Integration von Web-Verzeichnissen in algorithmische Suchmaschinen

in: Leitbild Informationskompetenz: Positionen - Praxis - Perspektiven. 27. DGI-Online-Tagung, Proceedings. Frankfurt am Main, 2005

Dirk Lewandowski, Düsseldorf

1 Einleitung

Suchmaschinen sind bemüht, den Nutzern möglichst hochwertige Ergebnisse zu liefern. Dass dies nur zum Teil gelingt, zeigen neben der persönlichen Erfahrung wohl eines jeden Suchmaschinennutzers zahlreiche Retrievaltests (s. z.B. Griesbaum et al. 2002; Griesbaum 2004), denen übereinstimmend zu entnehmen ist, dass es den Suchmaschinen nur zum Teil gelingt, relevante Treffer in die Ergebnislisten zu bringen.

Eine Qualitätsbewertung findet im Ranking statt, indem mit klassischen textstatistischen Verfahren relevante Dokumente gefundenen werden, deren Reihenfolge dann ergänzend durch von linktopologischen Verfahren errechnete statische oder dynamische Qualitätsmaße verändert wird. Neben der Auswertung der Verlinkungsstruktur des Web werden ergänzende Qualitätsmaße vorgeschlagen, um das Ranking weiter zu verbessern (Mandl 2005).

Neben dem alleinigen Ranking, welches alle im Datenbestand der Suchmaschine enthaltenen Dokumente berücksichtigt, berücksichtigen einige Suchmaschinen auch Quellen, die gesondert gelistet werden (Lewandowski 2004, 186ff.). Dies können sowohl externe Quellen sein, die als besonders hochwertig angesehen werden als auch Quellen aus dem Angebot der Suchmaschine selbst. Der letzte Fall ist vor allem bei Suchmaschinen, die gleichzeitig ein Portalangebot betrieben, zu sehen (z.B. bei Yahoo).

Während im Ranking einzelne *Dokumente* nach ihrer Qualität in Relation zur Suchanfrage gelistet werden, werden im letztgenannten Fall ganze Informationsressourcen (also *Quellen*) bevorzugt angezeigt. Bei den populären Suchmaschinen sind dies jeweils nur wenige, für ihren Themenbereich hoch relevante Quellen, wobei es sich zusätzlich um solche Quellen handelt, die von den Suchmaschinen nicht oder nur schwer erschlossen werden können bzw. die wegen ihres Umfangs von den Suchmaschinen nicht vollständig erschlossen werden. Ein Beispiel hierfür ist die Patentdatenbank des US-Patentamts, deren Dokumente (in Form von einzelnen HTML-Seiten) zwar grundsätzlich in Google enthalten sind. Bei der Eingabe des Worts *patent* und einer Patentnummer verweist Google aber oberhalb der regulären Suchergebnisse zusätzlich auf die Patendatenbank.[1]

Ein Hinweis auf qualitativ hochwertige Quellen erfolgt schließlich in den Trefferlisten der Suchmaschinen durch die Anzeige von Treffern aus einem Web-Verzeichnis bzw. durch die Anzeige von passenden Verzeichniskategorien. Wie auch bei den eingebundenen Informationsressourcen werden die Quellen in den Verzeichnissen manuell ausgewählt und bieten somit eine geprüfte Qualität. Damit sind solche Treffer auch weniger anfällig für Spamming-Versuche.

1. Die Patennummernsuche ist bisher nur auf Google.com verfügbar. Andere auch auf der deutschen Suchoberfläche eingebundene Quellen werden unter http://www.google.de/intl/de/features.html aufgeführt.

In den letzten Jahren sind die allgemeinen Web-Verzeichnisse gegenüber den Suchmaschinen deutlich ins Hintertreffen geraten. Alleinige Verzeichnisse bestehen nur noch selten, meist werden sie in Verbindung mit einer algorithmischen Suchmaschine angeboten. Aber auch bei den Suchmaschinen sind die Verzeichnisse inzwischen weniger prominent platziert; das vielleicht deutlichste Beispiel ist Yahoo, dessen ursprüngliches Angebot ja nur aus einem Verzeichnis bestand. Inzwischen findet sich das Verzeichnis nur noch wenig prominent platziert unter zahlreichen anderen Angeboten.

Dass Verzeichnistreffer gerade für eine hochwertige Suche in algorithmischen Suchmaschinen geeignet sind, soll in diesem Aufsatz gezeigt werden. Als größtes Hindernis für die Nutzung der Verzeichnistreffer ist deren bisher nur mangelhafte Einbindung in die Trefferlisten zu sehen. Damit wird der große Nutzen, der sich aus diesen intellektuell ausgesuchten Informationsressourcen ziehen ließe, nicht vollständig ausgenutzt.

Klassisch werden von Suchmaschinen und Web-Verzeichnissen unterschiedliche Such-Paradigmen erfüllt. Zur Verdeutlichung sollen hier noch einmal kurz die Paradigmen der Websuche nach Dennis, Bruza u. McArthur (2002) dargestellt werden. Diese sind

1. die ununterstützte Stichwortsuche (*unassisted keyword search*)
2. die unterstützte Stichwortsuche (*assisted keyword search*), wobei die Unterstützung vor allem durch automatisch generierte Vorschläge zur Einschränkung der Suche erfolgt.
3. die verzeichnisbasierte Suche (*directory-based search*)
4. das Auffinden ähnlicher Dokumente (*query-by-example*)

Suchmaschinen unterstützen Punkt 1, teilweise auch Punkt 2 und Punkt 4. Punkt 3 betrifft die Web-Verzeichnisse. Im Folgenden soll es nach der Beschreibung der bisherigen Ansätze der Kombination von Suchmaschine und Verzeichnis um die Frage gehen, wie sich die verzeichnisbasierte Suche vor allem mit der einfachen Stichwortsuche verbinden lässt.

2 Erschließung des Web mittels Suchmaschinen und Verzeichnissen

Hauptunterscheidungsmerkmal zwischen Web-Verzeichnissen und Suchmaschinen ist, dass erstere von Menschen erstellt werden, d.h. dass Redakteure für Auswahl und Erschließung geeigneter Sites sorgen. Daher kann gegenüber den Suchmaschinen nur eine relativ geringe Zahl von Sites erfasst werden. Während die Suchmaschinen Indizes bis zu einer Größe von etwa acht Milliarden Dokumenten aufgebaut haben[2], gibt das größte Webverzeichnis an, über vier Millionen Websites erschlossen zu haben[3]. An dieser Stelle ist es allerdings wichtig, zwischen der Indexierung von Web-Seiten, wie sie in Suchmaschinen geschieht, und der Indexierung von Web-Sites, wie sie von Web-Verzeichnissen durchgeführt wird, zu unterscheiden. Eine einzige Site kann aus tausenden von Seiten bestehen; die Anzahl der in Suchmaschinen und Verzeichnissen erschlossenen Dokumente lässt sich also nur bedingt vergleichen. Individuelle Dokumente werden in Verzeichnissen in der Regel nicht erschlossen.

