Dirk Lewandowski

Suchmaschinen als Konkurrenten der Bibliothekskataloge: Wie Bibliotheken ihre Angebote durch Suchmaschinentechnologie attraktiver und durch Öffnung für die allgemeinen Suchmaschinen populärer machen können

Die elektronischen Bibliotheksangebote konkurrieren zunehmend mit den allgemeinen Websuchmaschinen und spezialisierten Wissenschaftssuchmaschinen um die Gunst der Nutzer. Dieser Aufsatz gibt einen kritischen Überblick über die bisherigen Initiativen zum Einsatz von Suchmaschinentechnologie im Bibliothekskontext sowie zur Sichtbarmachung von Bibliotheksinhalten in allgemeinen Suchmaschinen. Darauf aufbauend wird dargestellt, wie Bibliotheksangebote tatsächlich für den Nutzer attraktiver werden können und ihre Sichtbarkeit in Suchmaschinen erhöht werden kann. Grundlegend muss der OPAC zum zentralen Nachweisinstrument für alle in der jeweiligen Bibliothek verfügbaren Informationen gemacht werden. Außerdem müssen die Katalogdaten deutlich angereichert werden, um die Stärken der Suchmaschinentechnologie auch tatsächlich nutzen zu können. Auf der Basis dieser erweiterten Textmenge kann schließlich auch leicht die Auffindbarkeit in allgemeinen Suchmaschinen verbessert werden.

Electronic and online library services are competing increasingly with the general web search engines and with specialized academic search engines over the same clientele. This article provides a critical survey of the latest initiatives to apply search engine technology to the library field and make libraries and their services more visible through web search engines. Based upon these observations, the article then describes how libraries can in fact be made more attractive and their visibility raised via search engines. Fundamentally, the OPAC needs to become the central source of reference regarding all information available in a library. Furthermore, catalog data must be distinctly enriched in order to make good use of the strengths of search engine technology. On the basis of this broadened textual data, it will then be easy to improve the »findability« or rate of search returns via search engines.

Einleitung

Bibliothekare blicken oft neidisch auf die allgemeinen Suchmaschinen. Diese stellen mit die populärsten Angebote im Web dar und stehen teils in direkter Konkurrenz zu den Bibliothekskatalogen und weiteren elektronischen Angeboten der Bibliotheken. In diesem Aufsatz sollen einerseits die Erfolgsfaktoren der allgemeinen wie der spezialisierten Suchmaschinen aufgezeigt werden, andererseits sollen die bisherigen Strategien der Bibliotheken beschrieben werden, wie durch Ausnutzung der populären Suchmaschinen die Nutzer zurückgewonnen oder neue hinzugewonnen werden sollen.

Im Wesentlichen handelt es sich dabei um zwei Strategien, die von den Bibliotheken verfolgt werden: Einerseits versuchen sie, so zu sein wie die Suchmaschinen, indem sie aus der Suchmaschinenwelt kommende Technologien auf ihre Katalogdaten anwenden. Andererseits versuchen sie, vom Erfolg der Suchmaschinen zu profitieren, indem sie versuchen, ihre Bestände in die Indizes der großen allgemeinen Suchmaschinen zu bringen. Beide Ansätze sind mit Problemen verbunden, auf die ausführlich eingegangen werden wird.

Schließlich werden Empfehlungen gegeben, wie Bibliotheken ihre Angebote für die eigenen Nutzer populärer gestalten können und wie andererseits Nutzer, die bevorzugt Suchmaschinen verwenden, für die Bibliotheksangebote gewonnen werden können.

Foto privat

Dirk Lewandowski

Nutzung von elektronischen Bibliotheksangeboten

Die elektronischen Bibliotheksangebote werden von den Nutzern nicht in dem Maße angenommen, wie dies von Bibliotheksseite gewünscht wird. Nutzer verwenden stattdessen vor allem allgemeine Suchmaschinen wie Google, Yahoo oder MSN, in zunehmendem Maße aber auch spezialisierte Suchdienste wie Google Scholar oder Scirus. Eine weitere Konkurrenz besteht mit Online-Buchhändlern (vor allem Amazon), die umfangreiche Zusatzinformationen zu den von ihnen angebotenen Büchern anbieten und hier als Referenz für die Möglichkeiten des *catalogue enrichment* angeführt werden sollen.

Nutzer verwenden allgemeine Suchmaschinen

Die Recherche nach Fachinformationen findet vor allem auf Studierendenseite zu einem großen Teil über allgemeine Suchmaschinen statt.[1] Die SteFi-Studie aus dem Jahr 2001 fasst das Verhalten der Studierenden folgendermaßen zusammen:

»Von einem systematischen, professionellen Gebrauch dieses Mediums [Internet] kann aber kaum die Rede sein. Statt das gesamte Spektrum fachspezifischer elektronischer wissenschaftlicher Medien zu nutzen, beschränken sich die Studierenden häufig auf das ›Browsen‹ im Internet mit Hilfe freier Suchmaschinen wie Lycos oder Web.de. Ob sie dabei auf wertvolle und hilfreiche Informationen stoßen, bleibt ihnen verschlossen, weil sie ihre Kenntnisse im Umgang mit den neuen Medien nicht systematisch im Rahmen ihres Studiums, sondern im Selbstlernverfahren erworben haben.«[2]

Umgang mit neuen Medien im Selbstlernverfahren

Während die überwältigende Mehrheit der Studierenden dieser repräsentativen Befragung zufolge

Suchmaschinen für die Recherche nach wissenschaftlichen Informationen einsetzt, nutzt nur knapp die Hälfte der Befragten überhaupt den OPAC – und von ihnen nur 16,2 Prozent häufig.[3]

Es ist nicht davon auszugehen, dass sich die Situation in den letzten Jahren gebessert hat. Es herrscht ein dringender Handlungsbedarf seitens der Bibliotheken, wenn sie von den Studierenden und Lehrenden nicht nur als Bücherspeicher, sondern mit ihren OPACs auch als Instrument zur systematischen Recherche anerkannt werden wollen. Wesentliche Gründe für die mangelnde Attraktivität der OPACs sind sicherlich in den wenig umfangreichen Titelaufnahmen und der unvollständigen Repräsentation des Gesamtbestands der jeweiligen Bibliothek[4] zu sehen. Auf dieses Thema wird noch genauer eingegangen werden.

mangelnde Attraktivität der OPACs

Ein weiteres Problem besteht durch die mangelnde Auffindbarkeit der Bibliotheksangebote von außen: Zwar können die Webseiten der Bibliotheken in der Regel gut durch die Suchmaschinen indexiert werden, die tatsächlichen *Inhalte* der Bibliotheken (bzw. deren Repräsentanten in Form von Katalogeinträgen) können die Suchmaschinen jedoch nicht erfassen. Diese liegen im sog. *Invisible Web,* also in dem Bereich des Web, der zwar für Menschen ohne größere Probleme erreichbar ist, nicht jedoch für Suchmaschinen.

