Handbuch Internet-Suchmaschinen
D. Lewandowski (Hrsg.)
© 2009, AKA Verlag Heidelberg

Spezialsuchmaschinen

Dirk Lewandowski
Hochschule für Angewandte Wissenschaften Hamburg
Fakultät Design, Medien und Information
Department Information
Berliner Tor 5
20099 Hamburg
dirk.lewandowski@haw-hamburg.de

Abstract. In diesem Kapitel werden die Ansätze von Spezialsuchmaschinen systematisch dargestellt und ein Überblick über Spezialsuchmaschinen in unterschiedlichen Themenbereichen gegeben. Der Schwerpunkt dieses Überblicks liegt auf den Anbietern von Websuchmaschinen, die zunehmend Spezialsuchen in ihr Angebot integrieren.

Keywords. Suchmaschine, Spezialsuchmaschine, Bildersuchmaschine, Videosuchmaschine, Bibliothekssuchmaschine, Wissenschaftssuchmaschine, Buchsuchmaschine, Nachrichtensuchmaschine, Blogsuchmaschine, Lokale Suche.

Einleitung

In den letzten Jahren ist eine zunehmende Ergänzung der Websuchmaschinen um spezielle Kollektionen und Suchoberflächen für besondere Inhalte festzustellen. Die gesonderten Angebote reichen von Nachrichtensuchmaschinen bis hin zu Suchmaschinen für Musikstücke und Videoclips. Während viele der Inhalte der Spezialsuchmaschinen aus dem Web kommen und damit entweder im regulären Webcrawl mit erfasst werden können oder aber durch einen eigenen Webcrawl erfasst werden, gibt es auch Inhaltstypen, die erst durch die Verbindung von Inhalten des freien Web mit solchen aus Datenbanken sinnvoll durchsuchbar gemacht werden können. So hat sich etwa gezeigt, dass eine lokale Suche ohne strukturierte Daten aus Branchenbüchern und Kartendaten nicht praktikabel ist.

Dieses Kapitel gibt auf der einen Seite einen systematischen Überblick über die Ansätze (und damit verbundenen Probleme) der Spezialsuchmaschinen, auf der anderen Seite werden die Spezialsuchen in einzelnen Themenbereichen genauer betrachtet.

Der Fokus des Kapitels liegt sowohl auf den Interessen der Wissenschaftler, die sich mit Suchmaschinen beschäftigen, als auch auf den Interessen der Information Professionals, für die die Spezialsuchmaschinen eine Ergänzung bei ihren Recherchen in den Websuchmaschinen und in Fachdatenbanken darstellen können. Insofern wird besonders im zweiten Teil auch auf die Recherchefunktionen der einzelnen Angebote eingegangen. Ergänzt wird die Besprechung der Einzelangebote um weiterführende Literatur, die Ansatzpunkte für eine vertiefende Beschäftigung bietet. Die Literaturauswahl ist dabei stark praxisorientiert und hat nicht den Anspruch, den Stand der Forschung bei der Erstellung von Spezialsuchmaschinen aufzubereiten. Vielmehr

geht es um eine Verdeutlichung der jeweils speziellen Probleme in den Themenbereichen in einem Überblicksrahmen, wie er der Länge und dem Überblickscharakter dieses Kapitels entspricht.

Dieser Aufsatz ist folgendermaßen aufgebaut: Im ersten Abschnitt wird die Notwendigkeit von Spezialsuchmaschinen dargestellt, in zweiten Abschnitt erfolgt eine Systematisierung der unterschiedlichen Suchmaschinentypen mit einer Abgrenzung der Spezialsuchmaschinen von den allgemeinen Suchmaschinen und Datenbanken. Der Rest des Artikels widmet sich einem Überblick über Spezialsuchmaschinen in einzelnen Erschließungsbereichen. Das Fazit gibt eine Zusammenfassung der gegenwärtigen Situation und einen Ausblick auf Entwicklungen im Bereich der Spezialsuchmaschinen.

1. Warum sind Spezialsuchmaschinen notwendig?

Für die allgemeinen Web-Suchmaschinen bzw. den Nutzer dieser Suchmaschinen ergeben sich vier Problembereiche, die Spezialsuchmaschinen notwendig machen. Diese werden im Folgenden dargestellt.

1.1. Technische Probleme

Bei den technischen Problemen ist zunächst einmal der Aufwand zu nennen, der nötig ist, um das „komplette“ Web abzudecken. Allgemeine Suchmaschinen versuchen, potentiell alle im Web vorhandenen Dokumente zu erschließen; Ausnahmen sind Spam-Seiten und verbotene Inhalte, die gefiltert werden [1]. Aus der immensen Größe des Web ergeben sich Probleme hinsichtlich der Datenhaltung, aber auch der Aktualität der Datenbestände [2]. Um eine Vorstellung von der Größe des Web – verlässliche Zahlen hierzu liegen nicht vor – zu bekommen, sei nur auf die Indexgrößen der Suchmaschinen verwiesen (welche aber keinesfalls das ganze Web abdecken). Sie geben Indexgrößen von mehreren Milliarden Dokumenten an; Schätzungen sprechen von über 20 Milliarden Dokumenten in den führenden Suchmaschinen.

Selbst wenn eine Suchmaschine nun alle im Web erreichbaren Dokumente indexieren würde, bliebe immer noch der Bereich des sog. Invisible Web [3] mit den nicht durch die Suchmaschinen auffindbaren Dokumenten. Hierbei handelt es sich vor allem um die Inhalte von Datenbanken (ob kostenlos oder kostenpflichtig, siehe [4]). Suchmaschinen sind nicht in der Lage, in den Formularen dieser Datenbanken sinnvolle Abfragen einzugeben, um an die dahinter liegenden Inhalte zu gelangen. Die Schätzungen über die Größe des Invisible Web gehen weit auseinander [5, 6], man kann aber auf jeden Fall davon ausgehen, dass es dem „Oberflächenweb“ ebenbürtig ist.

Im Bereich des Invisible Web sind die kostenpflichtigen Inhalte von besonderer Bedeutung: Heute können die meisten öffentlich zugänglichen elektronischen Informationen über das Web abgerufen werden, jedoch ist eine Authentifizierung nötig. Allgemeine Suchmaschinen lassen diese Inhalte außen vor, da sie für den allgemeinen Nutzer nicht von Interesse sind. Hier setzen Spezialsuchmaschinen an, die gerade diese Inhalte erschließen und oft als sog. Hybridsuchmaschinen mit Inhalten des Oberflächenweb verbinden (siehe Abschnitt 2).

1.2. Finanzielle Hürden

Obwohl hier vornehmlich von technischen Problemen die Rede ist, sind damit auch erhebliche finanzielle Probleme verbunden. Nicht nur die zahlreichen Rechner, die den Index speichern und durchsuchbar machen, kosten Geld, sondern vor allem auch die für den Betrieb einer großen Suchmaschine nötige Bandbreite [7].

1.3. Ausrichtung am „Durchschnittsnutzer"

Allgemeine Suchmaschinen richten sich an einem angenommenen „Durchschnittsnutzer" aus. Die Rankingverfahren der Suchmaschinen berücksichtigen in starkem Maße die Popularität von Dokumenten als Rankingfaktor [1], um dem Nutzer das zu präsentieren, was die Mehrheit der Nutzer wünscht. Dies führt allerdings dazu, dass eher allgemeine, man könnte auch sagen: oberflächliche, Informationen in den Trefferlisten auftauchen. Allerdings ist dieses Vorgehen für allgemeine Suchmaschinen auch sinnvoll, da sie in erster Linie eine große Nutzermasse befriedigen wollen.