2. Diese Zahl wird von Google für den eigenen Index angegeben. Schätzungen zufolge handelt es sich dabei um den weltweit größten Suchmaschinen-Index; die meisten anderen der großen Anbieter veröffentlichen keine Zahlen zu ihrem Datenbestand.
3. Die Angaben stammen von der Startseite des Open Directory Project (http://www.dmoz.org)

Die Größe des Web liegt zwischen etwa 32 und 42 Millionen Servern (vgl. How much Information 2003). Die Anzahl der auf diesen Servern enthaltenen Dokumente variiert sehr stark, so dass sich aufgrund ermittelter Durchschnittszahlen aus Stichproben keine guten Schätzungen über die Gesamtzahl der Dokumente ableiten lassen. Die Abdeckung der Suchmaschinen liegt im Vergleich zu den Web-Verzeichnissen deutlich höher; allerdings sollten die Zahlen in der Relation betrachtet werden. Die durchschnittliche Anzahl der Seiten pro Server liegt nach der Untersuchung von Lawrence u. Giles (1999) bei 289. Damit ergäbe sich, dass die momentan größte Suchmaschine 27,6 Millionen Server indexiert. Will man diese Zahl nun mit der Zahl der von den Verzeichnissen indexierten Servern vergleichen, ergeben sich folgende Probleme: Es ist davon auszugehen, dass die Verzeichnisse bevorzugt Websites erschließen, die viele Dokumente enthalten. Andererseits gibt es Server, die eine große Anzahl von Websites enthalten, beispielsweise Hosting-Angebote für private Homepages.[4] Außerdem wird in einer solchen Berechnung davon ausgegangen, dass die Suchmaschinen jeden Server komplett indexieren.

Verlässliche Zahlen über den Anteil der indexierten Dokumente in Suchmaschinen und Web-Verzeichnissen sind also nicht zu ermitteln. Letztlich wird von Verzeichnissen oft auch nicht angestrebt, einen möglichst hohen Anteil des Web zu erfassen, sondern es wird gezielt auf die Qualität der erschlossenen Quellen geachtet.

Eine weitere Unterscheidung zwischen Suchmaschinen und Verzeichnissen zeigt sich in der hierarchischen Anordnung der Dokumente innerhalb von Verzeichnissen. Jedes Dokument wird hier einer oder mehrerer Klassen zugeordnet. Suchmaschinen bieten keine vergleichbare Einordnung. Ein weiterer großer Unterschied zwischen den beiden Formen der Erschließung des Web ist der Grad der Indexierung. Während Suchmaschinen den Volltext jeder gefundenen Seite indexieren, beschränken sich die Verzeichnisse auf eine kurze Beschreibung des Inhalts der kompletten Site. Dafür wird diese Beschreibung intellektuell erstellt und bietet über den Volltext hinausgehende Metainformationen zu der erfassten Website.

Es gibt sowohl umfassende (allgemeine) als auch themenspezifische Webverzeichnisse. Allgemeine Verzeichnisse wie das Open Directory Project (ODP) oder das Yahoo-Verzeichnis versuchen, Sites zu allen möglichen Themen zu erschließen und gehen weniger in die Tiefe als spezifische Verzeichnisse. Diese enthalten zu ausgesuchten Themen meist eine wesentlich höhere Anzahl von Quellen und erschließen diese wesentlich genauer.

Keine der großen Suchmaschinen hat bisher spezifische Verzeichnisse in seiner Trefferlisten integriert, während eine rudimentäre Integration allgemeiner Verzeichnisse die Regel ist.

Relativ viele Arbeiten beschäftigen sich mit den Themen automatische Klassifikation von Webseiten (vgl. u.a. Chung u. Noh 2003) und automatische Einordnung von Webseiten in eine bestehende Klassifikation (vgl. u.a. Wätjen 1998). Die Integration von bestehenden Webverzeichnissen in Suchmaschinen wird allerdings in der aktuellen Forschung nicht diskutiert. Dies mag mit der Annahme zusammenhängen, dass mit der bisher schon bestehenden rudimentären Integration der Webverzeichnisse in Suchmaschinen das Problem gelöst sei. Im Folgenden wird jedoch angenommen, dass durch eine verbesserte Integration der Verzeichnisergebnisse die Websuche effektiver gestaltet werden kann.

4. Angebote wie geocities.com hosten tausende von Homepages.

3 Web-Verzeichnisse und ihre Integration in Suchmaschinen

Der Ansatz der Webverzeichnisse, die Quellen durch Menschen erschließen zu lassen, beschränkt die Erschließung auf ausgewählte Websites. Alle Verzeichnisse haben Richtlinien für die Aufnahme der Sites in das Verzeichnis[5] und versuchen, nur Sites, die eine bestimmte Qualität erreichen, zu listen. Ob es den Verzeichnisbetreibern gelingt, tatsächlich nur Seiten von hoher Qualität in die Kataloge aufzunehmen, kann hier nicht umfassend diskutiert werden. Qualitätsprobleme sind allerdings in der Hinsicht vorhanden, dass auch in Verzeichnissen teilweise Sites von schlechter Qualität oder sogar Spam-Sites auftauchen, allerdings weit seltener als in den Trefferlisten der Suchmaschinen. Im Folgenden wird angenommen, dass die Kategorien der Verzeichnisse in der Regel eine Auswahl qualitativ hochwertiger Sites enthalten und diese Kategorien deshalb als ein guter Ausgangspunkt für themenbezogene Anfragen dienen können.

