Technologie-Trends im Bereich der WWW-Suchmaschinen

Dirk *Lewandowski*, Düsseldorf

1 Einleitung

Die Suchmaschinen-Szene wurde in den Jahren 2003 und 2004 vor allem durch zwei große Themen auf Trab gehalten: einerseits durch die Aufkäufe und Fusionen (vgl. Lewandowski 2003a, Lewandowski 2003b), andererseits durch den Kampf der Suchmaschinen-Betreiber gegen Spam, also unerwünschte Inhalte in ihren Datenbanken.

Die Datenbanken wurden zunehmend mit werblichen, „inhaltslosen" Seiten überflutet. Dieses grundsätzliche Problem war schon länger bekannt, neu war jedoch im Jahr 2003 der enorme Zuwachs und die Tatsache, dass nun auch die Publikumspresse darauf aufmerksam wurde (vgl. u.a. Karzaunikat 2003). Im Jahr 2003 gaben 82 Prozent der deutschen Suchmaschinen-Betreiber an, dass das Spam-Problem im Vergleich zum letzten Jahr zugenommen habe (Machill, Welp 2003, 82).

Zu den technischen Details der Spam-Bekämpfung ist wenig bekannt, da die Suchmaschinen ein berechtigtes Interesse haben, dass ihre Methoden nicht bekannt werden; deren Kenntnis könnte wiederum leicht dazu genutzt werden, um Spam zu verbreiten.

Trotz dieses großen Problems haben die Suchmaschinen-Betreiber auch im Jahr 2003 (und natürlich auch im laufenden Jahr) die Weiterentwicklung ihrer Angebote nicht vernachlässigt. Diese Technologie-Trends sollen im vorliegenden Aufsatz diskutiert werden.

Für deren Identifizierung wurden nicht nur die neuen Funktionen bei den etablierten Suchmaschinen beobachtet, sondern auch Veröffentlichungen, Patente und Patentanmeldungen ausgewertet. Weiter wurden die neu am Markt auftretenden Suchmaschinen berücksichtigt und Aussagen von Experten (z.B. Beal 2004, Seuss 2004) ausgewertet. Aus dieser umfassenden Auswertung konnten vier Trends abgeleitet werden, die die Entwicklung der Suchtechnologie auf kurze bis mittlere Frist kennzeichnen:

1. **Desktop-Suche**: Die Internet-Suche wird nicht mehr nur über das Ansteuern einer Suchmaschine oder die Verwendung einer zuvor installierten Toolbar möglich sein.
2. **Informationsressourcen**: Suchmaschinen werden nicht mehr nur Web-Dokumente finden, sondern auf verschiedene Datenbestände zurückgreifen und auch auf Informationsressourcen hinweisen.
3. **Lokalisierung**: Neben der regulären Suche wird eine lokale Suche möglich sein.
4. **Personalisierung**: Suchmaschinen werden ihre Ergebnisse dem einzelnen Nutzer bzw. einer Nutzergruppe anpassen.

Im Text werden zu den jeweiligen Technologien auch Anwendungsfälle bei den Suchmaschinen vorgestellt. So wird auch ein Überblick über einige neue Funktionen und deren Implementierung gegeben. Ausdrücklich nicht berücksichtigt werden allgemeine Entwicklungen im Information Retrieval, die *auch* dazu dienen können, die Qualität der Suchmaschinen zu verbessern.

In einem Ausblick werden noch einmal die grundlegenden Probleme der WWW-Suchmaschinen angesprochen, und es wird gezeigt, wo die Suchmaschinen-Forschung und -Entwicklung im Jahr 2004 angekommen ist.

2 Desktop-Suche

Die bisherige Entwicklung der Suche im Internet lässt sich in Hinblick auf die Stellung der Suchfunktion im Arbeitsumfeld des Nutzers in vier Phasen unterteilen (vgl. Abbildung 1).

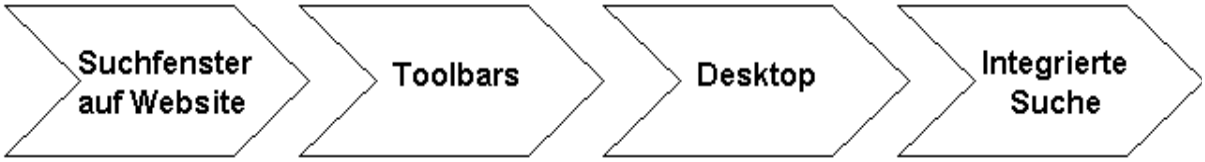

Abbildung 1: Entwicklung des Zugriffs auf Suchmaschinen durch den Benutzer

Traditionell greift der Nutzer auf eine Suchmaschine zu, indem er die betreffende Website der Suchmaschine anwählt und dort seine Suchanfrage in ein Formular einträgt, welches aus einem oder mehreren Suchfeldern besteht. In einem zweiten Schritt wurden sog. Toolbars entwickelt, also Browser-Plugins, die eine Nutzung der jeweiligen Suchmaschine möglich machen, ohne deren Website ansteuern zu müssen. Für die Suchmaschinenbetreiber sind die Toolbars insbesondere aus Aspekten der Kundenbindung interessant. Manche Toolbars bieten aus diesem Grund auch Funktionen, die nicht zur eigentlichen Suche gehören (wie z.B. einen Pop-Up-Blocker) aber einen weiteren Anreiz bieten, die betreffende Toolbar zu installieren. Mittlerweile bieten alle wichtigen Suchmaschinen Toolbars an, weiterhin gibt es Toolbars, die auf mehrere Suchmaschinen zugreifen können (vgl. Sullivan 2004, Notess 2004). Als Nachteile von Toolbars sind zu nennen, dass sie vom Benutzer installiert werden müssen und in der Regel nur für einzelne Browser-Typen verfügbar sind (meist nur für den Internet Explorer unter Windows).

