen werden nur die wenigsten genutzt. Die Bereitschaft, erweiterte Suchfunktionen einzusetzen bzw. diese erst einmal zu erlernen, ist gering.

Ebenfalls gering ist die Kenntnis der Nutzer über die Funktionsweise von Suchmaschinen. Zu diesem Ergebnis kommt sowohl die Studie von Machill et al. als auch eine Untersuchung der Verbraucherorganisation Consumer Web Watch (*Marable* 2003). Hier zeigten sich Suchmaschinennutzer unter anderem auch überrascht von den Businessmodellen der Suchmaschinen und der Platzierung von Werbelinks.

11 Fazit

Der Bereich Web Information Retrieval stellt sich als ein dynamisches Forschungsfeld mit einem umfangreichen Ausstoß an Forschungsergebnissen dar. Ein großes Problem ist allerdings darin zu sehen, dass ein großer Teil der innerhalb von Unternehmen erzielten Forschungsergebnisse nicht veröffentlicht wird, um dem eigenen Unternehmen Vorteile gegenüber der Konkurrenz zu verschaffen. Für die Zukunft ist eine verstärkte Zusammenarbeit zwischen akademischer Forschung und Suchmaschinen-Unternehmen zu wünschen.

Für den deutschsprachigen Bereich ist eine verstärkte Forschung im Bereich Web Information Retrieval zu wünschen. Diese anzuregen ist das wesentliche Ziel dieses Beitrags.

Information Retrieval, Forschung, Suchmaschine, Übersichtsbericht

Swets

Der Autor

Dirk Lewandowski

promoviert zur Zeit an der Heinrich-Heine-Universität Düsseldorf mit einer Arbeit im Bereich Web Information Retrieval. Er betreut die Rubrik „Suchmaschinen-News“ in der Zeitschrift Password und unterrichtete in den vergangenen Jahren in Düsseldorf und Köln. Studium 1994-2001 (Bibliothekswesen in Stuttgart; Informationswissenschaft und Philosophie in Düsseldorf); beruflichen Stationen im Wirtschaftsministerium des Landes NRW sowie bei der NRW Medien GmbH.

Heinrich-Heine-Universität
Institut für Sprache und Information
Abteilung Informationswissenschaft
Universitätsstraße 1, 40225 Düsseldorf
E-Mail: dirk.lewandowski@uni-duesseldorf.de
www.durchdenken.de/lewandowski

Literatur

Amitay, E.; *Har'El*, N.; *Sivan*, R.; *Soffer*, A. (2004): Web-a-Where: Geotagging Web Content. SIGIR '04, July 25-29, 2004, Sheffild, South Yorkshire, UK. 273-280

Bergman, M. K. (2001): The Deep Web: Surfacing Hidden Value. In: The Journal of Electronic Publishing 7(1). http://www.press.umich.edu/jep/07-01/bergman.html [13.1.2004]

Broder, A. (2002): A taxonomy of web search. In: SIGIR Forum 36(2). http://www.acm.org/sigir/forum/F2002/broder.pdf [12.7.2004]

Broder, A.; *Kumar*, R.; *Maghoul*, F.; *Raghavan*, P.; *Rajagopalan*, S.; *Stata*, R.; *Tominks*, A.; *Wiener*, J. (2000): Graph Structure in the Web http://www.almaden.ibm.com/webfountain/resources/GraphStructureintheWeb.pdf [5.3.2004]

Brooks, T. A. (2003): Web Search: how the Web has changed information retrieval. In: Information Research 8(3), http://informationr.net/ir/8-3/paper154.html [18.3.2004]

Chakrabarti, S. (2003): Mining the Web: Discovering Knowledge from Hypertext Data. Amsterdam (u.a.): Morgan Kaufmann

Chowdhury, G. G. (1999): The Internet and Information Retrieval Research: A Brief Review. In: Journal of Documentation 55(2), 209-225

Chu, H. (2003): Information Representation and Retrieval in the Digital Age. Medford, NJ: Information Today

Culliss, G. (2000): The Direct Hit Popularity Engine Technology. A White Paper. http://web.archive.org/web/20010619013748/www.directhit.com/about/products/technology_whitepaper.html [10.2.2004]

Dean, J. (2004): Google: A Behind-the-scenes Look. Presentation at the University of Washington 21 October 2004. http://norfolk.cs.washington.edu/htbin-post/unrestricted/colloq/details.cgi?id=274 [26.11.2004]

Dean, J. A.; *Gomes*, B.; *Bharat*, K.; *Harik*, G.; *Henzinger*, M. (2002): Methods and Apparatus for employing Usage Statistics in Document Retrieval/Google Inc. US Patent Application Nr. US2002/0123988 A1