Webverzeichnisse sind vor allem für die folgenden Zwecke nützlich:

- Webverzeichnisse können das Problem mehrdeutiger Anfragen einschränken. Durch Benutzung der Klassifikation kann man die Anfrage auf eine passende Klasse (und deren Unterklassen) eingeschränken und Polysemie-Probleme dadurch mindern; eine Trennung zwischen kommerziellen und nicht kommerziellen Treffern kann erfolgen.
- Verzeichnisse können genutzt werden, wenn keine geeigneten Suchbegriffe für das Themenfeld bekannt sind. Hierzu wird auf die Navigation entlang der Verzeichnisebenen zurückgegriffen; die Eingabe von Suchbegriffen ist nicht nötig.
- Mit Hilfe von Webverzeichnissen lassen sich thematisch verwandte Dokumente finden. Ausgehend von einer bekannten Website, welche in einem Verzeichnis enthalten ist, können weitere Sites gefunden werden, welche derselben Klasse zugeordnet sind. Hier zeigt sich ein wesentliches Problem der bisherigen Verzeichnisintegration: Wenn der Nutzer eine Suche innerhalb aller Quellen einer Verzeichnisklasse ausführen will, so muss er jede Site einzeln anwählen und mittels der dort vorhandenen Site-Suche durchsuchen. Die Suchmaschinen bieten ihm keine Möglichkeit, alle Dokumente einer Klasse direkt zu durchsuchen.
- Die Struktur von Webverzeichnissen kann genutzt werden, um eine hierarchische Visualisierung zu unterstützen und um Navigationshilfen zu erstellen (Chakrabarti 2003, 126).

Im Folgenden sollen die ersten drei Punkte genauer behandelt werden, die im letzten Punkt genannten Anwendungen gehen über die Zielsetzung dieses Aufsatzes hinaus.

Suchmaschinen binden Verzeichniseinträge auf zwei verschiedene Arten ein. Am häufigsten wird in den Trefferlisten zu jedem Eintrag eine Verzeichniskategorie angezeigt, sofern eine solche vorhanden ist. Eine solche Integration findet sich beispielsweise in den großen Suchmaschinen Google und Yahoo. Damit lassen sich zu einem Treffer verwandte Seiten finden, die in der gleichen Klasse des Verzeichnisses einsortiert sind. Ähnliche Sites bzw. Seiten können teils auch über automatisierte Verfahren („related pages“) gefunden werden; diese arbeiten jedoch bei weitem nicht so zuverlässig wie die manuelle Klassifikation.

Die zweite bisher genutzte Möglichkeit ist es, passende Kategorien oberhalb der Trefferlisten mit den algorithmischen Ergebnissen anzuzeigen. Eine solche Anwendung findet sich

5. Z.B. http://help.yahoo.com/help/us/dir/basics/basics-09.html [13.12.2004]

zum Beispiel bei Yahoo, allerdings nur in der Verzeichnis-Suche. Es erscheint verwunderlich, dass ein solcher Hinweis auf eine passende Kategorie (also einer Linksammlung zum Thema) – auch bei anderen Suchmaschinen – nicht in der regulären Suche genutzt wird. Algorithmische Ansätze wie Kleinbergs HITS (Kleinberg 1999) versuchen, von Menschen erstellte Linksammlungen zu finden und an prominenter Stelle auf den Ergebnisseiten anzuzeigen. Eine Anwendung hierfür ist die Suchmaschine Teoma[6], die neben den algorithmischen Ergebnissen im Hauptteil der Trefferliste in einer gesonderten Spalte Hinweise auf Linksammlungen zum Thema gibt. Diese Linklisten kommen nicht unbedingt aus den großen Verzeichnissen, sondern sind im Regelfall singuläre Linklisten, die nicht unbedingt eine systematische Aufarbeitung eines Themenbereichs bieten. Das Verfahren von Kleinberg bezieht sich wie alle anderen linktopologischen Verfahren auch allerdings auf einzelne Dokumente und betrachtet nicht ganze Sites als Informationsressourcen, was auch verschiedentlich kritisiert wurde (u.a. in Mandl 2003).

Schon heutige Anwendungen von Verzeichnisdaten gehen allerdings über die alleinige Bereitstellung eines kompletten Verzeichnisses innerhalb der Seiten einer Suchmaschine hinaus. So reichert etwa Google die von ODP übernommenen Verzeichnisdaten mit seinen eigenen PageRank-Werten (vgl. Page et al. 1998) an. Die Sites werden innerhalb einer Kategorie nicht wie in anderen Suchmaschinen oder in ODP selbst in alphabetischer Ordnung angezeigt, sondern werden nach ihrem PageRank-Wert sortiert. Dies soll gewährleisten, dass auch innerhalb der Verzeichnisklassen die wichtigsten Sites zuerst angezeigt werden. Eine solche Qualitätsmessung könnte auch dafür eingesetzt werden, einen Schwellenwert zu bestimmen, bis zu welchem Verzeichniseinträge in einer Suche berücksichtigt werden. Damit könnten beispielsweise aus großen Verzeichnisklassen nur die besten Einträge für eine weitere Suche verwendet werden, um eine „Qualitätssuche" durchzuführen.

4 Erschließung der Sites in Web-Verzeichnissen

In den allgemeinen Web-Verzeichnissen werden die einzelnen Websites nur knapp beschrieben; neben dem Link, der Kategorienzuordnung und der Beschreibung werden keine weiteren Informationen erfasst. Auch die Beschreibungen selbst sind nicht einheitlich verfasst, so dass der Informationsgehalt stark variiert. Viele der Beschreibungen sind von den Anbietern der entsprechenden Websites selbst erstellt worden und wurden von den Verzeichnissen nach Prüfung einfach übernommen. Ebenso wird die Kategorie meist von den Website-Betreibern vorgeschlagen, so dass sich ähnliche Seiten oft in unterschiedlichen Kategorien wiederfinden.

Auch die von den Editoren der Verzeichnisse geschriebenen Beschreibungen der Sites sind keineswegs einheitlich oder verwenden gar ein kontrolliertes Vokabular. Vielmehr geht es um kurze, prägnante Beschreibungen, die es dem Nutzer ermöglichen, schon beim Querlesen der Ergebnisseite die für ihn relevanten Sites zu erkennen (vgl. Hamdorf 2004, 224).

Stock u. Stock (2000) kritisieren die bei den großen Verzeichnissen verwendeten Klassifikationssysteme. Anstatt auf etablierte Systeme zurückzugreifen, haben sowohl Yahoo! als auch Open Directory eigene Klassifikationen entwickelt, die mit der Zeit „gewuchert“ seien, so

6. www.teoma.com [29.3.20005]

dass von einem einheitlichen Aufbau nicht mehr gesprochen werden könne. Die Klassifikation von Yahoo! ist zum Teil polyhierarchisch aufgebaut; bei ODP finden sich häufig Klassen, deren Unterklassen schlicht die Buchstaben des Alphabets tragen. Stock u. Stock (2000, 30) sehen dies als „Kapitulation vor den Problemen einer thematischen Ordnung."