Von Seiten der Bibliotheken werden zwei Ansätze verfolgt, um ihre Angebote bei den Nutzern populärer zu machen: Einerseits können die Nutzer durch Werbemaßnahmen oder Aufklärung über Inhalt und Nutzen der Bibliotheksangebote dazu gebracht werden, diese für ihre Recherchen einzusetzen. Dabei kann auf den ganzen Bereich des *Invisible Web* verwiesen werden, in dem sich zahlreiche nützliche Angebote finden, die über generelle Suchmaschinen nicht erschlossen werden – speziell natürlich auf denjenigen (kostenpflichtigen) Bereich, auf den exklusiv über Bibliothekslizenzen zugegriffen werden kann.[5]

Auf der anderen Seite wird versucht, die Nutzer dort »abzuholen«, wo sie (zumindest bisher) ihre Recherchen bevorzugt durchführen, nämlich bei den allgemeinen Suchmaschinen. Dazu müssen die für die Suchmaschinen nicht sichtbaren Inhalte sichtbar gemacht werden, indem aus den dynamischen Inhalten der Kataloge statische HTML-Seiten generiert werden, die von den Suchmaschinen erschlossen werden können.

Stellung des OPAC innerhalb der Angebote von Bibliotheken

Abbildung 1 zeigt den Weg zu den gewünschten Informationen in einer Bibliothek am Beispiel der Universitäts- und Landesbibliothek Düsseldorf. Unabhängig davon, dass die umfangreichen Angebote der Bibliothek tatsächlich nur auf verschiedenen Wegen erreichbar sind und keine Oberfläche existiert, die einen gleichzeitigen Zugriff auf alle Quellen erlaubt, dürfte das Schaubild bei einem ungeübten Nutzer erst einmal Ratlosigkeit hinterlassen.

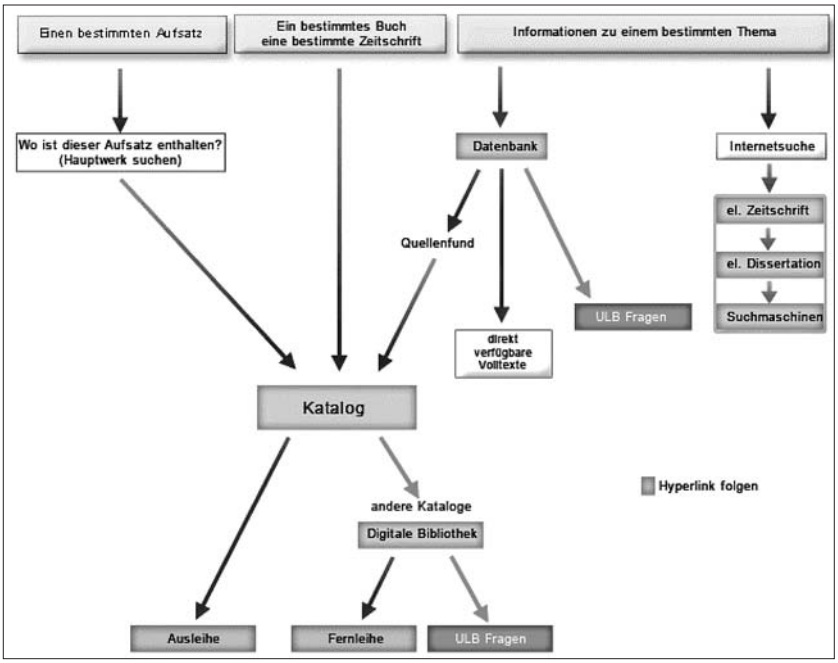

Abb.1: Darstellung des Wegs zur gewünschten Information (Universitäts- und Landesbibliothek Düsseldorf)[6]

Zu erkennen ist allerdings, welches das zentrale Angebot der Bibliothek ist, das in der Mitte des Schaubilds platziert ist und auf das von drei Pfeilen verwiesen wird: der Online-Katalog. Er stellt *das* Nachweisinstrument der Bibliotheksangebote dar bzw. wird wenigstens von den Nutzern als solches angesehen.[7] Das Schaubild zeigt aber auch, dass über den OPAC mitnichten alle in der Bibliothek vorhandenen bzw. über die Bibliothek erreichbaren Angebote recherchiert werden können. Solange dies so ist, ist es nicht verwunderlich, wenn Nutzer annehmen, dass sie alles im OPAC finden könnten, sich nach einer Recherche jedoch enttäuscht von diesem abwenden, weil sie nichts bzw. zu wenig gefunden haben.[8] Im Gegensatz zu diesen Recherchen suggerieren die Suchmaschinen zu nahezu jeder Anfrage eine umfangreiche Treffermenge, auch wenn die Präzision gerade bei speziellen Anfragen oft mangelhaft ist.

Als gravierender Nachteil der OPACs ist also anzusehen, dass sie nur einen Teil des tatsächlich vorhandenen Informationsangebots nachweisen. Neben der »Datenbank der Bücher« sind für den Nutzer aber vor allem Nachweise und Volltexte von Zeitschriftenaufsätzen relevant, die er im OPAC nicht (oder nur zu

einem gewissen Teil) finden kann. Auch Angebote wie die *Digitale Bibliothek* kommen diesem Mangel nur zum Teil entgegen.