Während Toolbars vom Benutzer installiert werden müssen, besitzen manche Browser bereits Suchfelder, die auf eine oder mehrere ausgewählte Suchmaschinen zurückgreifen.[1] Zwar wer-

1. So zum Beispiel Opera (www.opera.com) und Safari (http://www.apple.com/safari/). Opera bietet die Suche in unterschiedlichen Suchmaschinen und E-Commerce-Marktplätzen; Safari enthält ein Suchfenster für Google.

den hier im Gegensatz zu den Toolbars nicht die vollen Funktionalitäten der Suchmaschinen geboten, dafür ist die Integration der Suche in das tägliche Arbeitsumfeld des Nutzers einen Schritt weiter gekommen: Die Suchfunktion ist „schon da“ und braucht nicht erst installiert zu werden. Eine ähnliche Integration einer Suchfunktion findet in den neuen Versionen von Microsoft Office statt: hier werden allerdings die kostenpflichtigen Dienste von Factiva, Genios, usw. abgefragt (Quint 2003).

Mit dem im Jahr 2003 angekündigten Einstieg von Microsoft in den Suchmaschinenmarkt (bzw. mit der Ankündigung, nun eine eigene Suchtechnologie für das MSN-Portal entwickeln zu wollen) begannen Spekulationen über eine von Microsoft angestrebte Integration der Suchfunktion auf dem Desktop. Nutzer hätten damit die Möglichkeit, direkt von ihrer Benutzeroberfläche aus die Websuche zu starten, ohne extra einen Browser starten zu müssen, die Suche wäre an prominenter Stelle integriert. Google reagierte auf diese Spekulationen mit einer eigenen „Deskbar“ für das Windows-Betriebssystem. Das Programm muss wiederum vom Nutzer installiert werden und integriert ein Suchfenster ähnlich der Toolbar in die Taskleiste von Windows. Die Suchergebnisse erscheinen in einem eigenen Fenster; der Start des Browsers wird erst nötig, wenn ein Ergebnis tatsächlich angezeigt werden soll.

Als letzte Stufe der Integration der Suche in das Arbeitsumfeld des Nutzers ist die tatsächliche integrierte Suche zu nennen. Nutzer suchen nicht nur nach Dokumenten, die im Web frei verfügbar sind, sondern auch nach kostenpflichtigen Inhalten, im Firmenintranet und nicht zuletzt in ihren eigenen Dokumenten, seien dies Office-Dokumente oder E-Mails. Bisher ist für jeden Inhaltstyp eine eigene Suchanfrage zu starten, dazu kommt das Problem der wenig komfortablen und funktional eingeschränkten Suche innerhalb der Betriebssysteme wie Windows oder MaxOS, da diese keinen invertierten Index der Dokumente erstellen.

Seit März 2004 bietet HotBot eine Toolbar an, die nicht nur das Web durchsucht, sondern auch E-Mails und Dateien auf dem Rechner des Benutzers durchsuchen kann[2]. Diese werden in regelmäßigen Abständen indexiert, so dass die Suche weniger zeitaufwendig ist als die in den in Windows bzw. Outlook integrierten Suchfunktionen.

Microsoft Research arbeitet an einer ähnlichen Lösung (Dumais et al. 2003), die die Suche nach allen Inhaltstypen unter einer Oberfläche integriert. Besonderes Augenmerk wird auf das *Wieder*finden von Dokumenten gelegt; die Forscher gehen davon aus, dass zwischen 60 und 80 Prozent aller Besuche auf Webseiten Aufrufe solcher Seiten sind, die vom Nutzer schon einmal besucht wurden. Daher sei es von besonderer Bedeutung, diese Seiten so zu erschließen, dass es dem Nutzer möglich ist, gezielt bereits bekannte Dokumente wiederzufinden.

Es ist davon auszugehen, dass sich die Suche auf dem Desktop bzw. in dem weiteren Arbeitsumfeld des Nutzers (z.B. auf Handhelds und Mobiltelefonen) etablieren wird. Besonders

2. Die HotBot-Toolbar funktioniert allerdings nur unter Windows mit Internet Explorer und Microsoft Outlook oder Outlook Express.

interessant ist die Möglichkeit, unterschiedliche Informationsquellen in die Suche zu integrieren. Dies könnten einerseits persönliche Informationsquellen sein (wie etwa die eigenen E-Mails) oder aber Informationsquellen, die nur für einen bestimmten Nutzerkreis zugänglich sind. Vorbild sind hier die Firmenlösungen der Suchmaschinen-Anbieter (vgl. z.B. Lervik 2004), die unter einer Oberfläche neben der Websuche weitere Quellen integrieren, die der Firma zugänglich sind. Dies könnten u.a. Inhalte des Firmen-Intranets, abonnierte Datenbanken und News-Feeds sein. Ähnliches lässt sich auf privater Ebene realisieren, indem dem Nutzer zugängliche Datenbanken mit der Websuche und persönlichen Dokumenten zusammen abgefragt werden.