Dies schlägt sich auch bei den Einschränkungsmöglichkeiten in der Suche nieder. Zwar bieten alle Suchmaschinen erweiterte Suchformulare und Operatoren, allerdings nicht in dem für den Profi-Rechercheur wünschenswerten Maß [8]. Die Suchfunktionen können bei weitem nicht mit denen professioneller Datenbanken mithalten. Auch hier liegt also ein Ansatzpunkt für Spezialsuchmaschinen, die (für einen beschränkten Datenbestand) gezieltere Einschränkungsmöglichkeiten anbieten können.

Allgemeine Suchmaschinen unterscheiden bei der Präsentation der Ergebnisse nicht nach dem Kenntnisstand der Nutzer. Die alleinig leitende Annahme ist hier, dass fortgeschrittene Nutzer auch spezifischere Anfragen eingeben und dadurch zu den spezielleren Dokumenten gelangen würden. Für die gleiche Anfrage erhalten aber alle Nutzer auch das gleiche Ergebnis in der gleichen Reihenfolge. Abhilfe soll hier die Personalisierung schaffen, die die Ergebnisse entsprechend dem früheren Verhalten eines individuellen Nutzers anpassen soll (siehe den Beitrag von Riemer und Brüggemann in diesem Band).

1.4. Probleme der Erschließung

Aufgrund ihrer Fokussierung auf das ganze Web verwenden allgemeine Suchmaschinen auch nur eine Form der Erschließung. Die Dokumente werden im Volltext erfasst und der Dokumententext in der Regel um die Ankertexte der auf das Dokument verweisenden Dokumente ergänzt. Dadurch ergibt sich eine erweiterte Beschreibung, die teils weit über das ursprüngliche Dokument hinausgeht. Es können Dokumente, die vorwiegend oder ausschließlich aus Bildern bestehen, beschrieben werden, Seiten, die in nicht erschließbaren Dateiformaten vorliegen, erfasst werden, und es können Dokumente mittels fremdsprachiger Bezeichnungen oder Akronymen gefunden werden, sofern diese in den Ankertexten vorkommen (vgl. [1]).

Nachteilig an dieser Erschließungsmethode ist vor allem, dass sie alle Dokumente gleich behandelt. Zwar gibt es durchaus Dokumente, die sinnvolle Erschließungsangaben in den Metatags enthalten, diese werden aber nicht ausgewertet, da Metatags in der Vergangenheit vor allem missbraucht wurden, um Dokumente bei den Suchmaschinen künstlich in den Trefferlisten zu platzieren. Falsche oder

zumindest irreführende Angaben sind auch heute noch an der Tagesordnung, und keine allgemeine Suchmaschine kann es sich erlauben, sich auf solche Angaben zu verlassen. Spezialsuchmaschinen hingegen können sich auf einen zuvor definierten, als zuverlässig erkannten Bereich des freien Web beschränken und von den dort vorhandenen Dokumenten auch die Metaangaben auswerten.

Eine weitere Möglichkeit der Erschließung der Dokumente ergibt sich aus der bereits im Dokumententext wiedergegebenen Erschließung. So bieten Verlage und Fachgesellschaften auf ihren Webseiten zu den dort veröffentlichten Fachaufsätzen in der Regel Schlagworte und klassifikatorische Angaben. Diese werden von den allgemeinen Suchmaschinen aber nicht ausgewertet, können jedoch von Spezialsuchmaschinen verwendet werden. Dazu ist eine Einpassung der Erschließung auf die Struktur der zu erschließenden Angebote nötig, die im Rahmen einer auf eine bestimmte Domäne beschränkten Suchmaschine möglich, für das gesamte Web jedoch als nicht praktikabel anzusehen ist.

1.5. Vorteile der Spezialsuchmaschinen

Aus den beschriebenen Problembereichen der allgemeinen Suchmaschinen ergeben sich die folgenden Vorteile der Spezialsuchmaschinen: Sie beschränken sich thematisch und machen damit eine zielgenaue Recherche möglich. Sie können das Ranking speziell auf die von ihnen erschlossenen Dokumente anpassen, ebenso wie die sachliche Erschließung der Dokumente, die wiederum in das Ranking einfließen kann. Als letzter Punkt ergeben sich Vorteile in der Darstellung der Ergebnisse, welche auch auf den individuellen Zweck der Spezialsuchmaschine sowie auf das Niveau der Zielgruppe angepasst werden kann.

1.6. Unterschiede zu Datenbanken

Aus dem Vorangegangenen mag die Frage entstehen, was eine Spezialsuchmaschine von einer einfachen Datenbank, die vielleicht sogar über eine Web-Schnittstelle zugreifbar ist, unterscheidet. Als erstes wichtiges Unterscheidungsmerkmal ist die *Suchmaschinentechnologie* zu nennen. Dies bedeutet, dass Dokumente mittels Crawling erschlossen werden, also durch einen regelmäßigen Besuch des Crawlers im Volltext erschlossen werden. Eventuell werden die im Crawling ermittelten Dokumente um solche aus strukturierten Datenbanken ergänzt (sog. Hybrid-Suchmaschine). Weiterer Bestandteil der Suchmaschinentechnologie ist das Ranking der Dokumente.

2. Systematisierung

Suchwerkzeuge lassen sich grob nach der Art ihres Indexaufbaus unterscheiden, hier stehen sich manuell erstellter Index (bei Webkatalogen und Social Bookmarking Systemen) und maschinell erstellter Index (bei den algorithmischen Suchmaschinen) gegenüber (siehe den Beitrag von Griesbaum, Bekavac und Rittberger in diesem Band). Betrachtet man nun die allgemein vorherrschende Form von Suchwerkzeugen, nämlich die algorithmischen Suchmaschinen, so kann man diese wiederum in Universalsuchmaschinen, Spezialsuchmaschinen, Archivsuchmaschinen [1] unterteilen.

Universalsuchmaschinen versuchen, potentiell das gesamte Web zu erschließen. Sie kennen keine Grenzen hinsichtlich der Sprachen, Länder oder Themen.

Prominenteste Beispiele sind die Websuchmaschinen Google, Yahoo und Live.com (MSN). Spezialsuchmaschinen hingegen haben selbst auferlegte Beschränkungen hinsichtlich eines der genannten Bereiche. Zwar mögen auch Universalsuchmaschinen beschränkt sein, indem sie ihr Ziel einer vollständigen Web-Abdeckung nicht erreichen (können), diese Beschränkung ist jedoch unfreiwillig.

Archivsuchmaschinen schließlich stellen einen Sonderfall dar: Sie indexieren Web-Dokumente und speichern sie *dauerhaft*, um so einen Zugriff auf im aktuellen Web nicht mehr vorhandene Dokumente zu ermöglichen (vgl. [1]).

Spezialsuchmaschinen im Sinne dieses Aufsatzes sind solche, die sich thematisch oder anhand formaler Dokumentenmerkmale (Bsp. Dateityp) beschränken. Suchmaschinen, die sich auf die Dokumente eines Landes oder einer Sprache beschränken, sind zwar genau genommen Spezialsuchmaschinen, von ihrem Anspruch her (wenn sie thematisch nicht beschränkt sind) aber eher den Universalsuchmaschinen zu vergleichen.