Als Konsequenz für die Gestaltung von OPACs ist daher die Integration sowohl von weiteren Dokumenttypen wie Aufsätzen und Webquellen als auch die inhaltliche Anreicherung der einzelnen Datensätze zu sehen. Wo bisher schon Aufsätze in die OPACs aufgenommen werden, geschieht dies in der Regel in Form reiner bibliographischer Angaben. Um die Stärken von Suchmaschinentechnologie ausreizen zu können (und natürlich auch zum Nutzen des Recherchierenden) ist aber auch bei diesen Daten eine Anreicherung um Abstracts und/oder um aus dem Volltext automatisch generierte Begriffe notwendig. Erst durch die Anreicherung der Datensätze kann die Attraktivität der OPACs für die Recherche und gezielte Quellenauswahl gesteigert werden.

Suchmaschinentechnologie

Die Technologie der gängigen Suchmaschinen beruht im Wesentlichen auf zwei Eigenarten: Erstens werden die Dokumente in einem sog. Crawl eingesammelt, zweitens findet bei der Darstellung der Treffer ein Relevance Ranking statt. Beim Crawling handelt es sich um ein Verfolgen der Linkstruktur des Web, um auf bisher unbekannte Dokumente zu stoßen, Veränderungen an bestehenden Dokumenten festzustellen und gelöschte oder verschobene Dokumente aus dem Index zu nehmen.

Im Gegensatz zu klassischen Datenbanken ist dieses Verfahren nötig, da Dokumentbestände im Web weder direkt abgefragt werden können noch strukturiert vorliegen.[9] Zu unterscheiden ist also zwischen Datenbank-Abfragen (OPAC) und Text-Retrieval (Suchmaschinen).

Das Relevance Ranking ist keine Spezialität der Suchmaschinen, allerdings wurde es erst im Webkontext mit der überwältigenden Anzahl der potenziell relevanten Treffer zu nahezu jeder Suchanfrage unverzichtbar. Die Stärken von Rankingverfahren liegen vor allem bei der Anordnung nicht oder nur schwach strukturierter Texte, während Rankingverfahren für ungewichtete Deskriptoren bzw. Schlagwörter nur einen geringeren Wert haben.[10]

Im Gegensatz zur Technologie der Suchmaschinen beruhen Bibliothekskataloge auf dem klassischen Datenbankmodell. Da zwar die Einstiegsseiten der Datenbanken von den Suchmaschinen erfasst werden können, die Suchmaschinen jedoch nicht in der Lage sind, Anfragen an Datenbanken zu stellen und damit zu den eigentlichen Inhalten zu gelangen, zählen Bibliothekskataloge zum sog. *Invisible Web*.[11] Insbesondere in Hinblick auf die (potenziellen) Bibliotheksnutzer, die Suchmaschinen bevorzugt oder gar an Stelle der Bibliothekskataloge für die Literaturrecherche benutzen, ergibt sich ein massives Problem: Die Inhalte der Bibliotheken werden durch die allgemeinen Suchmaschinen nicht erfasst und können damit auch nicht gefunden werden.

OPACs sind für Suchmaschinen »invisible«

So wie die OPACs für die Suchmaschinen »invisible« sind, so sind es die von den Bibliotheken erworbenen Fachdatenbanken und elektronischen Zeitschriften meist für die OPACs. Ähnlich wie bei den Suchmaschinen werden auch hier nur die Einstiegsseiten der Datenbanken nachgewiesen (bzw. deren generelles Vorhandensein), nicht aber die dahinter stehenden Inhalte. Hier findet für den Nutzer ein Bruch statt: Er muss erkennen, dass er auf der einen Seite »echte Treffer« (Bücher und teils Zeitschriftenaufsätze) angezeigt bekommt, auf der anderen Seite aber weitere Sucheinstiege (Datenbanken und Zeitschriften).[12] In den Trefferlisten wird zwischen den beiden Treffertypen nicht unterschieden, so dass die Hinweise auf Sucheinstiege in der Regel eher in der Gesamttreffermenge untergehen.

Spezielle Suchmaschinen

Einige Suchmaschinenbetreiber haben sich Gedanken gemacht, wie man den Komfort der Websuchmaschinen für die Recherche nach wissenschaftlichen Informationen umsetzen kann. Dabei wird auf der einen Seite das freie Web berücksichtigt, in dem sich eine große Zahl wissenschaftlicher Artikel findet, dazu die Homepages von Wissenschaftlern und Forschergruppen. Die Inhalte des freien Webs bieten den Vorteil der Aktualität, zusätzlich finden sich Informationen über Wissenschaftler und Forschungsvorhaben sowie Reports u.ä., die nie den Weg in eine reguläre Veröffentlichung finden. Der Nachteil ist vor allem in der fehlenden Qualitätskontrolle zu sehen.

Relevance Ranking

Aber die Websuchmaschinen gehen mit ihren wissenschaftlichen Spezialsuchen einen Schritt weiter und erfassen auch kommerziell von den Verlagen angebotene Inhalte. Mit der Kombination beider Inhaltstypen bilden sie Hybridsysteme, die wissenschaftliche Aufsätze in größerem Umfang nachweisen können als Aufsatzdatenbanken oder Suchmaschinen für das freie Web. Die technische Lösung ist entweder eine Mischung aus Crawling und Datenbankabfrage oder eine Ausdehnung des Crawlings auf Verlagsinhalte durch Kooperationen mit den Verlagen.

Im Folgenden sollen die beiden bedeutendsten Suchmaschinen für wissenschaftliche Inhalte allgemein vorgestellt und bewertet werden. Diesen ist gemeinsam, dass es sich um Angebote aus dem kom-

merziellen Bereich handelt; eine vergleichbare Suchmaschine, die auf Initiative von oder durch eine Bibliothek oder sonstige Einrichtung der Öffentlichen Hand betrieben wird, existiert nicht. Auf fachspezifische Suchmaschinen soll nicht weiter eingegangen werden.

Google Scholar

Google Scholar ist eine Suchmaschine für wissenschaftliche Inhalte (insbes. Aufsätze im Volltext und bibliographische Nachweise von Büchern), mit der ein Teilbestand des Google-Web-Index recherchiert werden kann. Dazu kommen Angebote von Verlagen und Fachgesellschaften, die in das Angebot mit aufgenommen wurden. Mit vielen der großen Wissenschaftsverlage und Fachgesellschaften wurden Vereinbarungen über die Indexierung deren Inhalte getroffen, es sind aber keineswegs alle großen Verlage mit im Boot. Die wichtigste Ausnahme dürfte Elsevier sein, auf dessen Inhalte nicht zugegriffen werden kann.