3 Informationsressourcen

Die traditionelle Aufgabe von Suchmaschinen ist der Nachweis von Dokumenten. Hier unterscheiden sie sich auch erheblich von den Web-Katalogen bzw. Verzeichnissen. Diese weisen traditionell Informationsressourcen nach. Während also die Suchmaschinen im Idealfall das gewünschte Dokument selbst ausgeben, weisen die Verzeichnisse auf einen neuen Sucheinstieg hin.

Suchmaschinen ist es üblicherweise nicht möglich, Dokumente nachzuweisen, die in Datenbanken abgelegt sind. Dieser Bereich gehört zum sogenannten „Invisible Web“ (Sherman, Price 2001), über dessen Größe unterschiedliche Schätzungen existieren (Bergman 2001, Sherman 2001, Stock 2003). Die genannten Autoren sind sich aber einig, dass die Informationsmenge des Invisible Web die des „surface web“ deutlich übersteigt. Weiterhin ist davon auszugehen, dass die Inhalte des Invisible Web zu einem großen Teil von hoher Qualität sind (Lewandowski 2002, 560). Verzeichnisse haben hier den Vorteil, dass sie Invisible-Web-Quellen nachweisen können, indem sie auf die Einstiegsseiten dieser Quellen (also in der Regel auf die jeweilige Datenbank-Suchmaske) verweisen. Suchmaschinen können diese Einstiegsseiten zwar auch indexieren, bei der Suche gehen solche Seiten aber oft in der Menge der gefundenen Dokumente unter.

Die Suchmaschine Turbo10 bietet eine Möglichkeit, Datenbanken des Invisible Web zu erschließen (Hamilton 2003). Sie ist in der Lage, Datenbank-Suchmasken als solche zu erkennen und Anfragen automatisch an diese Datenbanken weiterzuleiten. Der Benutzer der Suchmaschine kann sich aus den bereits bekannten Datenbanken ein individuelles Portfolio zusammenstellen und seine Suchanfrage an die ausgewählten Datenbanken schicken. Wie bei einer regulären Meta-Suchmaschine werden die unterschiedlichen Ergebnisse neu gerankt und als einheitliche Liste zurückgegeben. So elegant dieser Ansatz das Problem angeht, bestehen auch hier weiterhin zwei große Probleme. Erstens ist die Auswahl auf nur zehn Quellen beschränkt. Da jede Quelle einzeln abgefragt werden muss, wären die Antwortzeiten bei einer hohen Anzahl von zu berücksichtigenden Quellen schlicht inakzeptabel. Zweitens ist es bei Turbo10 nötig, die zu durchsuchenden Quellen bereits zu kennen bzw. diese aus einer hinterlegten Liste

auszuwählen. Der große Vorteil dieser Suchmaschine ist also allein in der gleichzeitigen Abfrage mehrerer bereits bekannter Quellen zu sehen. Allerdings werden auch hier die Eigenheiten und individuellen Abfragemöglichkeiten der einzelnen Datenbanken nicht berücksichtigt, so dass die schon von den Meta-Suchmaschinen bekannten Nachteile bestehen.

Einen anderen Weg als Turbo10 gehen etablierte Suchmaschinen wie Google und Yahoo. Sie bieten einerseits die Suche in unterschiedlichen Datenbeständen, die in der Regel über Registerkarten ausgewählt werden können, an. Neben dieser Möglichkeit, die Suche auf einen Datenbestand zu beschränken, finden sich in vielen Fällen auch Hinweise auf besondere Datenbestände oberhalb der regulären Trefferlisten (Sullivan 2003). Die Integration unterschiedlicher Datenbestände einer Suchmaschine ist in Tabelle 1 am Beispiel von Google dargestellt. Deutlich wird, dass die Suchmaschinen immer mehr Datenbanken zusätzlich zum Kernbestand der textuellen Web-Dokumente unterhalten, um besser auf die Bedürfnisse ihrer Nutzer eingehen zu können. Dabei ist die Integration dieser Bestände in die reguläre Suche von großer Bedeutung, da die Nutzer nur in seltenen Fällen selbst einen bestimmten Bestand für ihre Suche auswählen.

Neben der Suche in eigenen (und ggf. zugekauften) Datenbeständen setzen die Suchmaschinen bei bestimmten Anfragen mittlerweile auch auf die Anzeige von Hinweisen auf Informationsressourcen des Invisible Web. Gibt man zum Beispiel bei Google das Wort „patent“ gefolgt von einer Patentnummer ein, so erfolgt ein Hinweis auf die Datenbank des US Patent and Trademark Office (siehe Abbildung 2).[3] Dabei wird allerdings nicht überprüft, ob unter der eingegebenen Nummer tatsächlich ein Patent vorliegt. Der Hinweis erfolgt bei jeder eingegebenen Nummer, eine Meldung, dass unter dieser Nummer kein Patent existiert, kommt erst nach dem Anklicken des Links direkt aus der Patent-Datenbank.