Spezialsuchmaschinen lassen sich wiederum unterscheiden nach den Dokumentkollektionen, die sie erfassen:

1. *Spezialsuchmaschinen für bestimmte Bereiche des Web* erfassen durch sog. „focused crawling“ (vgl. [9]) ähnlich den Universalsuchmaschinen Dokumente aus dem Web. Technisch ist die Erfassung die gleiche wie bei diesen, allerdings arbeiten sie mit sog. „white lists“, also Listen von Webservern, die für den Aufbau des Datenbestands berücksichtigt werden. Alle anderen Server bleiben außen vor, so dass durch die Vorauswahl der Server bereits eine Qualitätskontrolle durchgeführt und ein thematischer Fokus festgelegt wird. Focused crawling kann auch thematisch angelegt werden und die white list um „umgebende Dokumente“ (also solche, die von den Servern der white list aus verlinkt werden) erweitert werden.
   Beispiele für Suchmaschinen, die auf focused crawling basieren, sind Forschungsportal.net (http://www.forschungsportal.net; die white list besteht hier aus den Servern der deutschen Hochschulen und Forschungseinrichtungen) und die U.S. Government Search von Google (http://www.google.com/ig/usgov; die white list besteht hier aus den Servern der US-Regierungseinrichtungen).
2. *Spezialsuchmaschinen für bestimmte Dokumenttypen* können durch ihre Beschränkung die für den jeweiligen Dokumenttyp spezifischen Eigenschaften besser auswerten. So können Nachrichtensuchmaschinen die typische Struktur der Meldungen für ihr Ranking ausnutzen und beispielsweise auch zuverlässig Datumsinformationen erfassen, was bei der Gesamtheit der Web-Inhalte nicht möglich wäre. Bei Suchmaschinen für bestimmte Dokumenttypen kann man auch von einem focused crawling sprechen (wie etwa in dem Beispiel der Nachrichtensuche), allerdings gehen sie über dieses hinaus, indem sie aus dem fokussierten Crawl beispielsweise nur die Bilder extrahieren.
3. *Spezialsuchmaschinen für aktualitätskritische Dokumente* crawlen gezielt bestimmte Dokumenttypen, die zwar auch in einem allgemeinen Crawl berücksichtigt, allerdings nicht in der gewünschten Aktualität erfasst werden könnten oder aber in der Flut der im allgemeinen Web-Crawl gefundenen Dokumente untergehen würden. Beispiel für eine solche Spezialsuche sind die Nachrichten- und die Blogsuchmaschinen. In beiden Fällen handelt es sich um aktuelle Dokumente, die von den Suchmaschinen zeitnah erfasst werden

müssen. Ein gesonderter Crawler besucht diese Seiten in kurzen Abständen und baut einen eigenen, häufig aktualisierten Index auf. Alle bekannten Universalsuchmaschinen haben eigene Nachrichtensuchmaschinen aufgebaut, um diesen gerecht zu werden.

4. *Hybridsuchmaschinen* verbinden gecrawlte Inhalte aus dem freien Web mit solchen aus dem Invisible Web; dies können sowohl kostenlose als auch kostenpflichtige Inhalte sein. Erfasst werden können die Dokumente sowohl durch Crawling (der Suchmaschinen-Crawler bekommt Zugang zu einem ansonsten kostenpflichtigen Bereich des Web; ein Beispiel hierfür ist Google Scholar, siehe Abschnitt 3.4) als auch durch das Einlesen strukturierter Datenbanken.

In der Praxis zeigen sich durchaus Überschneidungen zwischen den genannten Bereichen. So soll diese Unterteilung auch keine Trennschärfe zeigen, sondern mögliche Ansätze für Spezialsuchmaschinen.

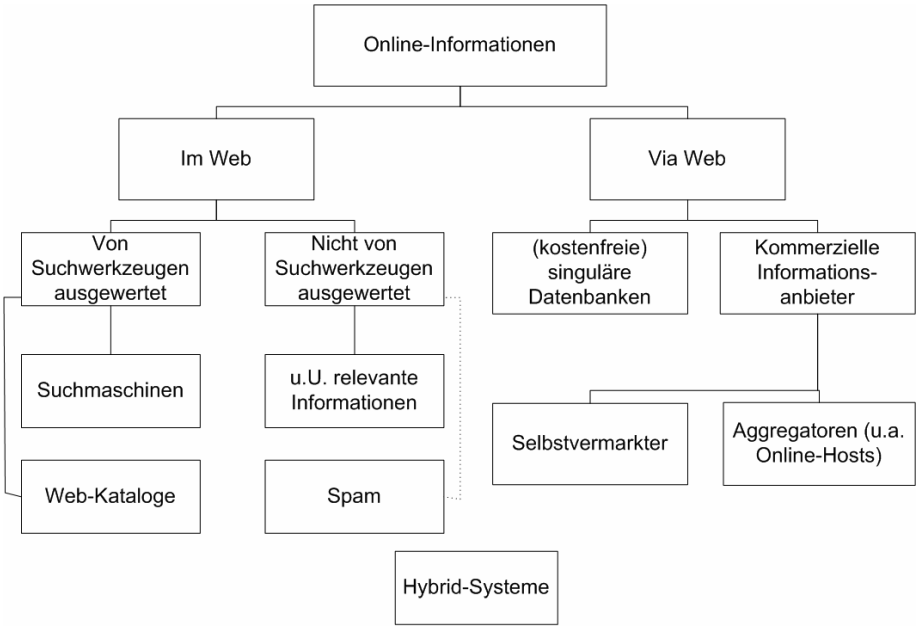

Abbildung 1. Taxonomie der digitalen Online-Information [10]

Abbildung 1 fasst die Unterteilung der Online-Informationen mit den jeweiligen Suchwerkzeugen nach Wolfgang G. Stock [10] zusammen. Unterschieden wird zwischen den Informationen, die im Web stehen und solchen, die nicht direkt im Web stehen, jedoch über das Web erreichbar sind.

Die Informationen im Web werden entweder von (allgemeinen oder spezialisierten) Suchwerkzeugen ausgewertet, es besteht aber stets die Gefahr, dass in dem großen Datenraum nicht alle Informationen gefunden werden. Auf der einen Seite ist diese mangelnde Vollständigkeit problematisch, auf der anderen Seite schützt ein gezielter Ausschluss von Spam-Seiten vor Ballast in den Suchergebnissen.

Die über das Web erreichbaren Informationen werden entweder als kostenfreie Datenbanken (frei zugängliche Inhalte des Invisible Web) oder aber von kommerziellen

Informationsanbietern als Einzeldatenbanken oder aggregiert in einer Sammlung von Datenbanken angeboten. Inwieweit es sich dabei um Spezialsuchmaschinen handelt und inwieweit um (aggregierte) Datenbanken, hängt im Wesentlichen von der dahinter stehenden Technologie ab.

Viele der in der Praxis eingesetzten Spezialsuchmaschinen fallen unter die Kategorie der Hybridsuchmaschinen und verbinden Inhalte aus dem freien Web mit Inhalten des Invisible Web.

3. Überblick Spezialsuchmaschinen

In diesem Abschnitt wird ein Überblick über die wichtigsten Formen von Spezialsuchmaschinen gegeben. Dabei geht es nicht um einen umfassenden Anbieterüberblick; vielmehr soll jeweils ein kurzer Überblick über die besonderen Probleme bei der Erschließung der Inhalte des jeweiligen Bereichs gegeben werden. Dazu werden jeweils einige exemplarische Angebote genannt und auf Besonderheiten bei der Recherche hingewiesen.

So vielfältig wie die Landschaft der Spezialsuchmaschinen ist auch die Zahl und Ausrichtung der Anbieter. Einerseits werden Spezialsuchmaschinen von den Betreibern der großen Universalsuchmaschinen angeboten. So bietet etwa Google etwa einerseits Spezialsuchen, die sich schlicht auf einen Teil des Webbestands beschränken (so die Mac-Suche: http://www.google.de/mac), andererseits spezielle Kollektionen wie die Nachrichten-, Bilder- und Blogsuche.

Neben den großen Suchmaschinenbetreibern bieten auch viele kleinere Unternehmen Spezialsuchen an. Aufgrund des hohen finanziellen und technischen Aufwands für den Betrieb einer Universalsuchmaschine sind Spezialsuchmaschinen nicht nur unter wirtschaftlichen Aspekten für kleinere Unternehmen sinnvoll, sondern sie können auch als Showcases für neu entwickelte Technologie dienen. Begrenzte Kollektionen sorgen auch dafür, dass die im Web allgegenwärtige Spamproblematik (bzw. allgemeiner gefasst die Qualitätsproblematik) weitgehend umgangen werden kann.