Elsevier nicht in Google Scholar

Google Scholar deckt prinzipiell alle Fächer ab, bisher ist allerdings ein Schwerpunkt bei den Naturwissenschaften und der Technik festzustellen. Dies mag aber auch auf die bisher erfassten Quellen zurückzuführen sein – der Schwerpunkt entspricht durchaus dem Angebot der großen Verlage. Ebenso verhält es sich bei der Sprache der erfassten Dokumente: Auch hier werden prinzipiell alle Sprachen erfasst, der Schwerpunkt liegt jedoch klar bei englischsprachigen Dokumenten.

Neben Aufsätzen im Volltext werden Bücher erfasst, diese allerdings nur in Form von Literaturzitaten, die den im Volltext erschlossenen Aufsätzen entnommen wurden. Für jeden Buchtreffer gibt es die Möglichkeit, die Suche im Online-Buchhandel, im Web oder im OCLC-Katalog fortzusetzen.

Kritisch ist die Erfassung unterschiedlicher Texttypen zu sehen: Es werden sowohl Zeitschriftenaufsätze, die das Peer-Review-Verfahren durchlaufen haben, als auch Konferenzberichte, Preprints, Postprints, Reports, usw. erfasst. Dies reicht bis hin zu Seminararbeiten, deren wissenschaftliche Qualität ja durchaus angezweifelt werden kann. Alle diese Texttypen werden durchmischt und nicht ausgewiesen, d. h. es ist oft für den Nutzer nicht feststellbar, welchen Qualitätsstandards die gefundene Information entspricht.

Qualitätsstandard der gefundenen Informationen nicht eindeutig

Der Gesamtumfang von Google Scholar liegt zwischen zwei und sieben Millionen Dokumenten,[13] eine weitere Schätzung beläuft sich auf acht Millionen Dokumente.[14] Festzuhalten ist, dass die Dokumentenzahl im Vergleich zu anderen wissenschaftlichen Angeboten relativ gering ausfällt.

Die Inhalte von Google Scholar werden mit Ausnahme der Bücher im Volltext erfasst. Dabei werden Autorennamen und der Titel der Zeitschrift, in der der Aufsatz erschienen ist, extrahiert. Allerdings zeigt sich hier die Unzuverlässigkeit des Systems: Selbst bei Angeboten, deren Metadaten immer in der gleichen Form präsentiert und die von Google Scholar in großem Maß erschlossen werden (z. B. ACM Portal), finden sich haufenweise falsch extrahierte Autorennamen und Zeitschriftentitel. Weiterhin erfolgt keine Übernahme der auf den Verlagsangeboten vorgegebenen Schlagwörter, Klassifikationsstellen o. ä. Es werden auch keine linguistischen Verfahren angewendet, um die Erschließung zu verbessern.

Der Erfolg von Google Scholar bei den Nutzern ist in seiner einfachen Bedienbarkeit, der Geschwindigkeit und vor allem in der Integration der verschiedenen Inhaltstypen wie Büchern, Aufsätzen von Verlagen und Inhalten aus dem freien Web zu sehen, die alle über ein einziges Suchformular gefunden werden können und in einer einzigen Trefferliste zusammengefasst werden.

Betrachtet man die Vor- und Nachteile von Google Scholar, so ist festzustellen, dass die Nachteile im Wesentlichen auf der inhaltlichen Ebene liegen: Es sind nur relativ wenige Dokumente indexiert, die Erschließung ist mangelhaft und es ist unklar, in welchem Umfang die Quellen erschlossen sind. So stellt sich bei einer Recherche etwa die Frage, ob damit tatsächlich alle Jahrgänge einer Zeitschrift, aus der Artikel gefunden wurden, abgedeckt sind oder, ob auch nur ein einziger Jahrgang dieser Zeitschrift komplett abgedeckt ist.[15]

Die Vorteile des Systems liegen dagegen auf der Ebene des Interfaces und der Trefferanzeige: Das System ist intuitiv bedienbar und zeigt Treffer aus unterschiedlichen Bereichen in einer einzigen, sortierten Trefferliste an. Ähnlich wie auch in der regulären Websuche suggeriert das System dem Nutzer für einen Großteil der Suchanfragen eine große Treffermenge, auch wenn über die Qualität der Treffer damit wenig ausgesagt ist.

Scirus

Scirus[16] ist eine wissenschaftliche Suchmaschine, die von Elsevier betrieben wird. Die Technologie stammt von FAST. Der Index besteht sowohl aus freien Web-Inhalten als auch aus Verlagsinhalten (Elseviers Science Direct), Artikeln aus Open-Access-Archiven und Patenten. Scirus ist die weltweit umfangreichste Suchmaschine für wissenschaftliche Informationen.

Anders als bei Google Scholar ist der Index nicht auf Literatur in Form von Volltexten oder Zitationen beschränkt, sondern die Aufgabe ist es vielmehr, das

gesamte wissenschaftliche Web zu erfassen, das heißt, auch Homepages von Wissenschaftlern und ähnliches. Die Suche lässt sich – auch im Gegensatz zu Google Scholar – auf einzelne Quellen oder eine individuelle Quellenauswahl beschränken.

Deutlicher noch als bei Google Scholar zeigt sich das Problem der unterschiedlichen Zuverlässigkeit der Dokumente: Bei Scirus wird schlicht alles indexiert, was auf irgendeinem Uni-Server abgelegt ist.

Problem der unterschiedlichen Zuverlässigkeit der Dokumente auch bei Scirus

Forschungsportal.net

Das Forschungsportal[17] indexiert die Websites der in Deutschland öffentlich geförderten Forschungseinrichtungen sowie die bei Der Deutschen Bibliothek erfassten Online-Dissertationen. Aufgabe ist es, die deutschen Forschungsserver tiefer zu indexieren als dies allgemeine (Google etc.) oder spezielle (Scirus) Suchmaschinen tun. Ob dies gelingt, wurde bisher noch in keinem wissenschaftlichen Test nachgewiesen.

Die vorgestellten wissenschaftlichen Suchmaschinen kranken an den von den allgemeinen Suchmaschinen bekannten Problemen: Die Abdeckung und die Tiefe der Indexierung sind teils unklar, der Datenbestand ist (bewusst) beschränkt (hier das freie wissenschaftliche Web, dort spezielle Kollektionen). Keine der Suchmaschinen bietet einen vollständigen integrativen Ansatz der kompletten Erfassung aller für den Wissenschaftler relevanten Dokumente.