Google nutzt diese Funktion, um Suchanfragen, die Zahlen enthalten, besser beantworten zu können. Im Fall der Patentanfrage ist auch die Eingabe des fokussierenden Suchbegriffs „patent“ nötig, in anderen Fällen wie beispielsweise bei Telefonvorwahlen reicht die Eingabe der Zahl allein. Neben den genannten Beispielen können auch Paketnummern unterschiedlicher Anbieter, Barcode-Nummern und Flugnummern eingegeben werden.

3. Diese Funktion ist bisher nur in der US-Version von Google verfügbar. Auch alle weiteren Beispiele beziehen sich auf Google.com und sind mit Google.de zur Zeit nicht reproduzierbar.

Tabelle 1: Datenbestände einer Suchmaschine am Beispiel von Google

Datenbank	Inhalt	Quelle	Integrationsform
Web	Textdokumente (HTML, PDF, Word, usw.)	eigener Datenbestand	Standardfunktion
Bilder	Bilder	eigener Datenbestand	Registerkarte
Groups	Newsgroup-Postings	eigener Datenbestand	Registerkarte
Verzeichnis	hierarchischer Website-Katalog	Open Directory Project	Registerkarte, Anzeige oberhalb und ergänzend innerhalb regulärer Trefferlisten
News	Nachrichten aus gecrawlten Online-Publikationen	eigener Datenbestand	Registerkarte, Anzeige oberhalb der regulären Trefferlisten
Froogle	Produkte (Preisvergleich)	eigener Datenbestand	Anzeige oberhalb der regulären Trefferlisten *
Phone Book	US-Telefonnummern aus Telefonbuch und Gelben Seiten	Daten von Fremdanbieter	Anzeige oberhalb der regulären Trefferlisten *
Catalogs	eingescannte Kataloge von Versandhäusern	eigener Datenbestand	keine Integration
Local	Informationen zu lokalen Unternehmen	zugekaufte Informationen aus Gelben Seiten und eigener Bestand an Webseiten	Integration oberhalb der Trefferlisten *
Answers	von „Experten“ beantwortete Fragen	eigener Datenbestand	Hinweis auf Answers, wenn in der regulären Suche keine Treffer gefunden werden *

Print	(unvollständige) bibliographische Informationen und Klappentexte von Büchern	eigener Datenbestand	Integration in die regulären Suchergebnisse
Wörterbuch	Worterklärungen	Dictionary.com	In der Wiederholung der Suchanfrage oberhalb der Trefferliste sind Wörter anklickbar, zu denen Erklärungen vorliegen. Verweis auf Dictionary.com. *
numerische Informationen	u.a. Nachverfolgung von Paketen Patente Flugdaten	autoritative Datenbanken	Verweise auf entsprechende Datenbanken, wenn die Suchanfrage einen fokussierenden Suchbegriff und eine Zahl enthält *

* nur in der US-Version (google.com)

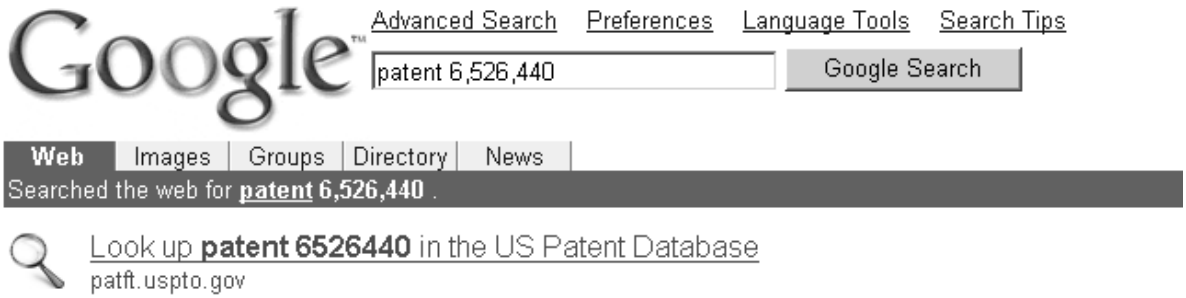

Abbildung 2: Hinweis auf die US-Patentdatenbank bei Google

Ähnliche Möglichkeiten bietet Yahoo.com, wo die fokussierenden Suchbegriffe bereits seit längerem eingesetzt werden. Aufgrund vieler eigener bzw. zugekaufter Daten kann Yahoo! hier aus dem eigenen Angebot schöpfen. So werden bei der Eingabe von „Wetter“ und dem Namen einer Stadt oberhalb der regulären Trefferliste die konkreten Wetterinformationen angezeigt (siehe Abbildung 3). Dabei werden die Daten in diesem Fall direkt eingebunden, in anderen Fällen erfolgt ein Verweis wie im oben genannten Google-Beispiel.