In der Tat ist die Ordnung der Verzeichnisse als problematisch auch für deren Einbindung in Suchmaschinen zu sehen. Vor allem im Open Directory, das wegen seiner freien Nutzbarkeit für alle Suchmaschinen als Verzeichnis attraktiv wäre, finden sich ähnliche bzw. zusammengehörende Einträge oft in unterschiedlichen Klassen. Dies trifft zum Beispiel bei der Suche nach den Hochschulinstituten der Informationswissenschaft zu: Diese werden teils unter „Wissenschaft: Geisteswissenschaften: Fakultäten und Institute", teils aber auch unter „Wissenschaft: Informatik: Fakultäten und Institute: Deutschland" oder „Wissen: Bildung: Hochschulen: Deutschland: Nordrhein-Westfalen: Fachhochschule Köln" gelistet.

Das gleiche Beispiel, diesmal im Yahoo-Verzeichnis, zeigt als weiteres großes Problem die mangelnde Vollständigkeit. Zwar existiert in diesem Verzeichnis eine eigene Kategorie, in der die Institute zusammen aufgeführt sind, dort finden sich jedoch nur neun der insgesamt 15 vom Hochschulverband Informationswissenschaft aufgeführten deutschen Institute[7].

Fragwürdig ist auch, ob sich die Kategorie an der Stelle in der Hierarchie findet, an der der Nutzer sie vermuten würde. Bei Yahoo liegt sie auf der Hierarchieebene „Nachschlagen > Bibliotheken > Bibliotheks- und Informationswissenschaft > Ausbildung und Beruf > Hochschulinstitute".

5 Datenbankauswahl bei Online-Hosts

Datenbank-Hosts erfüllen zum Teil eine ähnliche Aufgabe wie Web-Verzeichnisse: Sie bieten unter einer Oberfläche ein Verzeichnis relevanter Quellen, die für die Recherche ausgewählt werden können: „The Web directories are aggregators – they do for Web sites what proprietary online services do for individual databases" (O'Leary 1998, 79). Im Folgenden soll beschrieben werden, welchen Nutzen die Hosts bei der Auswahl geeigneter Quellen für die Recherche und die Einschränkung der Suche auf bedeutende Quellen bieten. Daraus werden Möglichkeiten für Suchmaschinen abgeleitet, ihre Qualitätseinschränkungen auf ähnliche Art zu verbessern.

Alle Hosts haben erkannt, dass eine gleichzeitige Suche in *allen* verfügbaren Quellen nur selten die vom Nutzer gewünschten Ergebnisse bringt. Vielmehr ist eine gezielte Quellenauswahl mit für den Sucherfolg entscheidend. Bei Lexis-Nexis findet sich beispielsweise eine Datenbankgruppe „Major World Publications", die die als am wichtigsten angesehenen Nachrichtenquellen der Welt enthält. Eine ähnliche Datenbank-Gruppe ist die „Manager-Kombi" bei Genios, die die wichtigsten deutschsprachigen Zeitungen und Nachrichtenmagazine enthält. Die Quellen, die in diesen Datenbankgruppen enthalten sind, wurden von Hand ausgewählt. Dabei ist die Auswahl der Quellen für eine Datenbankgruppe für den Nutzer nachvollziehbar: Welches die bedeutendsten deutschen Tageszeitungen sind oder welche Fachzeitschriften einer Rubrik wie „Medien und Kommunikation" zuzuordnen sind, dürfte relativ unstrittig sein.

7. http://www.informationswissenschaft.org/institutionen/intro.htm [23.3.2005]

Eine weitere Möglichkeit bieten Funktionen wie die „Cross-Suche“, die eine Recherche über alle Quellen des Hosts ermöglichen, wobei nicht die Trefferlisten mit den Dokumenten angezeigt werden, sondern nur die Zahl der in der jeweiligen Datenbank vorhandenen Dokumente. Diese Art der Suche soll es dem Nutzer erleichtern, die für seine Recherche wichtigsten Quellen auszuwählen. Dies werden in der Regel diejenigen sein, die die meisten Dokumente zum Thema enthalten bzw. diejenigen fachlich spezialisierten Quellen, die zumindest eine gewisse Anzahl von passenden Dokumenten enthalten.

Der Ansatz der „Top-Quellen“ beruht auf der Einsicht, dass Suchanfragen zu einem bedeutenden Teil eher zu viele als zu wenige Treffer liefern. Es erfolgt eine Konzentration auf die wichtigen Quellen, gleichzeitig werden die weniger bedeutenden Quellen ausgeschlossen, um Ballast zu vermeiden.

Die „Cross-Suche“ eignet sich hingegen sowohl für Suchanfragen, die nur wenige Dokumente ergeben, als auch für solche, die zu viele Treffer bringen. Als Mittel zum Auffinden der wenigen Treffer eignete sich die „Cross-Suche“ vor allem in Systemen, die keine direkte Suche über alle Datenbanken zuließen. Mittlerweile kommt der „Cross-Suche“ allerdings eher Bedeutung in Bezug auf trefferreiche Anfragen zu; hier können gezielt Quellen ausgewählt werden, die besonders viele Treffer zum Thema enthalten, gleichzeitig aber schon als Quelle für das Thema relevant sind oder einen besonderen Blickwinkel eröffnen.

Überträgt man die Auswahl der „Top-Quellen“ bzw. die „Cross-Suche“ auf die Web-Suche, so zeigt sich bei den gängigen Suchmaschinen, dass der Ansatz, bei einer Recherche erst einmal die wichtigsten Quellen zu finden, von diesen negiert wird, denn die Suchmaschinen zeigen in ihren Trefferlisten in der Regel nur zwei Treffer pro Server, also pro Quelle, an. Zwar ist es möglich, durch Folgen eines Links unterhalb dieser Treffer die weiteren Ergebnisse auf dem gleichen Server zu sehen, die Server mit vielen Dokumenten zum Thema werden jedoch nicht bevorzugt gelistet oder besonders hervorgehoben. Wie viele Dokumente tatsächlich hinter den entsprechenden Links stehen, wird nicht mir angegeben. Für den Benutzer ist es also nicht ersichtlich, ob es sich tatsächlich um eine umfangreiche Quelle zum Thema handelt.

Daraus ergibt sich bei der Websuche ein Paradox, nämlich dass man bei dieser vermeintlich einfachen Suche schon im Voraus die wichtigsten Quellen kennen sollte. Und kennt man sie, so hat man doch nicht die Möglichkeit, die Suche auf diese zu beschränken. Daraus ergibt sich, dass man die Websuche zumindest zum Teil auch als Quellensuche betrachten sollte. Dabei können Informationen aus Verzeichnissen, aber auch andere aus dem Web extrahierte Informationen nützlich sein.

6 Auffinden von „Top-Quellen” in Suchmaschinen

Der Nutzen der Einschränkung der Suche auf „Top-Quellen“ konnte im vorangegangenen Abschnitt gezeigt werden. Nun soll untersucht werden, wie sich dieses Konzept mittels der Integration von Daten aus Web-Verzeichnissen in Suchmaschinen umsetzen lässt.