Suchmaschinentechnologie für Bibliotheksangebote

Einige Bibliotheken haben erkannt, dass die Suchmaschinentechnologie solch entscheidende Vorteile bietet, dass sie um deren Einsatz nicht mehr herumkommen. Ihre Nutzer sind mittlerweile die einfachen Suchinterfaces der Suchmaschinen gewöhnt und erwarten auch von der Suche in Bibliothekskatalogen, dass ohne viel Nachdenken bei der Anfrageformulierung relevante Treffer zurückgegeben werden. Im Folgenden werden die drei nach Ansicht des Verfassers interessantesten Ansätze aus dem deutschsprachigen Raum vorgestellt. Diesen ist gemeinsam, dass sie sich alle noch in einem Teststadium befinden und in ihren Datenmengen noch begrenzt sind. Sehr wohl zeigen sie aber bereits die Vorteile dieser Erschließungsform, aber auch die Grenzen der Anwendbarkeit der Suchmaschinentechnologie auf reine Katalogdaten.

BASE

Mit BASE[18] steht eine Suchmaschine bereit, die Daten des Bibliothekskatalogs (Bereich Mathematik) der UB Bielefeld mit ca. 160 Quellen vor allem aus dem Open-Access-Bereich einheitlich durchsuchbar macht. Die Technologie stammt von FAST, entsprechende Anpassungen an das Umfeld der wissenschaftlichen Informationen wurden vorgenommen. Von den Dokumenten werden neben den bibliographischen Angaben im Fall der Open-Access-Quellen auch die Abstracts erfasst und durchsuchbar gemacht. Die für die Recherche verwendbare Textmenge wird damit gegenüber einfachen Bibliothekskatalogen erhöht, wenn auch nicht in den Maße, wie es für den Nutzer wünschenswert wäre.

Die Suchfunktion wird ergänzt durch auf die Treffermenge angepasste Browsing-Elemente, wodurch eine Einschränkung der Treffermenge einfach möglich ist. Dabei handelt es sich um ein übliches Verfahren zur Einschränkung von Treffermengen, wie es auch von allgemeinen Web-Suchmaschinen verwendet wird.[19]

Der Schritt, den Bibliothekskatalog mit Daten aus weiteren Quellen zu kombinieren und mittels Suchmaschinentechnologie verfügbar zu machen, ist als innovativ anzusehen und der UB Bielefeld hoch anzurechnen. Um das volle Potenzial des Systems aufzuzeigen, wäre jedoch dringend eine deutliche Erweiterung des Quellenspektrums nötig; insbesondere ist es bei einem System, welches laut Aussage der Betreiber gerade durch seine Skalierbarkeit besticht, unverständlich, warum aus dem eigenen Bibliothekskatalog nur ein kleiner Ausschnitt verfügbar gemacht wird.

Dandelon

Das Portal Dandelon[20] bietet eine Suchmaschinenartige Recherche in Bibliotheksbeständen. Neben der Verwendung von Suchmaschinentechnologie ist besonders die Anreicherung der Katalogeinträge durch das Einscannen von Inhaltsverzeichnissen zu erwähnen. Durch Verfahren der maschinellen Indexierung in Kombination mit der Verwendung fachspezifischer Thesauri wird eine wesentlich breitere Erschließung erreicht, als sie manuell von Bibliotheken geleistet wird und geleistet werden kann. Für den Nutzer ergibt sich der Vorteil, dass die Wahrscheinlichkeit wesentlich erhöht wird, durch eine Suchanfrage »in eigenen Worten« relevante Treffer angezeigt zu bekommen. Die eingesetzten Verfahren des Relevance Ranking greifen in diesem System sehr gut, da im Gegensatz zu reinen bibliographischen Angaben genug Text für ein solches Ranking zur Verfügung steht.

Anreicherung der Katalogeinträge

Leider steht auch in diesem System, welches zu Beginn vor allem von der Vorarlberger Landesbibliothek bestückt wurde[21], noch nicht die für den Erfolg kritische Masse an Dokumenten zur Verfügung. Durch das Hinzukommen neuer Projektpartner ist aber mit einem schnellen Anwachsen des Datenbestands zu

fehlende kritische Masse an Dokumenten

rechnen. Der Ansatz ist vor allem in Hinblick auf seine Benutzerfreundlichkeit zu begrüßen.

HBZ Suchmaschine

angereicherte bibliografische Angaben bei Amazon

Die HBZ-Suchmaschine[22] ist das jüngste der drei beschriebenen Angebote und kann bereits auf die Erfahrungen mit BASE und Dandelon aufbauen. Ebenso wie BASE basiert es auf der Technologie von FAST, erschlossen wird der Bestand des HBZ-Verbundkatalogs. Auch hier können die Trefferlisten mit (optional anwählbaren) Auswahlmenüs leicht auf das gewünschte Maß eingeschränkt werden. Das Ranking nach Relevanz und Ähnlichkeitssuchen werden angeboten, aufgrund der rein aus bibliographischen Angaben und der sachlichen Erschließung bestehenden Datensätze kann sich das Ranking allerdings nicht in seiner vollen Nützlichkeit entfalten. Hier ist auf eine Verwendung der Daten aus dem 18oT-Projekt[23] zu hoffen.

Alle drei Anwendungen zeigen den Nutzen des Einsatzes von Suchmaschinentechnologie für (angereicherte) Bibliotheksdaten auf. Typische Probleme der Web-Suchmaschinen wie das Erkennen von Manipulationsversuchen auf unterschiedlichen Ebenen oder die Notwendigkeit einer Qualitätsbewertung der Dokumente[24] entfallen. Die Stärken der Suchmaschinentechnologie lassen sich aber erst vollständig ausspielen, wenn einerseits die Datensätze entsprechend angereichert werden und andererseits auch bisher Katalog-untypische Dokumente eingebunden werden.