Die folgenden Abschnitte geben einen Überblick über die Entwicklungen und die Anbieter in den wichtigsten Bereichen der Spezialsuche.

3.1. Nachrichtensuchmaschinen

Bei Nachrichtensuchmaschinen handelt es sich in der Regel um ein Ergänzungsangebot zur regulären Websuche, das mittlerweile von allen nennenswerten Suchmaschinen angeboten wird. Nachrichtensuchmaschinen können als eine direkte Reaktion auf die gravierenden Aktualitätsmängel der allgemeinen Suchmaschinen angesehen werden. Deutlich wurden diese Mängel besonders, als nach den Terroranschlägen vom 11. September 2001 auch Nachrichten im Web sehr stark nachgefragt wurden. Dieser Bedarf konnte von den Suchmaschinen nicht oder nur unzureichend befriedigt werden [11]. In der Folge wurden daher eigene Nachrichtenkollektionen aufgebaut, die zum einen durch die Quellenauswahl definierten, welche Websites als Nachrichtenangebote anzusehen sind und zum anderen diese Websites in kurzen Zeitintervallen (wenige Minuten) erfassten. Heute sind Nachrichtensuchmaschinen nicht nur eigenständig zu nutzen, sondern ihre Ergebnisse werden bei einer normalen Websuche bei Bedarf auch oberhalb der Trefferliste angezeigt (siehe dazu den Beitrag von Lewandowski und

Höchstötter in diesem Band) oder direkt in die Trefferliste eingebunden (siehe dazu den Beitrag von Quirmbach in diesem Band).

Die bekanntesten Angebote der Nachrichtensuche sind Google News und Yahoo News. Google News bietet neben der eigentlichen Suche auch eine automatisch generierte Überblicksseite, die sich an den Ressorts konventioneller Zeitungen orientiert und damit als eine direkte Konkurrenz zu den Online-Zeitungen im Wettbewerb um Aufmerksamkeit anzusehen ist. Allerdings darf nicht vergessen werden, dass gerade die Online-Angebote konventioneller Medien einen beträchtlichen Teil ihrer Zugriffe über (Nachrichten-)Suchmaschinen und insbesondere über Google bekommen. Yahoo News präsentiert sich einerseits als Nachrichtensuchmaschine, die ebenso wie Google News die Angebote externer Nachrichtenseiten durchsucht, andererseits aber auch als Nachrichtenportal mit zugekauften Meldungen von Nachrichtenagenturen, die in dieser Form nicht mittels anderer Nachrichtensuchmaschinen zu finden sind.

Mit Nachrichtensuchmaschinen lassen sich in der Regel die Meldungen der ausgewerteten Online-Angebote der letzten dreißig Tage finden. Auf der einen Seite durchsuchen sie also nur die Meldungen, die tatsächlich online erschienen sind (und nicht die Archive der Print-Ausgaben), auf der anderen Seite bieten sie keine Archivrecherche. Dem begegnen die beiden genannten Anbieter mit gesonderten Angeboten: Google bietet ein „News Archive“ [12], welches sich sowohl Nachrichtenquellen aus dem freien Web bedient als auch den Inhalten einiger kostenpflichtiger Angebote. Im Rahmen der „Yahoo Search Subscriptions“ können kostenpflichtige Archive wie Factiva und Lexis-Nexis durchsucht werden. Für beide Angebote gilt jedoch, dass sie weder hinsichtlich der Suchfunktionen noch hinsichtlich der Abdeckung [13] (welche in beiden Fällen unklar ist) mit den Angeboten der Nachrichtenhosts wie Genios, Factiva oder Lexis-Nexis mithalten können.

Die Nachrichtenabdeckung ist generell als ein Problem anzusehen. Keine der bekannten Nachrichtensuchmaschinen bietet eine öffentliche Quellenliste an. Um im professionellen Kontext aber zu überprüfen, ob die Recherche überhaupt vollständig sein kann, wäre dies dringend nötig. Dies gilt insbesondere für die Archivsuche, bei der unklar bleibt, welche Zeitungsarchive wie weit zurückgehen. Als weiterer Punkt zu den Quellen ist schließlich noch anzusprechen, dass durch die Intransparenz der Quellenauswahl auch unklar bleibt, inwieweit Pressemeldungen bzw. PR-Artikel mit in den Ergebnissen auftauchen. Schließlich ist festzustellen, dass sich die Nachrichtenangebote im Netz sprachlich zunehmend an die Suchmaschinenalgorithmen anpassen, um optimal aufgefunden zu werden [14].

3.2. Blogsuchmaschinen

Bei den Artikeln aus Blogs verhält es sich auf den ersten Blick ähnlich wie bei den Nachrichten: Es geht hier um eine zeitnahe Erfassung und Verfügbarmachung der Einträge; bei keiner anderen Form dürfte die Geschwindigkeit, mit der neue Einträge erstellt und Diskussionen entfacht werden, größer sein.

Während die Indexierung neuer Artikel auf Nachrichtenwebsites dem konventionellen Crawlingansatz folgt (die Suchmaschine legt fest, wie häufig ein Angebot nach neuen Artikeln durchsucht werden soll, folgt den gefundenen Links und indexiert die so gefundenen HTML-Dokumente), ist es bei Blogs üblich, dass beim Einstellen eines neuen Beitrags von der Blogsoftware automatisch bestimmte Server durch einen sog. *ping* über den neuen Artikel informiert werden. So können

Blogsuchmaschinen neue Artikel ohne Zeitverlust indexieren und sind nicht auf ein eigenständiges Auffinden angewiesen.

Nicht nur aus der Geschwindigkeitsproblematik ergibt sich eine Notwendigkeit für Blogsuchmaschinen; die Trennung von den Nachrichtensuchmaschinen hat ihre Gründe eher in der unterschiedlich einzuschätzenden Zuverlässigkeit der Einträge. Während Nachrichtensuchmaschinen zumindest dem Anspruch nach eine Qualitätsbewertung durch die Auswahl der zu indexierenden Quellen vornehmen (auch wenn diese problematisch ist, siehe Abschnitt 3.1), können Blogs potentiell von jedem Internetnutzer geschrieben werden und folgen meist nicht den etablierten journalistischen Qualitätsstandards [15].

Bei der Indexierung von Blogeinträgen kommen den Suchmaschinen die durch die gängigen Werkzeuge zur Erstellung von Blogs (Wordpress, Blogger, usw.) vorgegebenen Strukturen zugute. So lassen sich neben Überschriften auch die Autorennamen und das Datum der Erstellung eines Artikels extrahieren. Zusätzliche Merkmale wie die Popularität des Blogs (oder eines einzelnen Eintrags, gemessen beispielsweise anhand der Anzahl der abgegebenen Kommentare) lassen sich ebenso in das Ranking integrieren. Die bei der Websuche selbstverständlichen linktopologischen Verfahren dagegen lassen sich allerdings nur in einer abgewandelten Form anwenden, die den Schwerpunkt auf das schnelle Wachstum bzw. die schnelle Etablierung von Verlinkungen (durch die sog. *trackbacks*, also die Verlinkung eines Artikels von anderen Blogs aus) legt.

Ebenso wie die Verlinkung im „normalen Web“ durch Suchmaschinenoptimierer manipuliert wird, um bestimmte Seiten in den Trefferlisten nach oben zu bringen, werden auch die Verlinkungen und *trackbacks* von Blogs manipuliert, um in den Blogsuchmaschinen nach oben zu kommen bzw., allgemeiner gefasst, den eigenen Beiträgen eine gewisse Prominenz zu verleihen. Dies geht bis zu sog. *Splogs* (aus Spam und Blog), deren alleiniges Ziel die Simulation einer „diskussionsstarken“ Verlinkung ist.