Bei der Suche in einer Suchmaschine mit eingebundenem Web-Verzeichnis können als Ergebnis direkt Verzeichnisklassen angezeigt werden. Dies kann auf Anfragen hin erfolgen, die entweder eine exakte Übereinstimmung mit der Klassenbezeichnung ergeben oder durch erweiterte Verfahren mit den Klassenbezeichnungen abgeglichen werden. Ein solches Ver-

fahren wird beispielsweise bei Yahoo eingesetzt, um auch nicht exakte Anfragen mit den Klassen abgleichen zu können (Wu 1999; vgl. auch Stock u. Stock 2000). Wichtig ist, dass bei solchen Treffern die weitere Auswahl von Top-Quellen meist nicht sinnvoll ist, da die Suche in den Quellen wiederum mit einem Teil der Klassenbezeichnung durchgeführt werden würde. So ist es zwar sinnvoll, bei einer Anfrage nach „Informationswissenschaft“ die entsprechende Klasse als Ergebnis anzuzeigen, eine Suche in den Sites dieser Klasse wäre aber nicht sinnvoll, da durch die Klassenbezeichnung ja schon klar ist, dass alle Quellen für den Begriff relevant sind.

Interessanter ist der Fall, wenn keine Übereinstimmungen zwischen Anfrage und Klassenbezeichnungen bestehen. Es wird im Folgenden von einer großen Treffermenge ausgegangen, die zumindest einige Quellen (Server) enthält, die jeweils viele zur Anfrage passende Dokumente enthalten. Diese würden in der regulären Trefferliste „geclustert“ werden, d.h. es würden nur zwei Dokumente pro Server angezeigt werden. Es sollen aber gerade die Quellen gefunden werden, die sowohl viele Dokumente enthalten als auch durch die Aufnahme in ein Verzeichnis eine gewisse Qualitätsprüfung durchlaufen haben. Abbildung 1 zeigt den Prozess der Quellenauswahl, der im Folgenden erläutert wird.

Nach der Überprüfung, ob es eine Übereinstimmung zwischen Anfrage und Verzeichnisklasse gibt, werden in einem ersten Schritt alle Server ermittelt, die entweder mindestens eine gewisse Anzahl von Dokumenten enthalten oder aber es werden die *n* Server mit den meisten Dokumenten ermittelt, wobei *n* einen Cut-Off-Wert darstellt, beispielsweise 20. Die ermittelte Menge der Server wird für die weitere Bearbeitung verwendet. Allerdings enthält diese Menge noch nicht allein die Top-Quellen, sondern schlicht alle Quellen, die viele Dokumente zum Thema enthalten. Zu diesen dürften in vielen Fällen auch für die Anfrage nicht relevante Quellen gehören; zum Beispiel solche, die versuchen, durch den Aufbau von komplexen Verlinkungsstrukturen in den Suchmaschinen ein besseres Ranking zu erhalten und deshalb eine hohe Anzahl von Dokumenten, die einen Suchbegriff enthalten, generieren. Auch muss vermieden werden, dass Quellen allein aufgrund ihres Umfangs als Top-Quellen angesehen werden.

Die so ausgewählten Quellen können nun mit einem oder mehreren Verzeichnissen abgeglichen werden. Es bietet sich an, sowohl ein allgemeines Verzeichnis (wie ODP) einzubinden als auch spezialisierte Verzeichnisse.

Als nächstes wird in jedem verwendeten Verzeichnis für jeden einzelnen Server geprüft, ob dieser enthalten ist. Die im Verzeichnis enthaltenen Server werden in der weiteren Auswertung berücksichtigt, die nicht im Verzeichnis enthaltenen Server werden ausgeschlossen. Durch die Qualitätskontrolle der Verzeichnisse (der menschlichen Redaktion) werden diejenigen Server ausgeschlossen, die die Qualitätsstandards des verwendeten Verzeichnisses nicht einhalten können. Allerdings werden auch alle Server ausgeschlossen, die im Verzeichnis nicht enthalten sind, etwa weil bisher kein Editor Zeit fand, diese mit aufzunehmen. Es ist allerdings davon auszugehen, dass die Verzeichnisklassen die wichtigsten Quellen zum Thema enthalten (vgl. auch Hamdorf 2004 zur Vorgehensweise beim Aufbau von Verzeichnissen). Des Weiteren wird eine Liste der gefundenen Kategorien erstellt, die auch die darin enthaltene Anzahl der überprüften Server enthält.

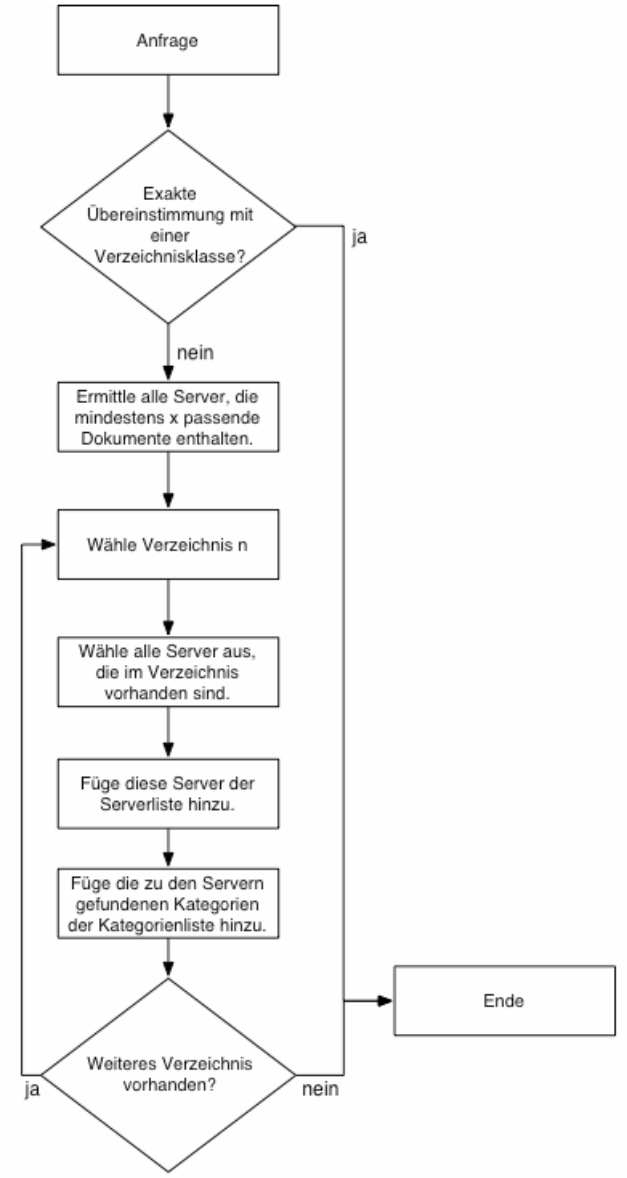

Abbildung 1: Automatische Auswahl der Verzeichnisquellen

Die neu ermittelte Servermenge erfüllt nun zwei Bedingungen: Erstens enthält sie nur Quellen, die eine gewisse Anzahl zur Suchanfrage passender Dokumente enthalten. Zweitens enthält sie nur Quellen, die in einer menschlichen Qualitätskontrolle für gut befunden wurden.