Bibliotheksdaten in allgemeinen Suchmaschinen

Titelaufnahmen in die Indizes der Suchmaschinen bringen

Bibliotheken versuchen, die Auffindbarkeit ihrer Inhalte nach außen zu verbessern, indem aus ihren Datenbanken statische HTML-Dokumente generiert werden, die von den Suchmaschinen erfasst werden können. Diese Dokumente weisen auf entsprechende Titel im Bibliotheksbestand hin und zeigen den Weg, wie diese beschafft werden können. Das verwendete Verfahren ist aus dem Bereich der kommerziellen Suchmaschinen-Optimierung bekannt: Neben der Beeinflussung des Rankings wird in einem ersten Schritt die Auffindbarkeit von Dokumenten aus dem zu optimierenden Angebot erhöht, indem unter anderem die für Suchmaschinen nicht indexierbaren Datensätze aus vorhandenen (Produkt-)Datenbanken als HTML-Seiten ausgelesen werden, damit sie für die Suchmaschinen erfassbar sind.

Ein Vorbild solcher Techniken im Buch-Bereich ist der Online-Buchhändler Amazon: Der gesamte Katalog liegt in Form von HTML-Seiten vor, die von den Suchmaschinen erfasst werden können und aller Erfahrung nach von diesen auch für viele Anfragen auf den vorderen Rängen der Trefferlisten angezeigt werden.

Amazon reichert die bibliografischen Angaben der Titel mit einer Vielfalt weiterer Daten an. Diese werden zu einem großen Teil erstens automatisch (und damit ohne weiteren Aufwand für Amazon) und zweitens durch die Nutzer selbst erfasst.

Im Idealfall enthält ein Datensatz dann[25]

- bibliografische Angaben
- klassifikatorische Angaben
- Schlagwörter
- Klappentext
- Besprechungen
- Hinweis auf ähnliche Bücher aufgrund des Kaufverhaltens der Amazon-Kunden
- Hinweis auf ähnliche Bücher aufgrund des Browsingverhaltens auf der Amazon-Website
- wichtige Mehrwortausdrücke aus dem Text
- Zitationen
- von Kunden vergebene *Tags*[26]
- von Kunden erstellte Themenlisten
- beschränkt zugänglicher Volltext (»Search Inside«).

Durch diese umfangreichen Anreicherungen entsteht auf der einen Seite ein für den Nutzer informatives Dokument, das ihm fundierte Informationen für seine Kaufentscheidung bereitstellt, auf der anderen Seite entsteht eine textreiche Seite, die gut von Suchmaschinen erfasst werden kann. Die Popularität von Amazon ist sicher zu einem großen Teil gerade auf diese informationslastigen Produktbeschreibungen zurückzuführen, nicht nur auf die Rolle des Unternehmens als Pionier des Online-Buchhandels.

Das »Virtuelle Bücherregal NRW«

Mit dem »Virtuellen Bücherregal NRW«[27] versucht das HBZ, Titelaufnahmen aus seinem Verbundkatalog in die Indizes der Suchmaschinen zu bringen. Dazu wird aus jeder Titelaufnahme der Verbunddatenbank des HBZ eine HTML-Seite erzeugt, die für die Suchmaschinen problemlos lesbar ist. Da mit diesem Angebot neue Nutzer (d.h. bisherige Nicht-Nutzer der Bibliotheken) erreicht werden sollen, sollen zusätzlich alle unverständlichen Abkürzungen, Sigla etc. aufgelöst und dem Nutzer im Klartext präsentiert werden. Von jeder Seite wird weiter auf die Digitale Bibliothek des HBZ verwiesen, wo nach der Verfügbarkeit und (Fern-)Leihmöglichkeiten recherchiert werden kann.

Nach den vom HBZ veröffentlichten Zahlen[28] sprechen die Suchmaschinen gut auf das Angebot an und indexieren die Seiten in großer Zahl. So soll die Suchmaschine Google 3,2 Millionen Seiten aus dem Virtuellen Bücherregal NRW indexiert haben. Diese Zahl konnte in einer Testrecherche nicht nachvollzogen

werden. Unklar ist, ob die Zahl der indexierten Dokumente inzwischen abgenommen hat und welche Gründe dies ggf. hat.

Die Treffer des Virtuellen Bücherregals werden einheitlich präsentiert, und zwar durch die Titelaufnahme aus dem Verbundkatalog inklusive Bestandsangaben der besitzenden Bibliotheken. Autor, Titel und besitzende Bibliothek sind verlinkt. Ein Klick auf Autor oder Titel führt zu einer Seite mit weiteren Informationen, wie der Titel beschafft werden kann; ein Klick auf die besitzende Bibliothek führt zum Siglenverzeichnis Der Deutschen Bibliothek.

Bibscout

Einen ähnlichen Ansatz wie das »Virtuelle Bücherregal NRW« verfolgt das Angebot »Bibscout«[29] des Bibliotheksservicezentrums Baden-Württemberg. Es sieht sich als »systematischer Katalog, themenorientiertes Internet-Verzeichnis und Buch- und Bibliotheksführer«. Es werden Titellisten nach der Systematik präsentiert, im Umfeld werden Schlagworte angezeigt. Durch diese Ergänzungen in Verbindung mit dem auf jeder Einzelseite aufgeschlagenen Systematikbaum entsteht pro Seite eine relativ große Textmenge, die zu entsprechend vielen Treffern durch Suchmaschinen führen dürfte.

Zu den einzelnen Titeln wird allerdings nicht mehr als die bibliographischen Angaben geboten. Klickt man auf den zu jedem Titel vorhandenen Link »Bibliotheken«, so wird die Anfrage an den Verbundkatalog des BSZ weitergereicht.

Sowohl beim »Virtuellen Bücherregal NRW« als auch bei »Bibscout« will sich der Nutzen nicht so recht erschließen. Ziel beider Angebote soll es sein, Nutzer, die normalerweise nicht im Bibliothekskatalog recherchieren, auf dieses Angebot zu lenken. Dies mag – betrachtet man die Aufrufstatistiken der beiden Angebote[30] – auch durchaus der Fall sein, jedoch werden bei beiden Angeboten nur Angaben über die Aufrufzahlen gemacht, nicht jedoch über die Konversionsrate. Während sich in der Suchmaschinenoptimierer-Szene die Ansicht durchgesetzt hat, dass es eben nicht allein auf die Zahl der Seitenaufrufe ankommt, sondern auf getätigte Abverkäufe (oder Transaktionen wie Newsletter-Bestellungen usw.), wird diese Zahl (zum Beispiel in Form von resultierenden Anfragen, Ausleihen oder Fernleihbestellungen) bei den bibliothekarischen Angeboten wohl nicht beachtet.