Eastman, C. M.; *Jansen*, B. J. (2003): Coverage, Relevance, and Ranking: The Impact of Query Operators on Web Search Engine Results. In: ACM Transactions on Information Systems 21(4), 383-411

Egghe, L. (2004). A universal method of information retrieval evaluation: The „missing" link M and the universal IR surface. In: Information Processing and Management, 40(1), 21-30

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

Gelernter, J. (2003): At the Limits of Google: Specialized Search Engines. In: Searcher 11(1), 26-31

Griesbaum, J. (2003): Unbeschränkter Zugang zu Wissen? Leistungsfähigkeit und Grenzen von Suchdiensten im Web. Zwischen informationeller Absicherung und manipulierter Information. In: *Schmidt* , R. (Hrsg.), Competence in Content. Proceedings 25. Online-Tagung der DGI, Frankfurt am Main, 37-50

Griesbaum, J. (2004): Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de. Information Research 9(4) paper 189. http://informationr.net/ir/9-4/paper189.html [3.8.2004]

Griesbaum, J., *Rittberger*, M., *Bekavac*, B. (2002): Deutsche Suchmaschinen im Vergleich: AltaVista.de, Fireball.de, Google.de und Lycos.de. In: *Hammwöhner*, R., *Wolff*, C., *Womser-Hacker*, C. (Hrsg.): Information und Mobilität. Optimierung und Vermeidung von Mobilität durch Information. Proceedings des 8. Internationalen Symposiums für Informationswissenschaft., 201-223

Griesbaum, J.; *Bekavac*, B. (2004): Web-Suche im Umbruch? Entwicklungstendenzen bei Web-Suchdiensten, in: Bernard *Bekavac*, Josef *Herget*, Marc *Rittberger* (Hg.): Information zwischen Kultur und Marktwirtschaft. Proceedings des 9. Internationalen Symposiums fur Informationswissenschaft (ISI 2004), Chur, 6.-8.Oktober 2004, 283-299

Hamilton, N. (2003): The Mechanics of a Deep Net Metasearch Engine. http://turbo10.com/papers/deepnet.pdf [11.3.2004]

Heinisch, C. (2003): Suchmaschinen des Surface Web als Promotoren für Inhalte des Deep Web – Wie Doorway-Pages als ªTeaser´ zu Datenbank-Inhalten in die Index-Files der Suchmaschinen gelangen. In: *Schmidt*, R. (Hrsg.), Competence in Content. Proceedings 25. Online-Tagung der DGI, Frankfurt am Main, 13-24

Henzinger, M., *Motwani*, R., *Silverstein*, C. (2002): Challenges in Web Search Engines. SIGIR Forum 36. http://www.acm.org/sigs/sigir/forum/F2002/henzinger.pdf [18.3.2004]

Hock, R. (2004): The Latest Field Trip: an Update on Field Searching in Web Search Engines. Online 28(5), 15-21

Huang, L. (2000): A Survey On Web Information Retrieval Techniques. http://citeseer.ist.psu.edu/cache/papers/cs/16461/http:zSzzSzwww.ecsl.cs.sunysb.eduzSztrzSzrpe8.pdf/huangoosurvey.pdf [18.3.2004

Karzauninkat, S. (2003): Die Suchmaschinenlandschaft 2003: Wirtschaftliche und technische Entwicklungen. In: Machill, M.; Welp, C. (Hrsg.): Wegweiser im Netz, S. 509-538

Kleinberg, J. (1999): Authoritative Sources in a Hyperlinked Environment. In: Journal of the ACM 46(5), 604-632

Lewandowski, D. (2003): Suchmaschinen-Update: Markttrends und Entwicklungsperspektiven bei WWW-Universalsuchmaschinen. In: *Schmidt*, R. (Hrsg.), Compentence in Content. Proceedings 25. Online-Tagung der DGI,. Frankfurt am Main: DGI, 25-35

Lewandowski,, D. (2004a): Abfragesprachen und erweiterte Funktionen von WWW-Suchmaschinen. In: Information: Wissenschaft und Praxis 55(2), 97-102

Lewandowski, D. (2004b): Datumsbeschränkungen bei WWW-Suchanfragen: Eine Untersuchung der Möglichkeiten der zeitlichen Einschränkung von Suchanfragen in den Suchmaschinen Google, Teoma und Yahoo. In: *Bekavac*, B.; *Herget*, J.; *Rittberger*, M.: Information zwischen Kultur und Marktwirtschaft: Proceedings des 9. Internationalen Symposiums für Informationswissenschaft (ISI 2004), Chur, 6.-8. Oktober 2004, S. 301-316

Machill, M.; *Welp*, C. (Hrsg.) (2003): Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen. Gütersloh: Verlag Bertelsmann Stiftung