Blogsuchmaschinen haben sich zuerst unabhängig von den bekannten Websuchmaschinen entwickelt, diese haben dann aber schnell eigene Angebote geschaffen. Aufgrund der geringeren Indexierungsproblematik sind die Unterschiede zwischen den Angeboten bei weitem nicht so groß wie bei der konventionellen Websuche. Auch aufgrund des technisch weit geringeren Anspruchs an solche Suchmaschinen wurde eine Vielzahl kleinerer Angebote erstellt, die durchaus mit den großen konkurrieren können.

Ein tiefgehender Überblick über Blogsuchmaschinen findet sich in [16]. Für die Recherche sind Blogs als zusätzliche Quellen zu den etablierten Nachrichtenmedien zu empfehlen [17], unter anderem, weil sie schneller auf Ereignisse reagieren können, zum anderen durch die Möglichkeit einer sehr kleinteiligen Berichterstattung (d.h. die Nachricht wird von den etablierten Medien als (noch) nicht bedeutend genug eingeschätzt), beispielsweise auch durch einen sehr starken lokalen Bezug [18].

3.3. Lokale Suche

Unter Lokaler Suche wird die Suche in der Umgebung des Nutzers verstanden [19], also beispielsweise die Suche nach einem nahe gelegenen Restaurant. Vor allem in Hinblick auf die zunehmend aufkommende mobile Nutzung von Internetdiensten ist die lokale Suche von zentraler Bedeutung für die zukünftige Akzeptanz von Suchanbietern.

Während die lokale Suche auf mobilen Endgeräten wie von anderen Suchmaschinen gewohnt aus einem Suchfeld besteht (der Standort des Nutzers kann durch das Gerät in der Regel automatisch ermittelt werden), hat sich bei der lokalen Suche auf dem PC eine Suche mittels zweier Suchfelder durchgesetzt, eines für den Suchbegriff, eines für die Ortsangabe.

Der Aufbau der Ergebnisseiten der lokalen Suche unterscheidet sich von der konventionellen Ergebnispräsentation vor allem durch die Anzeige der Ergebnisse auf einer Karte. Dabei wird der passende Kartenausschnitt gewählt, um eine bestimmte Anzahl relevanter Ergebnisse auf einen Blick zu visualisieren.

Bei der lokalen Suche wird noch stark von einer Suche entweder direkt nach Adressen oder Wegen (Routenplanung) ausgegangen oder aber von der Suche nach den Anbietern eines Produkts oder einer Dienstleistung in der gewünschten Umgebung. Dementsprechend basieren die gängigen lokalen Suchdienste auf strukturierten Branchendaten, die in der Regel von Branchenbuchanbietern zugekauft werden. Der von einigen Suchmaschinen ursprünglich verfolgte Ansatz, lokale Daten allein aus dem Web zu extrahieren, wurde inzwischen aufgegeben. Allerdings werden die Branchenbucheinträge durchaus um Informationen aus dem freien Web, dazu um Informationen aus Bewertungsportalen ergänzt. Ein drittes Element der lokalen Suche sind die Kartendaten, die von den Suchmaschinenbetreibern auch nicht selbst erstellt, sondern zugekauft werden. Insofern liegt ein großer Teil der Leistung der lokalen Suchmaschinen darin, die Daten aus diesen drei Bereichen sinnvoll zu kombinieren. Weitere Ergänzungen wie die Einbindung von lokalisierten Wikipedia-Einträgen und nutzergenerierten Fotos bei Google Maps sind optional.

Lokale Suchdienste werden sowohl von den großen Suchmaschinen als auch von den bekannten Portalen (wie Web.de, T-Online) angeboten [20]. Während bei der konventionellen Web-Recherche nicht zur Benutzung der Portale geraten werden kann (da sie keine eigenen Ergebnisse liefern, sondern ihre Treffer bei einer der großen Suchmaschinen zukaufen), haben die Portale teils eigene lokale Angebote aufgebaut, die entsprechend konkurrenzfähig sind.

Ein technisches Problem der lokalen Suchdienste liegt in der Rankingfunktion: Sie müssen auf Basis der Anfragen herausfinden, wie genau die Anfrage zu erfüllen ist und welcher maximale Weg für den Nutzer akzeptabel ist. Dabei ist zu beachten, dass ein Nutzer für ein seltenes Gut bereit sein dürfte, eine größere Entfernung zurückzulegen, als für Allerweltsdinge bzw. –dienste. Die Bestimmung der geeigneten Entfernung bleibt jedoch weiterhin eine Herausforderung [21].

3.4. Wissenschaftssuchmaschinen

Spezielle Suchmaschinen für den Wissenschaftsbereich werden von unterschiedlichen Unternehmen angeboten. Das prominenteste Beispiel ist sicherlich Google Scholar, auch wenn diese Suchmaschine nicht die erste Wissenschaftssuchmaschine war.

Viele wissenschaftliche Inhalte sind potentiell auch mit konventionellen Websuchmaschinen auffindbar, allerdings ist es oft mühevoll, die genuin wissenschaftlichen Inhalte aus den anderen Treffern herauszufiltern (ausführlich dazu in dem Beitrag von Pieper und Wolf in diesem Band). Allein daraus ergäbe sich schon die Notwendigkeit für Wissenschaftssuchmaschinen, die sich auf die Indexierung des „wissenschaftlichen Web“ beschränken. Allerdings ist ein Großteil der wissenschaftlichen Inhalte in Datenbanken oder in kostenpflichtigen Angeboten „vergraben“; man spricht hier vom „Academic Invisible Web“ [4].

Wissenschaftssuchmaschinen beschränken sich entweder auf die Datenbank-Inhalte oder den wissenschaftlichen Teil des Oberflächenweb, oder aber sie versuchen, Inhalte aus „beiden Welten“ unter einer Oberfläche suchbar zu machen.

Prototyp für eine solche Hybridsuchmaschine ist Scirus [22, 23]. Diese Suchmaschine erfasst mittels Crawling einen als wissenschaftlich klassifizierten Teil des Web und ergänzt diesen einerseits um Verlagsinhalte (Scirus ist ein Angebot von Elsevier), andererseits um Inhalte aus Open-Access-Repositories. In dieser Kombination ist Scirus einzigartig und kann wohl als die elaborierteste Wissenschaftssuchmaschine bezeichnet werden.

Der Ansatz von Google Scholar ist dagegen beschränkter [24, 25]. Diese Suchmaschine konzentriert sich vornehmlich auf die Erschließung wissenschaftlicher Artikel, dazu kommen Bücher und weitere Quellen, die über die Referenzen aus den Artikeln gewonnen wurden. Artikel werden mittels Crawling im freien Web, aber auch innerhalb der Angebote von Verlagen und Fachgesellschaften gefunden. Dabei erhält Google teilweise Zugriff auf ansonsten kostenpflichtige Bereiche der Verlagsangebote. Die Artikel werden im Volltext erfasst und durchsuchbar gemacht. Mittels automatischer Verfahren werden Autorennamen, Zeitschriftentitel, Erscheinungsjahre und die Literaturangaben extrahiert. Verschiedene Versionen eines Beitrags (beispielsweise Original-PDF beim Verlag, Preprint, Manuskript auf dem Server des Autors) werden zusammengefasst. Durch diese Aufbereitung ist eine Zitationsanalyse möglich, wobei die Zahl der Zitationen auch in das Ranking bei Google Scholar eingeht. Gerade diese Zitationsanalyse macht Google Scholar einzigartig, auch wenn sie wegen der fehlerhaften Zuordnung von zitierten und zitierenden Werken heftig kritisiert wurde [26]. Außerdem bietet Google Scholar hinsichtlich der Abdeckung der indexierten Quellen keinerlei Transparenz; empirische Untersuchungen stellen fest, dass Google Scholar nicht die Abdeckung von Fachdatenbanken erreicht (s. u.a. [27, 28]).