Ein weiterer Kritikpunkt ist die mangelnde Aussagekraft der in die Suchmaschinen eingespeisten HTML-Dokumente. Sie alleine sind wenig aussagekräftig und stellen als Weiterleitungen auf die Bibliothekskataloge reine Brücken- oder »Teaser«-Seiten dar[31]. Solche Seiten werden im Bereich der Produkte üblicherweise von »Affiliates« erstellt, die auf große Händler wie Amazon oder Conrad-Elektronik verweisen und sich durch Provisionen aus dort getätigten Verkäufen finanzieren. Es ist sicherlich zu diskutieren, inwieweit Teaser-Seiten prinzipiell »schlecht« sind oder gar als Spam betrachtet werden sollten[32]; Tatsache ist jedoch, dass solche Seiten zumindest im kommerziellen Bereich für die Suchmaschinen zum Problem geworden sind. Solange es nur zwei entsprechende Angebote von Bibliotheksseite gibt, mag man ihre – wohl von vielen Nutzern als irrelevant angesehenen – Treffer in den Suchmaschinen als zu vernachlässigen ansehen. Stellt man sich jedoch vor, dass andere Bibliotheken oder Bibliotheksverbünde dem Beispiel folgen, dürfte man bald die immergleichen Büchertreffer in den Suchmaschinen finden. Wie dies auch bei der Häufung einander ähnlicher kommerzieller Dokumente der Fall ist, dürften dann auch diese zu einem großen Teil von den Suchmaschinen ausgeschlossen werden.

Teaser-Seiten gut oder schlecht?

Integration der Informationsbestände in ein Recherchesystem

Aus den obigen Ausführungen ergeben sich mehrere Konsequenzen für den Aufbau von Bibliotheksangeboten, die sich an den tatsächlichen Gewohnheiten der Nutzer orientieren und dem Nutzer dahingehend entgegenkommen wollen, dass sie sein Nutzungsverhalten akzeptieren, anstatt ihn zu einem besseren Rechercheur erziehen zu wollen.

Orientierung an den tatsächlichen Gewohnheiten der Nutzer

Die zentrale Forderung lautet, dass der OPAC (oder die »Bibliotheks-Suchmaschine«) das zentrale Nachweisinstrument der Bibliothek sein muss. In ihm müssen *alle* in der Bibliothek vorhandenen Dokumente sowie die für den Nutzer der jeweiligen Bibliothek relevanten, anderweitig im Web verfügbaren Dokumente recherchierbar sein. Die Trennung zwischen dem Nachweis von Dokumenten und Sucheinstiegen muss aufgehoben werden. Dem Nutzer ist nicht zuzumuten, sich mit unterschiedlichen Rechercheeinstiegen zu beschäftigen – er muss dabei unterstützt werden, möglichst schnell und problemlos an die von ihm gewünschten Informationen zu gelangen.

Um Suchmaschinentechnologie sinnvoll einsetzen zu können, müssen die Katalogdaten angereichert werden. Ein alleiniger Einsatz bibliothekarischer Erschließungsinstrumente ist dafür nicht ausreichend. Verlagsdaten wie Inhaltsverzeichnisse und Klappentexte sollten durch weitere, durch die Nutzer »geschaffene« Daten ergänzt werden. Die automatische Einbindung von Fachthesauri kann das Vokabular für die Recherche deutlich erweitern, so dass sich

mangelnde Aussagekraft der in die Suchmaschinen eingespeisten HTML-Dokumente

»Suche wie in Google«

die Recherche in ihrem Vokabular dem Nutzer anpasst, anstatt dass sich der Nutzer mit Vorzugsbenennungen o.ä. aus einem kontrollierten Vokabular beschäftigen muss. Erst die Anreicherung und Erweiterung der Bibliotheksdaten ermöglicht eine »Suche wie in Google«, wie sie auch von bibliothekarischer Seite zunehmend als wichtig angesehen wird.

Neben der wesentlich verbesserten Recherche ist aber auch die Auffindbarkeit nach außen zu verbessern. Ähnlich wie es bereits von Amazon bekannt ist, sollten die *angereicherten* Daten der Bibliothekskataloge in einfache HTML-Seiten umgesetzt und damit den allgemeinen Suchmaschinen zur Indexierung zur Verfügung gestellt werden. Diese Aufgabe sollte aber nicht jede Bibliothek einzeln in Angriff nehmen, vielmehr sollte hier eine zentrale Lösung angestrebt werden, um unnötige (Nahezu-) Dubletten in den Suchmaschinen-Indizes zu vermeiden.

Erhöhung der Attraktivität der Bibliotheksangebote und Auffindbarkeit ihrer Inhalte

Mit den beschriebenen Maßnahmen ließe sich sowohl die Attraktivität der Bibliotheksangebote als auch die Auffindbarkeit ihrer Inhalte wesentlich verbessern. Sie könnten damit unter den durch die Suchmaschinen und das von diesen übertragene Nutzerverhalten veränderten Bedingungen (wieder) zu einem attraktiven Rechercheinstrument werden.

[1] Aber auch auf der Dozentenseite stellt die Studie große Defizite im Umgang mit elektronischen Informationsquellen fest. S. Klatt, Rüdiger; Gavriilidis, Konstantin; Kleinsimlinghaus, Kirsten; Feldmann, Maresa: Nutzung elektronischer wissenschaftlicher Information in der Hochschulausbildung: Barrieren und Potenziale der innovativen Mediennutzung im Lernalltag der Hochschulen (2001). www.stefi.de/download/bericht2.pdf [Stand 29.12.2005]

[2] Klatt et al. 2001.

[3] Klatt et al. 2001.

[4] Die unvollständige Repräsentation ist vor allem in der größtenteils nicht vorhandenen Erfassung der Zeitschriftenaufsätze, der Beiträge zu Sammelwerken zu sehen.

[5] Vgl. Devine, J.; Egger-Sider, F.: Beyond Google: The invisible web in the academic library. Journal of Academic Librarianship 30 (2004), S. 265–269.

[6] www.ub.uni-duesseldorf.de/ebib/ [14.12.2005]

[7] Inwieweit der OPAC von Nutzern als *alleiniges* Rechercheinstrument angesehen wird, konnte aus der Literatur nicht ermittelt werden. Es ist jedoch anzunehmen, dass es sich um einen signifikanten Anteil der Nutzerschaft handelt.