Im März 2004 kündigte Yahoo! an, nun auch Inhalte des Invisible Web direkt integrieren zu wollen (Quint 2004). Neben einem kostenpflichtigen Programm, mit dem kommerzielle An-

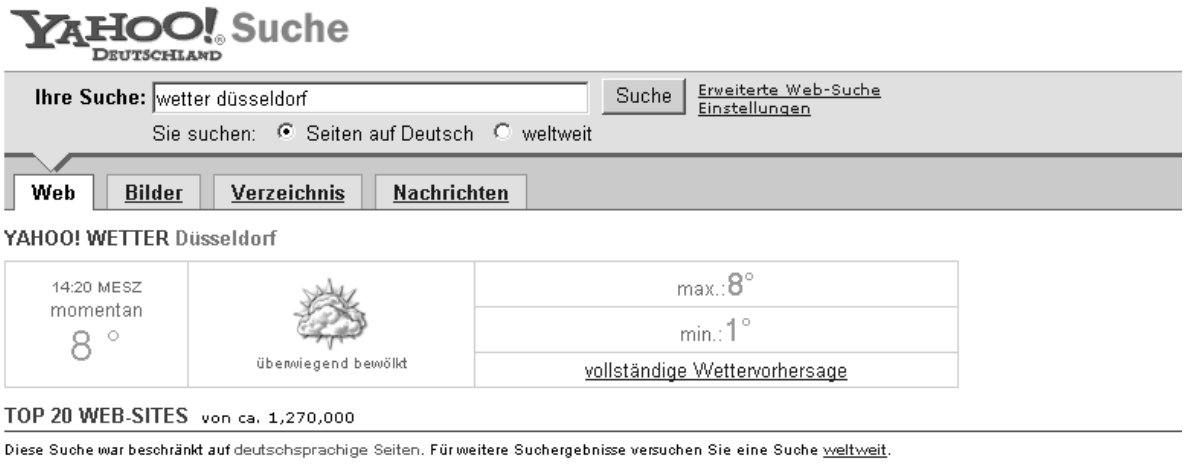

Abbildung 3: Integration von Wetterinformationen in die Ergebnisdarstellung

bieter ihre Datenbank-Inhalte erschließen lassen können, gibt es ein solches auch für Non-Profit-Organisationen. Bisherige Partner, deren Datenbank-Inhalte in die Yahoo!-Suche eingebunden sind, sind unter anderem die Library of Congress, Wikipedia und das Projekt Gutenberg.

Die großen Suchmaschinen-Anbieter fügen ihrem Kernbestand an Web-Dokumenten zunehmend weitere Datenbestände hinzu. Dies ist keine neue Entwicklung, die Datenbestände sind allerdings zunehmend spezialisiert. Teilweise bieten Suchmaschinen originäre (meist zugekaufte) Inhalte, die über andere Suchmaschinen nicht verfügbar sind.[4]

Neben diesen in das eigene Angebot integrierten Datenbeständen verweisen Suchmaschinen zunehmend auch auf weitere relevante Quellen. Dadurch gehen sie den Schritt vom reinen Nachweis von Dokumenten hin zu einem integrierten Nachweis von Dokumenten und Informationsressourcen. Dies stellt eine Möglichkeit dar, die Inhalte des Invisible Web zumindest teilweise zu erschließen. Für die Zukunft ist die Integration solcher Quellen in hohem Maße zu erwarten. Ähnlich wie bei den Internet-Verzeichnissen findet hier zudem eine Qualitätskontrolle statt, so dass nur tatsächlich wertvolle Quellen hinzugefügt und prominent platziert werden.

4 Lokalisierung

Nutzer, die sich für den Kauf eines Produkts oder eine Dienstleistung interessieren, möchten oft Anbieter in ihrer Nähe finden. Solche Suchanfragen können die Suchmaschinen nur schlecht bedienen; werden zur Suchanfrage ergänzende Ortsangaben mit eingegeben, führt

4. Beispiele hierfür sind der Zugriff auf die Corbis-Bilddatenbank unter AltaVista, der Zugriff auf Nachrichten von Nachrichtenagenturen über Yahoo! und der Zugriff auf die Inhalte von mehr als 700 Print-Publikationen über den „Find Articles“-Dienst von Looksmart.

dies oft nicht zu den gewünschten Ergebnissen. Selbst wenn relevante Ergebnisse gefunden werden, ist deren Vollständigkeit nicht garantiert. Im Gegensatz dazu kann sich der Nutzer bei einem Blick in die örtlichen Gelben Seiten bzw. in deren Online-Pendant in der Regel darauf verlassen, eine vollständige Auflistung der am Ort tätigen Unternehmen einer Branche zu bekommen.

Nachdem erste Versuche der lokalen Suche die Informationen der Branchenbücher noch außer Acht ließen, dürfte inzwischen klar sein, dass eine lokale Suchmaschine um diese Informationen nicht herum kommt. Bisher wird die lokale Suche von Google (local.google.com) und in einer anderen Form von Yahoo! (maps.yahoo.com) angeboten. Alle wichtigen Anbieter von Suchmaschinen-Technologie haben jedoch angekündigt, an Lösungen zu arbeiten und diese noch im Laufe des Jahres 2004 präsentieren zu wollen.