Weitere Ansätze von Wissenschaftssuchmaschinen versuchen, exklusiv die Inhalte der Verlage strukturiert zu erschließen, also direkt auf die Datenbanken der Verlage zuzugreifen und die dort schon vorhandenen Metadaten (Autorenangaben, Schlagwörter, Klassifikation) auszunutzen. Dies war bei der mittlerweile eingestellten Suchmaschine Windows Live Academic [29] der Fall. Die Beschränkung allein auf Web-Inhalte wird von Thomson-Reuters mit Thomson Scientific Web Plus verfolgt, welches als Ergänzung zu den eigenen Angeboten wie Web of Science zu sehen ist. Dieses Angebot ist nur zahlenden Kunden der entsprechenden Angebote zugänglich.

Die vorgestellten Angebote zeigen die unterschiedlichen Ansätze zur Indexierung des wissenschaftlichen Web. Allerdings ist es bisher keinem Anbieter gelungen, eine Suchmaschine aufzubauen, die Wissenschaftlern eine Recherche erlaubt, die andere Quellen (wie Fachdatenbanken oder konventionelle Websuchmaschinen) überflüssig machen würde [4]. Wissenschaftssuchmaschinen sollten bei einer Recherche also stets nur komplementär eingesetzt werden.

3.5. Suchmaschinen für Bücher / Digitalisate

Die Suche in Buchinhalten unterscheidet sich wesentlich von den anderen hier vorgestellten Bereichen, da die zu indexierende Kollektion nicht direkt im Web (oder in anderen Datenbanken) verfügbar ist, sondern erst aus der „Offline-Welt“ herübergeholt werden muss. Die Inhalte liegen nicht bzw. nur zum Teil bereits in digitaler Form vor und müssen für die weitere Verarbeitung erst eingescannt werden.

Initiativen mit dem Ziel, (gedruckte) Bücher digital zugänglich zu machen, gibt es bereits seit langer Zeit. Das *Project Gutenberg* hat zum Ziel, gemeinfreie Werke zugänglich zu machen; die Erfassung erfolgt(e) durch das Abtippen der Bücher. Digitalisierungsinitiativen im wissenschaftlichen Bereich richteten sich vor allem auf wertvolle Handschriften und Drucke bzw. rare Werke, die so der Forschung zur Verfügung gestellt werden sollen. Die Initiative der Suchmaschinenbetreiber – allen voran Google – hat hingegen eine neue Qualität: Es soll nicht mehr ein nur ein kleiner Bestand zugänglich gemacht werden, sondern vielmehr sollen möglichst viele Werke (nahezu wahllos) eingescannt werden [30, 31].

Zu unterscheiden ist zwischen älteren Werken aus Bibliotheken, die nicht mehr durch das Urheberrecht geschützt sind und daher ohne Einschränkungen zugänglich gemacht werden dürften, und neueren Werken, deren Rechteinhaber sich teilweise explizit gegen die Digitalisierung und Verfügbarmachung durch Dritte sperren. Die Kritik an Googles Buchsuche beruht zu einem großen Teil darauf, dass das Unternehmen die Rechteinhaber nicht nach einer Genehmigung zur Aufnahme in die Buchsuche gefragt hatte, sondern die entsprechenden Werke nur nach Aufforderung aus dem Programm nahm. Dabei geht Google davon aus, dass das Zeigen von Ausschnitten aus den Werken den entsprechenden Regelungen für Zitate entspricht. Entsprechend werden gemeinfreie und geschützte Werke in unterschiedlicher Form zugänglich gemacht: Gemeinfreie Werke können im Volltext angesehen und auch als PDF heruntergeladen werden, während von geschützten Büchern nur Ausschnitte angesehen werden können (neben Inhaltsverzeichnis und Deckblättern sind dies in der Regel einige Seiten, die den gesuchten Begriff umgeben).

Der Problematik der urheberrechtlich geschützten Werke entgehen andere Anbieter entweder dadurch, dass sie sich von Vornherein auf die gemeinfreien Werke beschränken oder aber dadurch, dass sie mit den Verlagen explizit zusammenarbeiten, für jedes zu digitalisierende Buch dort eine Genehmigung einholen und Vereinbarung über den Textanteil, der zugänglich gemacht werden darf, treffen. Den erstgenannten Weg beschreitet die Open Content Alliance (OCA), ein Zusammenschluss, der gemeinfreie digitalisierte Bücher verfügbar macht und für diese keinerlei Beschränkungen in Bezug auf die Weiterverwendung erhebt. Dies unterscheidet die OCA von Google. Die dort digitalisierten Werke dürfen nicht beliebig weiterverwendet werden, was auch zu teils heftiger Kritik führte, sofern Bibliotheken in diesem Projekt mit Google zusammenarbeiten. Die beiden großen Konkurrenten Yahoo und Microsoft haben beide die OCA in ihren Bemühungen unterstützt (Microsoft hat inzwischen seine eigene Buchsuche, die auf OCA-Titeln basierte, allerdings eingestellt). Die im Rahmen der Open Content Alliance digitalisierten Werke werden für jedermann über das Internet Archive zugänglich gemacht.

Den Weg über Vereinbarungen mit den Verlagen geht Amazon, wo die Volltexte ausgewählter Bücher über die Funktion „Search Inside“ verfügbar gemacht werden [32-34]. Ähnlich wie bei Google sind nur Ausschnitte verfügbar, die Bücher können aber im Volltext durchsucht werden und von der Fundstelle aus kann jeweils ein paar Seiten vor- und zurückgeblättert werden.

Bei der Erschließung der Bücher gehen alle genannten Anbieter den Weg der Volltexterschließung; Amazon extrahiert zusätzlich Kernsätze, Zitationen und Referenzen. Bei den Buchsuchmaschinen handelt es sich also um Volltextsuchmaschinen, die auf einer zuerst digitalisierten Kollektion von Volltexten aufbauen. Zusätzlich werden Metaangaben von Verlagen oder Bibliotheken eingebunden (in der Regel die bibliographischen Angaben).

Da die Kollektionen von jedem Anbieter individuell aufgebaut werden, lohnt sich die Suche bei mehr als einem Anbieter. Aufgrund der beschränkten Kollektionen kann die Buchsuche auch nicht die Recherche in Bibliotheken und Bibliographien ersetzen.

3.6. Bibliothekssuchmaschinen

Unter den in diesem Abschnitt diskutierten Bibliothekssuchmaschinen ist nicht exklusiv die Suche in Volltexten von Büchern zu sehen (siehe Abschnitt 3.5), sondern vielmehr der Ausbau von Bibliothekskatalogen (*Online Public Access Catalog, OPAC*) hin zu einer Suchmaschine, die einerseits die bibliographischen Angaben des OPAC um weitere Informationen anreichert, andererseits dessen Datenbestand mit weiteren Angeboten der Bibliotheken (wie lizenzierten Datenbanken) zusammenfasst (vgl. [35]).

Die Notwendigkeit, den elektronischen Bibliothekskatalog mittels Suchmaschinentechnologie auf eine neue Basis zu stellen, wird bereits seit einigen Jahren gesehen [36-38]. Um diese Technologie auch ausnutzen zu können, ist jedoch eine Anreicherung der bibliographischen Angaben um Inhaltsverzeichnisse, Klappentexte und im besten Fall sogar die Volltexte der Bücher notwendig. Dies geschieht seit einigen Jahren auch im Rahmen verschiedener Initiativen, wobei zumindest teilweise allerdings nur zusätzliche Texte bereitgestellt werden, die nicht durchsuchbar sind, sondern nur nach dem Aufrufen eines Datensatzes zur zusätzlichen Information genutzt werden können. Um die Stärken von Rankingfunktionen ausnutzen zu können, ist aber eine Mindestmenge an Text notwendig; Probleme bei den Bibliotheksinhalten ergeben sich auch dann, wenn zu unterschiedlichen Datensätzen eine unterschiedliche Informationsmenge zur Verfügung steht. So ist eine Rankingfunktion schwer auf Datensätze anzuwenden, wenn ein Datensatz nur aus bibliographischen Angaben besteht, während ein anderer Datensatz zusätzlich Text aus einem Inhaltsverzeichnis enthält.