Die ermittelte Kategorienliste enthält die relevanten Kategorien aus den ausgewerteten Verzeichnissen mit der Anzahl der dort enthaltenen Server, auf denen Dokumente gefunden wurden sowie die Anzahl der insgesamt in der jeweiligen Kategorie enthaltenen Quellen.

Als letzter Schritt bleiben nun noch die Art und der Umfang der Umsetzung der Verzeichnisquellen in ein Suchergebnis. Dabei stehen vier Möglichkeiten zur Verfügung:

1. Die Auswahl der Server wird beibehalten. Alle ermittelten Server werden unabhängig von ihrer Stellung im Verzeichnis für die Suche ausgewählt.
2. Die Klasse oder diejenigen Klassen, die am meisten relevante Server enthalten, werden ausgewählt. *Alle* Server der Klasse werden in der weiteren Suche berücksichtigt, unab-

hängig davon, ob sie in der ursprünglichen Treffermenge enthalten waren. Da der Umfang der Klassen sehr stark variiert, kann auch innerhalb der Klassen mit einem Cut-Off-Wert gearbeitet werden. Wie schon bei Google üblich, kann die Liste der Quellen nach einem statischen Wert ihrer Linkpopularität geordnet werden. Aufgrund dieser Ordnung kann in Kombination mit dem Cut-Off-Wert die Suche nur auf die populärsten Quellen beschränkt werden.

3. Die Auswahl der zu berücksichtigenden Klassen wird dem Nutzer überlassen. Ihm werden die Klassenbezeichnungen mit der Anzahl der relevanten Quellen angeboten.
4. Der Nutzer wählt die zu berücksichtigenden Server selbständig aus einer Liste aus.
5. Auf welche Art auch immer dieser Schritt ausgeführt wird, wird letztlich noch die modifizierte Suchanfrage wieder an den Suchmaschinen-Index gesendet. Die Anfrage wird dabei auf die ausgewählten Server beschränkt, so dass nur Treffer von diesen zurückgegeben werden. Dabei sollten die sonst im Ranking verwendeten statischen Werte für die Linkpopularität nicht bzw. nur eingeschränkt verwendet werden, da sie häufig grundsätzlich Dokumente aus einer Quelle gegenüber denen aus einer anderen Quelle bevorzugen (vgl Lewandowski 2005).

Das vorgeschlagene Verfahren soll anhand eines Beispiels verdeutlicht werden: Ein Nutzer sucht nach Informationen über den Lotuseffekt. Eine Suche in Google erbringt über 30.000 Treffer. Tabelle 1 zeigt alle Server aus den Top-500-Treffern bei Google, die einen Link auf weitere Dokumente, die auf demselben Server liegen, enthalten.[8]

Tabelle 1: Gefundene Server für die Suchanfrage „Lotuseffekt“, geordnet nach der Anzahl der gefundenen Treffer je Site (Die einleitenden Kategorien „World: Deutsch“ werden aus Gründen der Übersichtlichkeit beim ODP nicht mit aufgeführt.)

Server	**Beschreibung**	**Anzahl Treffer in Google**	**Kategorie in ODP**	**Kategorie in Yahoo**
www.botanik.uni-bonn.de	Botanisches Institut der Universität Bonn	224	-	Nordrhein-Westfalen > Universität Bonn > Botanisches Institut und Botanischer Garten
www.uni-protokolle.de	Prüfungsprotokolle, Hochschulnachrichten	181	Wissenschaft: Studium	Uni/FH > Hausarbeiten, Skripte und Klausuren
search.ebay.de	Auktionshaus	174	-	-

8. Die Tabelle enthält alle Sites, die mehr als zwei Treffer ergaben. Bei dem Beispiel wurden sowohl in Google als auch in den Verzeichnissen nur deutschsprachige Treffer ausgewertet. Wurden bei einer Site auffällig viele Treffer gefunden (in Beispielen waren dies in der Regel weit über 50.000), so wurde anhand der Trefferliste überprüft, wie weit diese tatsächlich ging. Dieser Wert wurde entsprechend in der Tabelle angegeben. Es handelt sich dabei um ein spezielles Problem von Google: Es werden alle Dokumente einer Site gezählt, auch wenn der Begriff beispielsweise auf jeder Einzelseite in der Navigation vorkommt. Solche Treffer werden erst bei der Anzeige der Trefferliste als Dubletten angesehen und entsprechenden nicht mit angezeigt, können jedoch über einen Link am Ende der Trefferliste aufgerufen werden.

www.baustoffchemie.de	Informationsportal zur Chemie der Werkstoffe im Bauwesen	170	Wissenschaft: Technologie: Bauingenieurwesen	-
cgi.ebay.de	Auktionshaus	112	-	-
www.werkzeug.dea1.de	Werkzeug-Shop-Führer	89	-	-
wohnen.listings.ebay.at	Auktionshaus	89	-	-
www.baulinks.de	Bauportal	88	-	Firmen > Bauwesen > Brancheninformation
www.bauzentrale.com	Newsportal Bau	55	-	-
www.3sat.de	3sat Fernsehen	41	Medien: Fernsehen: Sender: Öffentlich-rechtliche	Fernsehen > 3sat
idw-online.de	Nachrichten aus der Wissenschaft	37	Wissenschaft: Nachschlagewerke	Forschung und Wissenschaften > Zeitschriften und Online-Magazine
www.neuematerialien.de	Marktplatz der Werkstofftechnik	31	Wissenschaft: Technologie: Werkstoffe	-
www.thesu.de	Themensuchmaschine	31	-	-
www.dbu.de	Deutsche Bundesstiftung Umwelt	29	Regional: Europa: Deutschland: Gesellschaft: Umweltschutz	Umwelt und Natur > Organisationen
www.moerike-g.es.bw.schule.de	Mörike-Gymnasium Esslingen	24	-	-
www.innovations-report.de	Informationsplattform zur Förderung der Innovationsdynamik	24	Wissenschaft: Technologie: Zeitschriften und Online-Magazine Wissenschaft: Technologie: Erfindungen und Innovationen	Forschung > Portale und Linksammlungen
www.zeiss.de	Unternehmen	12	Wirtschaft: Industriegüter und -dienstleistungen: Optik	Firmen > B2B > Optik > Carl Zeiss Gruppe