Mandl, T. (2003). Neuere Entwicklungen bei der Evaluierung von Information Retrieval Systemen: Web- und Multimedia-Dokumente. In: Information: Wissenschaft und Praxis, 54(4), 203-210

Mandl, T. (2005): Qualität als neue Dimension im Information Retrieval: Das AQUAINT-Projekt. In: Information: Wissenschaft und Praxis 56(1), 13-20

Marable, L. (2003): False Oracles: Consumer Reaction to Learning the Truth About How Search Engines Work. Report, Consumer WebWatch. http://www.consumerwebwatch.org/news/searchengines/ContextReport.pdf [18.3.2004]

Notess, G. (2001): Freshness Issues and Complexities with Web Search Engines. In: Online, 25(6), 66-68

Notess, G. (2003a): Search Engine Statistics: Database Total Size Estimates. http://www.searchengineshowdown.com/stats/sizeest.shtml [7.7.2004]

Notess, G. (2003b): Search Engine Statistics: Relative Size Showdown. http://www.searchengineshowdown.com/stats/size.shtml[6.7.2004]

Notess, G. (2003c): Search Engine Statistics: Freshness Showdown. http://www.searchengineshowdown.com/stats/freshness.shtml [6.7.2004]

Notess, G. (2004): Search Engine Features Chart. http://www.searchengineshowdown.com/features/ [19.7.2004]

Ntoulas, A.; *Cho*, J.; *Olston*, C. (2004): What's New on the Web? The Evolution of the Web from a Search Engine Perspective. Proceedings of the Thirteenth WWW Conference, New York, USA. http://oak.cs.ucla.edu/~ntoulas/pubs/ntoulas_new.pdf [25.3.2004]

Page, L., *Brin*, S., *Motwani*, R., *Winograd*, T. (1998): The PageRank citation ranking: Bringing order to the Web. http://dbpubs.stanford.edu:8090/pub/1999-66 [26.10.2004]

Rasmussen, E. M. (2003): Indexing and Retrieval for the Web. Annual Review of Information Science and Technology 37, 91-124

Richardson, M.; *Domingos*, P.: The Intelligent Surfer: Probabilistic Combination of Link and Content Information in Page Rank. http://www.cs.washington.edu/homes/pedrod/papers/nips01b.pdf [9.8.2004]

Rieh, S. Y. (2004): On the Web at Home: Information Seeking and Web Searching in the Home Environment. In: Journal of the American Society for Information Science and Technology 55(8), 743-753

Savoy, J. (2002): Information Retrieval on the Web: A New Paradigm. Upgrade 3(3), 9-11

Schiff, F. (2003): Business models of news Web sites: A survey of empirical trends and expert opinion. First Monday 8(6), http://firstmonday.org/issues/issue8_6/schiff/index.html [28.11.2004]

Seiffert, F. (2003): Das „Virtuelle Bücherregal NRW": Literatursuche mit der einfachsten Suchstrategie: Google und Co. In: BuB 55(6), 379-397

Sherman, C.; *Price*, G. (2001): The Invisible Web: Uncovering Information Sources Search Engines Can't See. Medford, NJ: Information Today

Singhal, Amit (2004): Challenges in Running a Commercial Search Engine. http://www.research.ibm.com/haifa/Workshops/searchandcollaboration2004/papers/haifa.pdf [15.10.2004]

Spink, A. (2002): A user-centered approach to evaluating human interaction with Web search engines. In: Information Processing & Management 38(3), 410-426

Spink, A. (2003): Web Search : Emerging Patterns. In: Library Trends 52(2), S. 299-306

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

Stock, W. G. (2000): Qualitätskriterien von Suchmaschinen. In: Password 15(5), 22-31

Stock, W. G. (2003): Weltregionen des Internet: Digitale Informationen im WWW und via WWW. In: Password Nr. 16(2), 26-28

Sullivan, D. (2003): Search Engine Sizes. http://searchenginewatch.com/reports/article.php/2156481 [2.7.2004]

Thelwall, M.; *Vaughan*, L. (2004): New versions of PageRank employing alternative Web document models. ASLIB Proceedings 56(1), 24-33

Vaughan, L. (2004): New measurements for search engine evaluation proposed and tested. In: Information Processing and Management 40(4), 677-691

Vaughan, L.; *Thelwall*, M. (2004): Search Engine Coverage Bias: Evidence and Possible Causes. In: Information Processing & Management, 40(4), 693-707.

Veritest (2003): Inktomi Corp.: Web Search Relevance Test. http://www.veritest.com/clients/reports/inktomi/inktomi_web_search_test.pdf [19.10.2004]

Xie, H. (2004): Online IR system evaluation: online databases versus Web search engines. In: Online Information Review 28(3), 211-219