Neben diesen Problemen der Anreicherung (die für aktuelle Titel vielleicht noch zu leisten ist, retrospektiv aber schnell an Grenzen stößt) ergibt sich für Bibliotheken das Problem der verteilten Ressourcen. Entgegen den Erwartungen der Nutzer ist der OPAC nicht das zentrale Verzeichnis aller in der Bibliothek verfügbaren Materialien, sondern zeigt nur einen Teil. Weitere Informationen sind in lizenzierten Fachdatenbanken verfügbar, die bei einer OPAC-Recherche allerdings nicht mit durchsucht werden. Dazu kommen Web-Quellen und Hinweise auf Datenbanken, die u.U. auf den Webseiten der Bibliothek oder in gemeinsam gepflegten Verzeichnissen (wie DBIS) abrufbar sind. Die Herausforderung besteht darin, diese Kollektionen gemeinsam verfügbar zu machen und dem Nutzer die mühsame Recherche nach den *relevanten Quellen* abzunehmen (vgl. [39]).

Die Recherchefunktion ist als zentral für den Erfolg von Bibliotheksangeboten anzusehen. Insofern ist auch die an sich wünschenswerte Diskussion um den „Katalog 2.0“ kritisch zu betrachten, da dieser den bestehenden Katalog im Wesentlichen um Funktionen erweitert, die sicher wünschenswert sind, jedoch die Trefferliste selbst nicht erheblich verbessern.

3.7. Bildersuchmaschinen

Die Bildersuche ist inzwischen ein Standard für alle Websuchmaschinen; ihre Ergebnisse werden häufig auch in die regulären Webergebnisse eingebunden. Die Schwierigkeit bei der Bildersuche liegt darin, dass, während sich alle bisher in diesem

Kapitel vorgestellten Spezialsuchen auf Texte beziehen, zu den Bildern keine oder nur eingeschränkte textuelle Informationen vorhanden sind.

Bei der Erschließung von Bildern konkurrieren grundsätzlich drei Ansätze: Die intellektuelle Erschließung mit Hilfe von Metadaten, die automatische inhaltsbasierte Erschließung [40] sowie die Erschließung über Umgebungstexte (vgl. [41]).

Bildersuchmaschinen beziehen ihre Daten aus dem freien Web (nur Yahoo ergänzt die Bilder aus dem Webcrawl durch Bilder aus dem zum Unternehmen gehörenden Angebot Flickr). Aufgrund der Menge der zu erfassenden Bilder ist eine intellektuelle Erschließung nicht praktikabel, eine inhaltsbasierte Indexierung erreicht in einer thematisch nicht beschränkten Kollektion keine zufrieden stellenden Ergebnisse. Insofern bleibt die Erschließung mittels Umgebungstexten, die von den Websuchmaschinen entwickelt wurde und eine technisch relativ anspruchslose Lösung darstellt, die jedoch durch ihren praktischen Nutzen überzeugt. Die nach diesem Verfahren arbeitenden Bildersuchmaschinen sind zumindest in der Lage, befriedigende Ergebnisse zu liefern (vgl. [41]).

Für die Recherche nach Bildern gilt weitgehend, was auch für die konventionelle Web-Recherche gesagt werden kann: Die Verwendung verschiedener Suchmaschinen lohnt, da durch den Aufbau der Kollektionen durch Crawling unterschiedliche Bilder gefunden werden, die durch die sich unterscheidenden Rankingverfahren auch in verschiedener Reihung angezeigt werden.

Hinsichtlich der Suchoptionen unterscheiden sich die Bildersuchmaschinen ebenso. Keine der bekannten Suchmaschinen bietet das volle Spektrum an Suchfunktionen, so dass bei spezialisierten Anfragen die jeweils am besten geeignete Suchmaschine ausgewählt werden sollte [42, 43]. Dabei ist zu beachten, dass auch Suchmaschinen, bei denen sich eine Recherche im sonstigen Bestand aufgrund der anderswo zugekauften Ergebnisse nicht mehr lohnt, bei der Bildersuche durchaus noch sinnvoll verwendbar sein können. So bietet AltaVista zwar schon lange keinen eigenen Index mehr an, die genauen Formatbeschränkungen in der erweiterten Bildersuche machen diese Suchmaschine aber zumindest für eine beschränkte Zahl von Anfragen zur ersten Wahl.

Übliche Einschränkungen für die Bildersuche sind neben Größenbeschränkungen die Beschränkung auf Farb- oder Schwarzweißbilder, auf Abbildungen von Personen, auf den Dateityp und auf den Bildtyp (Foto, Grafik, Banner).

3.8. Audio- und Videosuchmaschinen

Die Ansätze für die Indexierung von Audio- und Videoinhalten entsprechen den in Abschnitt 3.7 beschriebenen. Auch hier verwenden die gängigen Suchmaschinen den technisch wenig anspruchsvollen, aber praktikablen Ansatz der Erschließung durch Umgebungstexte.

Eine Audiosuche im Web wird von den großen Suchmaschinen nicht mehr angeboten. Yahoo hat sein Angebot zugunsten von Yahoo Musik (Suche in einer kommerziellen Musikdatenbank) eingestellt, die Audiosuche ist aber weiterhin über die AltaVista-Suche erreichbar. Dort kann nach Dateiformat und Länge beschränkt werden, die Treffer kommen von Webseiten aus dem freien Web.

Eine Videosuche hingegen wird sowohl von Google, Yahoo als auch Ask (in den englischsprachigen Versionen) angeboten (s. auch [44]). Die Inhalte stammen jeweils aus dem freien Web, durch die allgemeine Weiterverwendbarkeit der Youtube-Inhalte hat Google keinen Vorsprung durch exklusive Inhalte. Die Videosuche bei Google

entspricht auch nicht der Suche bei Youtube, sondern bietet darüber hinaus anderswo im Web aufgefundene Inhalte.

3.9. Produktsuchmaschinen

Bei den Produkt- oder Shoppingsuchmaschinen handelt es um Angebote, die ihre Daten aus Produktkatalogen beziehen. Der Unterschied zwischen Einzelanbietern und Suchmaschinen liegt darin, dass die Suchmaschinen die Angebote unterschiedlicher Händler durchsuch- und vergleichbar machen.

Die Daten werden von den einzelnen Anbietern in der Regel per *bulk upload* eingespielt, d.h. ein kompletter Produktkatalog kann auf einmal eingespielt werden. Die Suchmaschine ergänzt bzw. aktualisiert diesen Bestand dann um die hochgeladenen Produkte. Dadurch, dass die Daten strukturiert angeliefert werden, ergeben sich bei der Recherche gute Einschränkungsmöglichkeiten. Üblich sind facettierte Drill Downs, die den Nutzer in Anschluss an seine Suchanfrage (und ein erstes Ergebnisset) mittels Browsing punktgenau zu den gewünschten Ergebnissen leiten können.

4. Fazit

Mit den Spezialsuchmaschinen haben die Anbieter der allgemeinen Websuchmaschinen ihre Angebote wesentlich ausgebaut und verfeinert. Es zeigen sich je nach Themengebiet unterschiedliche Ansätze, vor allem hinsichtlich der verwendeten Datenquellen. Während es bei einigen Spezialsuchen ausreicht, Inhalte aus dem freien Web zu crawlen und der Schwerpunkt der Spezialsuche auf der Beschränkung der Kollektion und der Suchoberfläche liegt (Beispiel: Nachrichtensuche), erfordern andere Angebote eine gezielte Kombination aus Inhalten des freien Webs mit solchen aus zugekauften Datenbanken (Beispiel: Lokale Suche).