Lokale Lösungen arbeiten im Gegensatz zu den regulären Suchinterfaces mit zwei Suchschlitzen; in den einen werden die Suchwörter, in den anderen wird die Ortsangabe eingetragen. Bei der Ergebnispräsentation dürfte sich Folgendes als Standard herausbilden: Als Primärergebnisse werden Einträge aus einem Branchenbuch angezeigt; dabei handelt es sich um von der jeweiligen Suchmaschine zugekaufte Daten, da keine der Suchmaschinen über einen eigenen Index solcher Daten verfügt. Diese Informationen werden jedoch mit Informationen aus dem Web-Index der jeweiligen Suchmaschine angereichert. Dies kann ein Hinweis auf die offizielle Seite des jeweiligen Unternehmens sein, aber auch weiterführende Angaben wie beispielsweise Urteile über das Unternehmen aus Kundensicht. Weiterhin wird zu den Suchergebnissen der Ausschnitt eines Ortsplans angezeigt, auf dermdas Unternehmen eingezeichnet ist.

Die Eingabe des gewünschten Orts in der lokalen Suche dürfte in der Zukunft entfallen bzw. optional werden. Diese Information kann nach einmaliger Eingabe leicht gespeichert und bei zukünftigen Suchanfragen automatisch ausgewertet werden. Besonders interessant wird eine lokale Suche natürlich dann, wenn die Suche über mobile Endgeräte erfolgt. Hier fallen sowieso Positionsdaten des Nutzers an, die für die Suche verwendet werden können. Unterwegs ließe sich also abfragen, wo sich das nächste Hotel oder das nächste libanesische Restaurant befindet. Solche Services können nicht allein von Suchmaschinen angeboten werden, diese versuchen allerdings jetzt schon, diese Marktlücke zu füllen. Grund hierfür dürfte vor allem das erwartete hohe Volumen für lokale Textanzeigen sein. Ähnlich wie bereits bei der regulären Suche ist auch bei der lokalen Suche der Einsatz solcher Anzeigen zu erwarten. Bisher lohnt sich dieser für Unternehmen mit einem nur lokalen Kundenkreis allerdings kaum.

Ansätze der lokalen Suche dürften in nächster Zeit weiter ausgebaut und verbessert werden. Alle lokalen Dienste sind bisher nur für Orte in den USA verfügbar; mit einer Erweiterung auf andere Regionen ist nach der weiteren Verbesserung dieser Dienste zu rechnen.

5 Personalisierung/Social Networks

Schon bei der lokalen Suche kann man von einer gewissen Form der Personalisierung sprechen, vor allem, wenn die Ortsangaben des Nutzers gespeichert werden. Ähnlich verhält es sich auch beim Erkennen des Nutzers durch seine IP-Adresse. Hier werden die Suchergebnisse zwar nicht auf den einzelnen Nutzer zugeschnitten, eine Art Personalisierung der Suchergebnisse findet aber durch die Ausgabe der Suchergebnisse beispielsweise in ausgewählten Sprachen statt (eingesetzt u.a. bei All the Web).

Alle Anbieter von Suchmaschinen-Technologie setzen auf Verfahren der Personalisierung. Es wird erwartet, dass dadurch die Qualität der Suchergebnisse gesteigert werden kann und unpassende Ergebnisse herausgefiltert werden können (vgl. u.a. US 6.539.377 B1). Als Argument wird oft angeführt, dass sich aus dem bisherigen Nutzerverhalten bei einer uneindeutigen Suchanfrage die gewünschte Bedeutung herausfinden ließe. So würde ein Nutzer, der in der Vergangenheit viele Suchanfragen zum Automobilbereich ausgeführt hat, bei einer Suche nach „Jaguar“ bevorzugt Informationen zur Automarke erhalten. Diese Unterscheidung verdeckt jedoch in erster Linie, dass bis heute keine Suchmaschine eine zuverlässige Homonymkontrolle einsetzt.

Wird eine Personalisierungsfunktion implementiert, so kann diese auf verschiedener Datenbasis aufbauen: auf den Daten einer Gruppe von Nutzern oder auf den Nutzerdaten bzw. den Präferenzen eines Einzelnen.

Werden die Daten einer Gruppe ausgewertet, so werden die von einem Nutzer aus den Trefferlisten ausgewählten Seiten als Empfehlung dieser Seiten an die anderen Nutzer gewertet. Dabei wird davon ausgegangen, dass innerhalb der Nutzergruppe die Interessenlage nicht allzu heterogen ist. Führt ein Nutzer eine Suche durch, so werden Seiten, die von anderen Mitgliedern seiner Gruppe positiv bewertet wurden (bzw. überhaupt aus der Trefferliste ausgewählt wurden), bevorzugt angezeigt und markiert. Allerdings ergibt sich bei einer großen Gruppe bzw. bei einer Gruppe mit heterogenen Interessen das Problem, dass die Suchmaschine hier die Anfragen nicht mehr nach dem jeweiligen Interesse unterscheiden kann. Eine eingeschränkte Lösung für dieses Problem bietet die Suchmaschine Eurekster: Hier kann die Suche auf bestimmte „Search Groups“, die jeweils einem Interessengebiet zugeordnet sind, eingeschränkt werden. Die Daten, die für die Personalisierung erhoben werden müssen, können im Fall der gruppenbezogenen Auswertung anonym gespeichert werden.