www.colour-europe.de	Fachzeitschrift „Phänomen Farbe“	12	Wirtschaft: Chemie: Beschichtungs- und Klebstoffe: Farben und Lacke: Zeitschriften und Online-Magazine	-
www.maschinen-markt.de	Fachzeitschrift	10	Wirtschaft: Industriegüter und -dienstleistungen: Maschinen und Werkzeuge: Zeitschriften und Online-Magazine	
www. dasumwelthaus.de	Informationen rund um zeitgemäßes Wohnen	6	-	
www.fassatec.de	Unternehmen aus dem Bereich Fassadentechnik	5	-	

Die in der Tabelle gezeigten Ergebnisse sind das Resultat des in Abbildung 1 dargestellten Verfahrens. Nun stellt sich die Frage, welche Ergebnismenge aus diesem Ergebnis gezogen werden soll. Gemäß den oben aufgeführten Möglichkeiten der Umsetzung wären dies:

Die Auswahl der Server wird beibehalten, alle auf diesen Servern gefundenen Dokumente bilden die Ergebnismenge. Es erfolgt ein neues Ranking, welches die Ergebnisse aller dieser Server mischt. Im Beispiel würde es sich anbieten, alle Server einzubeziehen, die von mindestens einem der Verzeichnisse gefunden werden.[9] Hier zeigt sich auch die Schwäche der Verzeichnisse: Offensichtlich sind auch manche hoch relevante Server nicht in beiden Verzeichnissen vorhanden. Allerdings gibt es auch keine relevanten Quellen, die in beiden Verzeichnissen fehlen. Irrelevante Sites wie die verschiedenen Ebay-Server mit Auktionsangeboten werden erfolgreich ausgeschlossen. Die gefundenen Server verteilen sich auf relativ viele unterschiedliche Klassen. Eine Einschränkung auf nur eine Klasse erscheint daher nicht sinnvoll; die Ausweitung auf alle Server einer Klasse damit auch nicht. Weitere Beispielanfragen müssen zeigen, ob eine solche Form der Einschränkung in anderen Fällen sinnvoll ist. Möglich wäre allerdings die Beschränkung der Recherche auf eine der obersten Hierarchieebenen. Bei ODP zeigt sich eine klare Unterteilung der Treffer in die Klassen „Wirtschaft“ und „Wissenschaft“. Dem Nutzer könnte die Wahl gegeben werden, seine Suche auf einen der Bereiche einzuschränken. Bei Yahoo ergibt sich diese Möglichkeit aufgrund der Verzeichnisstruktur nicht.

6. Die Auswahl der relevanten Klassen und die weitere Recherche in diesen durch den Nutzer ließe sich für das Beispiel realisieren, auch wenn der Vorteil für die Recherche hier nicht sicher erscheint.

9. Die Einbeziehung der Daten der beiden großen Webverzeichnisse dürfte in der Praxis nicht gelingen, da die ODP-Daten zwar kostenfrei zur Verfügung stehen, das Yahoo-Verzeichnis jedoch proprietär ist.

7. Eine Auswahl der relevanten Server durch den Nutzer ist in jedem Fall sinnvoll. Die in den Verzeichnissen aufgeführten Server könnten nach der Anzahl der Dokumente oder je nachdem, von wie vielen Verzeichnissen sie gefunden wurden, gelistet werden.

Ein ähnliches Ergebnis zeigt sich bei einem zweiten Beispiel (s. Tabelle 2), der Suchanfrage „WLAN“. Allerdings zeigt sich hier bei ODP eine Verzeichnisklasse („Computer und Technik > Zeitschriften und Online-Magazine“), in der drei hoch relevante Server enthalten sind. Hier könnte es sinnvoll sein, die Recherche auf alle in dieser Klasse enthaltenen Server auszuweiten.

Tabelle 2. Gefundene Server für die Suchanfrage „WLAN“, geordnet nach der Anzahl der gefundenen Treffer je Site

Server	**Beschreibung**	**Anzahl Treffer in Google**	**Kategorie in ODP**	**Kategorie in Yahoo**
www.botanik.uni-bonn.de	Botanisches Institut der Universität Bonn	224	-	Nordrhein-Westfalen > Universität Bonn > Botanisches Institut und Botanischer Garten
www.uni-protokolle.de	Prüfungsprotokolle, Hochschulnachrichten	181	Wissenschaft: Studium	Uni/FH > Hausarbeiten, Skripte und Klausuren
search.ebay.de	Auktionshaus	174	-	-
www.baustoffchemie.de	Informationsportal zur Chemie der Werkstoffe im Bauwesen	170	Wissenschaft: Technologie: Bauingenieurwesen	-
cgi.ebay.de	Auktionshaus	112	-	-
www.werkzeug.dea1.de	Werkzeug-Shop-Führer	89	-	-
wohnen.listings.ebay.at	Auktionshaus	89	-	-
www.baulinks.de	Bauportal	88	-	Firmen > Bauwesen > Brancheninformation
www.bauzentrale.com	Newsportal Bau	55	-	-
www.3sat.de	3sat Fernsehen	41	Medien: Fernsehen: Sender: Öffentlich-rechtliche	Fernsehen > 3sat
idw-online.de	Nachrichten aus der Wissenschaft	37	Wissenschaft: Nachschlagewerke	Forschung und Wissenschaften > Zeitschriften und Online-Magazine

www.neuematerialien.de	Marktplatz der Werkstofftechnik	31	Wissenschaft: Technologie: Werkstoffe	-
www.thesu.de	Themensuch-maschine	31	-	-
www.dbu.de	Deutsche Bundess-tiftung Umwelt	29	Regional: Europa: Deutschland: Gesellschaft: Umweltschutz	Umwelt und Natur > Organisationen
www.moerike-g.es.bw.schule.de	Mörike-Gymna-sium Esslingen	24	-	-
www.innovations-report.de	Informationsplatt-form zur Förderung der Innovationsdy-namik	24	Wissenschaft: Technologie: Zeit-schriften und Online-Magazine Wissenschaft: Technologie: Erfindungen und Innovationen	Forschung > Por-tale und Linksam-mlungen
www.zeiss.de	Unternehmen	12	Wirtschaft: Indus-triegüter und -dienstleistungen: Optik	Firmen > B2B > Optik > Carl Zeiss Gruppe
www.colour-europe.de	Fachzeitschrift „Phänomen Farbe“	12	Wirtschaft: Che-mie: Beschich-tungs- und Klebstoffe: Farben und Lacke: Zeitschriften und Online-Magazine	-
www.maschinen-markt.de	Fachzeitschrift	10	Wirtschaft: Indus-triegüter und -dienstleistungen: Maschinen und Werkzeuge: Zeitschriften und Online-Magazine	
www.dasumwelthaus.de	Informationen rund um zeitgemäßes Wohnen	6	-	
www.fassatec.de	Unternehmen aus dem Bereich Fassa-dentechnik	5	-	