Für den professionell Recherchierenden bieten Spezialsuchmaschinen viele Möglichkeiten, fokussiert zu recherchieren und damit schnell zu punktgenauen Ergebnissen zu gelangen. Für den ungeschulten Suchmaschinennutzer gehen die Betreiber der Suchmaschinen zunehmend dazu über, Ergebnisse aus den von ihnen angebotenen Spezialsuchen in die regulären Web-Ergebnisse einzubauen. Dies ist als eine Reaktion auf die Unkenntnis der Nutzer zu sehen und auf das offensichtliche Übersehen von Auswahlmöglichkeiten auf den Such- und Trefferseiten („tab blindness“).

Während die Darstellung der Spezialsuchmaschinen im zweiten Teil dieses Kapitels schon einige Mängel aufgezeigt hat, ist vor allem zu bedauern, dass bisher nur zu wenigen speziellen Suchbereichen systematische Evaluierungen vorliegen. So beschränken sich Suchmaschinenvergleiche leider meist auf die konventionellen Web-Ergebnisse, während die anderen Angebote außen vor gelassen werden.

Literaturangaben

1. Lewandowski, D.: Web Information Retrieval: Technologien zur Informationssuche im Internet. DGI, Frankfurt am Main (2005)

2. Lewandowski, D.: A three-year study on the freshness of Web search engine databases. Journal of Information Science **34** (2008)
3. Sherman, C., Price, G.: The Invisible Web: Uncovering Information Sources Search Engines Can't See. Information Today, Medford, NJ (2001)
4. Lewandowski, D., Mayr, P.: Exploring the academic invisible web. Library Hi Tech **24** (2006) 529-539
5. Bergman, M.K.: The Deep Web: Surfacing Hidden Value. Journal of Electronic Publishing **7** (2001)
6. Mayr, P., Walter, A.-K.: An exploratory study of Google Scholar. Online Information Review **31** (2007)
7. Patterson, A.: Why Writing Your Own Search Engine is Hard. ACM Queue **2** (2004) 49-53
8. Lewandowski, D.: Abfragesprachen und erweiterte Suchfunktionen von WWW-Suchmaschinen. Information Wissenschaft und Praxis **55** (2004) 97-102
9. Kwiatkowski, M., Höhfeld, S.: Thematisches Aufspüren von Web-Dokumenten - Eine kritische Betrachtung von Focused Crawling-Strategien. Information Wissenschaft und Praxis **58** (2007) 69-82
10. Stock, W.G.: Weltregionen des Internet: Digitale Informationen im WWW und via WWW. Password **18** (2003) 26-28
11. Wiggins, R.W.: The Effects of September 11 on the Leading Search Engine. First Monday **7** (2001)
12. Lewandowski, D.: Google News Archive : Googles Einstieg bei den bezahlten Nachrichten. Password (2006) 20-21
13. Machill, M., Lewandowski, D., Karzauninkat, S.: Journalistische Aktualität im Internet. Ein Experiment mit den "News-Suchfunktionen" von Suchmaschinen. In: Machill, M., Schneider, N. (eds.): Suchmaschinen: Herausforderungen für die Medienpolitik. Vistas, Berlin (2005) 105-164
14. Range, S., Schweins, R.: Klicks, Quoten, Reizwörter : Nachrichten-Sites im Internet – Wie das Web den Journalismus verändert. Friedrich-Ebert–Stiftung, Berlin (2007)
15. Neuberger, C., Nuernbergk, C., Rischke, M.: Weblogs und Journalismus: Konkurrenz, Ergänzung oder Integration? Media Perspektiven (2007) 96-112
16. Thelwall, M., Hasler, L.: Blog search engines. Online Information Review **31** (2007) 467-479
17. Pikas, C.K.: Blog searching for competitive intelligence, brand image, and reputation management. Online **29** (2005) 16-21
18. Starr, J.: Local information from a blog near you. Searcher **15** (2007) 24-29
19. Notess, G.R.: Locating Uses for Local Search. Online **29** (2005) 39-41
20. Lewandowski, D.: Lokale Suche allerorten - Mit Ausnahme von Web.de eine Enttäuschung. Password (2006) 34-35
21. Jones, R., Zhang, W.V., Rey, B., Jhala, P., Stipp, E.: Geographic Intention and Modification in Web Search. International Journal of Geographical Information Science **22** (2008) 1-20
22. Scirus White Paper: How Scirus works. (2004) http://www.scirus.com/press/pdf/WhitePaper_Scirus.pdf
23. Jacsó, P.: Scirus. (2006) http://www.jacso.info/PDFs/jacso-NY-Times-26-6.pdf
24. Lewandowski, D.: Google Scholar - Aufbau und strategische Ausrichtung des Angebots sowie Auswirkung auf andere Angebote im Bereich der wissenschaftlichen Suchmaschinen. (2005) http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/Expertise_Google-Scholar.pdf
25. Jacsó, P.: Google Scholar: The pros and cons. Online Information Review **29** (2005) 208-214
26. Jacsó, P.: Google Scholar revisited. Online Information Review **32** (2008) 102-114
27. Mayr, P., Walter, A.-K.: Abdeckung und Aktualität des Suchdienstes Google Scholar. Information Wissenschaft und Praxis **57** (2006) 133-140
28. Lewandowski, D.: Nachweis deutschsprachiger bibliotheks- und informationswissenschaftlicher Aufsätze in Google Scholar. Information Wissenschaft und Praxis **58** (2007) 165-168
29. Jacsó, P.: Windows Live Academic. http://www.jacso.info/PDFs/jacso-jst-scielo-microsoft-acad-live.pdf
30. Notess, G.R.: Searching books between the covers. Online **29** (2005) 41-44
31. Ojala, M.: Searching by and for the book. Online **31** (2007) 49-51
32. Bank, M.A.: Amazon.com opens the books. Online **28** (2004) 30-33
33. Jacsó, P.: Amazon. http://www.jacso.info/gale/amazon/amazon.htm
34. McDermott, I.E.: Search Inside the Book : a new reference tool. Searcher **12** (2004) 41-44
35. Lewandowski, D.: Suchmaschinen als Konkurrenten der Bibliothekskataloge: Wie Bibliotheken ihre Angebote durch Suchmaschinentechnologie attraktiver und durch Öffnung für die allgemeinen Suchmaschinen populärer machen können. Zeitschrift für Bibliothekswesen und Bibliographie **53** (2006) 71-78

36. Summann, F., Lossau, N.: Suchmaschinentechnologie und Digitale Bibliotheken: Von der Theorie zur Praxis. Zeitschrift für Bibliothekswesen und Bibliographie **52** (2005) 13-17
37. Summann, F., Wolf, S.: Suchmaschinentechnologie für digitale Bibliotheken. Information Wissenschaft und Praxis **56** (2005) 51-57
38. Hauer, M.: Neue Qualitäten in Bibliotheken: Durch Content-Ergänzung, maschinelle Indexierung und modernes Information Retrieval können Recherchen in Bibliotheken deutlich verbessert werden. ABI-Technik **24** (2004) 262-268
39. Lewandowski, D.: Search engine user behaviour: How can users be guided to quality content? Information Services & Use **28** (2008)
40. Schmitt, I., Nürnberger, A.: Inhaltsbasiertes Multimedia Retrieval: Überblick und Herausforderungen. Datenbank-Spektrum (2006) 6-13
41. Lewandowski, D., Höchstötter, N.: Wie effektiv sind Suchmaschinen zur Recherche nach Bildern von berühmten Persönlichkeiten? In: Ockenfeld, M. (ed.): Verfügbarkeit von Informationen. 30. DGI-Online-Tagung. DGI, Frankfurt am Main (2008)
42. Hassan, I., Zhang, J.: Image search engine feature analysis. Online Information Review **25** (2001) 103-114
43. Tomauolo, N.G.: When image is everything : finding and using graphics from the Web. Searcher **10** (2002) 10
44. McDermott, I.E.: Movement on my monitor : video on the Web. Searcher **14** (2006) 16-20