Bei der Auswertung der Nutzerdaten eines Einzelnen kann genauso verfahren werden, jedoch ist insbesondere bei Portalanbietern wie Yahoo davon auszugehen, dass diese die Nutzerdaten auch für Marketingzwecke speichern möchten. Portale verfügen bereits über eine große Anzahl von Nutzern, von denen sie, da diese sich in der Regel in das Angebot einloggen, eine große Menge an Daten personenbezogen erheben. Hier ist jedoch der Nachteil dieser Form der Personalisierung zu sehen. Die Nutzer müssen dem Angebot vertrauen, um es zu nutzen.

Bisher realisierte Angebote wie Eurekster (www.eurekster.com) und Google Personalized (labs.google.com/personalized) verfolgen unterschiedliche Ansätze. Google setzt auf Voreinstellungen des Nutzers, der seine Interessen in einem persönlichen Profil festlegt, welches in einem Cookie gespeichert wird. Die Interessen können aus einem Kategorienschema ausgewählt werden, welches allerdings sehr allgemein gehalten ist. Google ordnet gefundene Webseiten den Kategorien zu und bewertet die Übereinstimmungen in der Suchanfrage. In der Trefferlistenanzeige kann eingestellt werden, wie stark die Ergebnisse dem persönlichen Profil angepasst werden sollen. Probleme bestehen allerdings, wenn der Nutzer mehrere Interessen in seinem Profil ausgewählt hat. In diesem Fall kommt es oft zu Ergebnissen, die keine Verbesserung gegenüber den regulären Web-Ergebnissen darstellen.

Eurekster verfolgt den gruppenbasierten Ansatz und arbeitet relativ zuverlässig. Die aktuellen Forschungen im Bereich der „social networks“ (vgl. u.a. Huberman 2001) und deren Anwendung im Bereich der Community-Sites lassen diesen Ansatz als den auch für die Suche geeigneteren erscheinen. Ein weiterer Indikator für die Zukunft dieses Ansatzes ist das Interesse der Suchmaschinenbetreiber an derartiger Technologie. So betreibt etwa Google das Angebot Orkut, welches eine Social-Network-Community bietet und laut Aussagen von Google-Vertretern in nächster Zeit in das Stamm-Angebot integriert werden soll.

Eine abschließende Bewertung der Personalisierungsansätze und -angebote ist zum jetzigen Zeitpunkt noch nicht möglich, da schlicht die Erfahrung mit solchen Systemen fehlt. Ein Problem ist auf jeden Fall darin zu sehen, dass die Suchergebnisse durch die Personalisierung nicht mehr vergleichbar sind und auch soziale Netzwerke durch Spam überflutet werden können. Diese Ansätze sorgen also für einen zunehmenden Verlust an Transparenz. Schon heute wird allerdings oft die mangelnde Transparenz der Suchmaschinen beklagt (vgl. Machill, Welp 2003).

6 Ausblick

Die aufgezeigten Trends können dabei helfen, dem Nutzer die Suche zu erleichtern. Allein durch diese Anstrengungen wird sich die Qualität der Suchmaschinen auf Dauer jedoch nicht wesentlich verbessern lassen. Grundsätzliche Probleme wie die Abdeckung des Web, die Art der Dokumentrepräsentation und die eingeschränkten Suchmöglichkeiten (Chu 2003) bleiben bestehen. Anspruchsvolle linguistische Ansätze wie Informationsextraktion, die Verarbeitung natürlicher Sprache und die Beantwortung von Fragen (Chakrabarti 2003) stecken noch in den Kinderschuhen.

Bei den Suchmöglichkeiten (vgl. Lewandowski 2004) sind zwar teils Fortschritte festzustellen, auf der anderen Seite aber auch (wie bei den im März 2004 erfolgten Umstellungen bei All the Web) erhebliche Rückschritte. Von den Ansprüchen professioneller Retrievalsysteme sind die Suchmaschinen immer noch weit entfernt.

Eine Entwicklung, die sich seit Bestehen der WWW-Suchmaschinen fortsetzt, ist die hin zu immer größeren Indexe. Zur Zeit ist Google hier mit etwa vier Milliarden Dokumenten führend, allerdings ist zu erwarten, dass andere Anbieter (insbesondere Yahoo) hier nachziehen werden. David *Seuss*, Gründer und CEO von Northernlight, äußert sich allerdings kritisch über die Entwicklung der Indexgrößen: „The effort to compete on database size is a tremendous waste of intellectual, financial, and human capital" (Seuss 2004). Neue Anbieter am Markt starten zunehmend als Meta-Suchmaschinen, damit sie keinen eigenen Index unterhalten müssen. Weiterhin ist zu fragen, welche Größe der Index einer WWW-Suchmaschine haben sollte, um gute Ergebnisse liefern zu können.