7 Fazit

Zwar haben Web-Verzeichnisse im Lauf der Jahre an Popularität verloren, dies mag allerdings auch an der mangelhaften Integration ihrer Daten in die algorithmischen Suchmaschinen liegen. Es wurde ein Ansatz vorgestellt, wie sich das aus der „Datenbank-Welt“ bekannte Konzept der „Top-Quellen“ bzw. der „Cross-Suche“ auf das Web anwenden lässt.

Das vorgestellte Verfahren erscheint vielversprechend, es bedarf jedoch vor allem noch einer empirischen Überprüfung und ausführlicher Tests mit echten Nutzern und ihren Suchanfragen. Dies konnte im Rahmen der vorliegenden konzeptionellen Arbeit noch nicht geleistet werden. Es konnten aber durchaus anhand der beschriebenen Beispiele mögliche Anwendung des Verfahrens gezeigt werden. Es wäre wünschenswert, wenn sich die Forschung (wieder) mit Fragen der Integration von Verzeichnisdaten in Suchmaschinen beschäftigen würde. Dass für das Suchergebnis die Qualität der zugrunde liegenden Quellen von großer Bedeutung ist, ist unstrittig. Mit den Verzeichnisdaten liegt ein Instrument vor, die Qualität der Suchergebnisse zu erhöhen.

Bisher nicht behandelt wurde die Navigation innerhalb des Verzeichnisses auf Basis der gefundenen Verzeichnistreffer. Durch ein solches den Nutzer leitendes Verfahren könnte die Qualität der Suchergebnisse in einem weiteren Suchschritt weiter erhöht werden.

Von besonderer Bedeutung für das vorgestellte Verfahren ist die Qualität der zugrunde liegenden Verzeichnisse. Schon in den vorgestellten Beispielen wurde etwa deutlich, dass sich die Treffer aufgrund der inkonsistenten Klassierung teils nur eingeschränkt verwenden lassen. Insbesondere die Integration von spezialisierten Verzeichnissen erscheint vielversprechend: Für jede Abfrage müssten dann allerdings entsprechend viele Einzelverzeichnisse durchsucht werden.

Literatur

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

Chung, Y. M.; *Noh*, Y. (2003)
Developing a specialized directory system by automatically classifying Web documents. Journal of Information Science 29(2), 117-126

Dennis, S.; *Bruza*, P.; *McArthur*, R. (2002)
Web Searching: A Process-Oriented Experimental Study of Three Interactive Search Paradigms. Journal of the American Society for Information Science and Technology 53(2), 120-133

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

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

Hamdorf, K. (2004):
Jenseits von Google – Erschließung und Recherche von Internet-Angeboten durch Webkataloge. IWP Information Wissenschaft und Praxis 55(4), 221-224

How much Information? (2003): Internet. http://www.sims.berkeley.edu/research/projects/how-much-info-2003/internet.htm [29.3.2004]

Kleinberg, J. (1999)
Authoritative Sources in a Hyperlinked Environment. Journal of the ACM 46(5), 604-632
Lewandowski, D. (2004)
Technologie-Trends im Bereich der WWW-Suchmaschinen. In: Information Professional 2011: 26. Online-Tagung der DGI; Frankfurt am Main 15. bis 17. Juni 2004; Proceedings, 183-195
Lewandowski, D. (2005):
Bewertung von linktopologischen Verfahren als bestimmender Ranking-Faktor bei WWW-Suchmaschinen. In: Wissensorganisation und gesellschaftliche Verantwortung. 9. Tagung der Deutschen ISKO (Wissensorganisation'2004), Proceedings [i.Dr.]. http://www.durchdenken.de/lewandowski/doc/isko2004.pdf [14.12.2004]
Mandl, T. (2002)
Evaluierung von Internet-Verzeichnisdiensten mit Methoden des Web-Mining. In: Hammwöhner, R.; Wolff, C.; Womser-Hacker, C. (Hrsg.) (2002): Information und Mobilität: Optimierung und Vermeidung von Mobilität durch Information. Proceedings des 8. Internationalen Symposiums für Informationswissenschaft, 7.-10.10.2002. Konstanz: Universitätsverlag (Schriften zur Informationswissenschaft; 40)
Mandl, T. (2003)
Projekt Automatische Qualitätsabschätzung von Internet Ressourcen (AQUAINT). Arbeitsbericht 3/2003, Universität Hildesheim, Informationswissenschaft. http://www.uni-hildesheim.de/~mandl/Publikationen/Ab_aquaint02.pdf
Mandl, T. (2005)
Qualität als neue Dimension im Information Retrieval: Das AQUAINT-Projekt. Information: Wissenschaft und Praxis 56(1), 13-20
O'Leary, M. (1998)
Web directories demonstrate an enduring online law. Online 22(4), 79-81
Page, L., *Brin*, S., *Motwani*, R., *Winograd*, T. (1998)
The PageRank citation ranking: Bringing order to the Web. http://dbpubs.stanford.edu:8090/pub/1999-66 [26.10.2004]
Sherman, C. (2000)
Humans Do It Better: Inside the Open Directory Project. Online 24(4). http://www.onlinemag.net/OL2000/sherman7.html [30.3.2005]
Stock, M., *Stock*, W. G. (2000)
Klassifikation und terminologische Kontrolle: Yahoo!, Open Directory und Oingo im Vergleich. Password 14(12), 26-33
Sullivan. D. (2005)
Yahoo Directory Makes Changes & Further Directory Decline. http://blog.searchenginewatch.com/blog/050308-101342 [29.3.2005]
Wätjen, H. (1999)
GERHARD – Automatisches Sammeln, Klassifizieren und Indexieren von wissenschaftlich relevanten Informationsressourcen im deutschen World Wide Web. B.I.T. online. 1(4), 279-290
Wu, J. (1999)
Information retrieval from hierarchical compound documents / Yahoo Inc. Patent Nr. US 5,991,756 vom 23.11.1999