[8] Leider liegen keine Untersuchungen von OPAC-Logfiles vor, die das Nutzerverhalten in dieser Hinsicht beleuchten.

[9] Lewandowski, Dirk: Web Information Retrieval: Technologien zur Informationssuche im Internet. Frankfurt am Main: DGI, 2005.

[10] Die Nützlichkeit bzw. Akzeptanz von Rankingverfahren hängt auch mit der Art der Anfragen zusammen. In professionellen Datenbanksystemen wie Lexis-Nexis und Dialog wurden Rankingverfahren von den Nutzern nur schlecht angenommen, da diese spezielle Nutzerschaft in der Lage ist, ihre Suchanfrage präzise zu formulieren und damit die Treffermengen auf ein überschaubares Maß einzugrenzen.

[11] Sherman, Chris; Price, Gary: The Invisible Web: Uncovering Information Sources Search Engines Can't See. Medford, NJ: Information Today, 2001.

[12] Auch dieses Problem besteht in ähnlicher Weise im Bereich der Suchmaschinen. Manche Suchmaschinen (wie Teoma) bieten neben der durchmischten Trefferliste auch einen gesonderten Trefferbereich mit Hinweisen auf weitere Quellen (bei Teoma Quellenlisten, die unter Umständen auch Invisible-Web-Quellen enthalten).

[13] Lewandowski, Dirk: Google Scholar – Aufbau und strategische Ausrichtung des Angebots sowie Auswirkung auf andere Angebote im Bereich der wissenschaftlichen Suchmaschinen. www.durchdenken.de/lewandowski/doc/Expertise_Google-Scholar.pdf [Stand: 13.12.2005]

[14] Mayr, Philipp; Walter, Anne-Kathrin: Google Scholar – Wie tief gräbt diese Suchmaschine? In: In die Zukunft publizieren: Herausforderungen an das Publizieren und die Informationsversorgung in den Wissenschaften. Bonn, 2005.

[15] s. Mayr u. Walter 2005

[16] www.scirus.com/ [Stand: 19.12.2005]. Siehe auch Scirus White Paper: How Scirus works. [Stand: 29.12.2005]

[17] www.forschungsportal.net/ [Stand: 19.12.2005]

[18] www.base-search.net/ [Stand: 15.12.2005]. Lossau, Norbert: Suchmaschinentechnologie und Digitale Bibliotheken – Bibliotheken müssen das wissenschaftliche Internet erschließen. In: ZfBB 51(2004), S. 284–294. Summann, Friedrich; Lossau, Norbert. Suchmaschinentechnologie und Digitale Bibliotheken: Von der Theorie zur Praxis. In: ZfBB 52(2005), S. 13–17. Summann, Friedrich; Wolf, Sebastian:. Suchmaschinentechnologie für digitale Bibliotheken. In: Information Wissenschaft und Praxis 56(2005), S. 51–57.

[19] Zu den benutzerleitenden Verfahren siehe Lewandowski 2005, S. 139–167.

[20] www.dandelon.com [Stand: 16.12.2005]. Hauer, Manfred: Neue Qualitäten in Bibliotheken: Durch Content-Ergänzung, maschinelle Indexierung und modernes Information Retrieval können Recherchen in Bibliotheken deutlich verbessert werden. In: ABI-Technik 24 (2004), S. 262–268.

[21] Rädler, Karl: In Bibliothekskatalogen »googlen«: Integration von Inhaltsverzeichnissen, Volltexten und WEB-Ressourcen in Bibliothekskataloge. In: Bibliotheksdienst 38 (2004), S. 927–939.

[22] http://suchen.hbz-nrw.de/search/ [Stand: 16.12.2005]

[23] Großgarten, Astrid: Das 180T-Projekt in Köln oder wie verarbeite ich 180.000 Bücher in vier Monaten. In: Information Wissenschaft und Praxis 56(8), S. 454–456.

[24] Lewandowski 2005, S. 132–136, S. 191–216.

[25] Die Auflistung bezieht sich auf die Amazon.com-Website. Die deutsche Amazon-Version enthält nicht alle beschriebenen Elemente, für die Zukunft kann aber mit deren Einbindung gerechnet werden. Ein Beispiel mit der Einbindung der genannten Elemente findet sich unter www.amazon.com/gp/product/1558607544/104-0062207-8919120

[26] Hier wird eine Form der Verschlagwortung (allerdings ohne kontrolliertes Vokabular) durch die Nutzer vorgenommen. Diese Form der Erschließung wird zunehmend populär (Stichwort »Folksonomies«), s.a. Morville, Peter: Ambient Findability. Sebastopol: O'Reilly, 2005.

[27] http://kirke.hbz-nrw.de/dcb/ [Stand: 15.7.2005]. Seiffert, Florian: Das »Virtuelle Bücherregal NRW«: Literatursuche mit der einfachsten Suchstrategie: Google und Co. BuB 55(2003), S. 379–397.

[28] Seiffert 2003; Seiffert, Florian: Wie indexieren Google & Co 13 Milionen Seiten? (2004). www.florian-seiffert.de/2004/Bonn/Inetbib2004.pdf [Stand: 15.12.2005]

[29] http://bibscout.bsz-bw.de/bibscout/ [Stand: 19.12.2005]

[30] Die Nutzungsstatistik von Bibscout findet sich unter http://titan.bsz-bw.de/bibscout/statistics/index_html/view [Stand: 30.12.2005]; Zahlen zur Nutzung des »Virtuellen Bücherregals NRW« finden sich in Seiffert 2003.

[31] Heinisch, Christian: Suchmaschinen des Surface Web als Promotoren für Inhalte des Deep Web – Wie Doorway-Pages als »Teaser« zu Datenbank-Inhalten in die Index-Files der Suchmaschinen gelangen. In: Competence in Content, 25. Online-Tagung der DGI, Frankfurt am Main, 2003, S. 13–24.

[32] Heinisch 2003.

Der Verfasser

Dirk Lewandowski ist freier Berater zum Themenfeld Suchtechnologie und Lehrbeauftragter an der Heinrich-Heine-Universität Düsseldorf, Institut für Sprache und Information, Abt. Informationswissenschaft, Universitätsstr. 1, 40225 Düsseldorf, dirk.lewandowski@uni-duesseldorf.de