Bei der Dokumentrepräsentation scheint es nur wenige Fortschritte zu geben. Die zu den jeweiligen Dokumenten erfassten Informationen unterscheiden sich bei den Suchmaschinen nicht wesentlich. Zwar wurden Forschungen in diesem Bereich angeregt (Henzinger, Motwani, Silverstein 2002), genießen jedoch offensichtlich keine Priorität in den Entwicklungsabteilungen der Suchmaschinen-Anbieter.

Der Befund über den technologischen Stand der Suchmaschinen im Jahr 2004 ähnelt dem der vergangenen Jahre: Die Entwicklung steht erst am Anfang und für die Zukunft sind große Fortschritte zu erwarten. Dies lässt sich natürlich auch weniger positiv fassen; mit den Worten von David Seuss: „Ten Years Into the Web, and the Search Problem is Nowhere Near Solved." (Seuss 2004)

Literatur

Beal, A.: Ask Jeeves: What's the Future of Search? http://www.sewatch.com/searchday/article.php/3316801 [1.3.2004]

Bergman, M. K.: The Deep Web: Surfacing Hidden Value. The Journal of Electronic Publishing 7 (2001) 2. http://www.press.umich.edu/jep/07-01/bergman.html [13.1.2004]

Chakrabarti, S.: Mining the web: Discovering knowledge from hypertext data. Amsterdam (u.a.): Morgan Kaufmann 2003

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

Dumais, S.; *Cutrell*, E.; *Cadiz*, J. J.; *Jancke*, G.; *Sarin*, R.; *Robbins*, D. C.: Stuff I've Seen: A System for Personal Information Retrieval and Re-Use. SIGIR ,03 (2003). http://research.microsoft.com/~sdumais/ SISCore-SIGIR2003-Final.pdf [11.3.2004]

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

Henzinger, M., *Motwani*, R., *Silverstein*, C.: Challenges in Web Search Engines. SIGIR Forum 36 (2002). http://www.acm.org/sigir/forum/F2002/henzinger.pdf [11.3.2004]

Huberman, B . A.: The Laws of the Web: Patterns in the Ecology of Information. Cambridge Mass.: MIT Press, 2001

Karzaunikat, S.: Google zugemüllt: Spam überschwemmt die Suchergebnisse. In: c't (2003) 20, 88-92

Lervik, J. M.: Search Engine Industry Trends – Impact for Digital Libraries. 7th International Bielefeld Conference 2004. http://conference.ub.uni-bielefeld.de/proceedings/lervik.pps [11.3.2004]

Lewandowski, D.: Alles nur noch Google? Entwicklungen im Bereich der WWW-Suchmaschinen BuB - Forum für Bibliothek und Information 54 (2002) 9, 558-561 [2002]

Lewandowski, D.: Mega-Suchmaschine durch Aufkauf von Overture? Password 14 (2003) 9, 37 [2003a]

Lewandowski, D.: Suchmaschinen-Update: Markttrends und Entwicklungsperspektiven bei WWW-Universalsuchmaschinen. In: Schmidt, R. (Hrsg.): Compentence in Content. 25. Online-Tagung der DGI, Proceedings. Frankfurt am Main: DGI, 2003, S. 25-35 [2003b]
Lewandowski, D.: Abfragesprachen und erweiterte Funktionen von WWW-Suchmaschinen. Information: Wissenschaft und Praxis 55 (2004) 2, 97-102 [2004]
Machill, M.; *Welp*, C. (Hrsg.): Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen. Gütersloh: Verlag Bertelsmann Stiftung, 2003
Notess, G.: Toolbars: Trash or Treasures? Online 28 (2004) 1, 41-44
Quint, B.: Microsoft Office 2003 Opens New Market to Fee-Based Information Services. Information Today Newsbreak 17.3.2003. http://www.infotoday.com/newsbreaks/nb030317-1.shtml [30.3.2004]
Quint, B.: Yahoo! Pursues Inivisible Web Content for its Search Engine. http://www.infotoday.com/newsbreaks/nb040308-1.shtml [17.3.2004]
Seuss, D.: Ten Years Into the Web, and the Search Problem is Nowhere Near Solved. Computers In Libraries Conference, March 10-12, 2004. http://www.infotoday.com/cil2004/presentations/seuss.pps [15.3.2004]
Sherman, C.: Search for the Invisible Web. Guardian Unlimited 6.9.2001. http://www.guardian.co.uk/online/story/0,3605,547140,00.html [5.3.2004]
Sherman, C.; *Price*, G.: The Invisible Web: Uncovering Information Sources Search Engines Can't See. Medford, NJ: Information Today, 2001
Stock, W.: Weltregionen des Internet: Digitale Informationen im WWW und via WWW. Password 14 (2003) 2, 26-28
Sullivan, D.: Search Toolbars & Utilities (2004). http://searchenginewatch.com/links/article.php/2156381 [11.3.2004]
Sullivan, D.: Searching With Invisible Tabs (2003). http://searchenginewatch.com/searchday/article.php/3115131 [1.3.2004]
US 6.539.377 B1, Ask Jeeves, Inc., Personalized Search Methods, 25.3.2003 (USA)