In: Hammwöhner, Rainer; Wolff, Christian; Womser-Hacker, Christa (Hg.): Information und Mobilität, Optimierung und Vermeidung von Mobilität durch Information. Proceedings des 8. Internationalen Symposiums für Informationswissenschaft (ISI 2002), Regensburg, 8. – 11. Oktober 2002. Konstanz: UVK Verlagsgesellschaft mbH, 2002. S. 239 – 257

Evaluierung von Internet-Verzeichnisdiensten mit Methoden des Web-Mining

Thomas Mandl

Informationswissenschaft
Universität Hildesheim
Marienburger Platz 22
31141 Hildesheim
mandl@uni-hildesheim.de

Zusammenfassung

Dieser Artikel gibt einen Überblick über die Möglichkeiten und Chancen der Bewertung der Popularität von Internet-Angeboten durch die Analyse der Verlinkungsstruktur. Insbesondere die auf eine Seite verweisenden Hypertext-Verbindungen werden immer häufiger als wichtiger Maßstab für die Autorität und sogar die Qualität von Seiten herangezogen. Eine Methode für die Adaption dieses seiten-orientierten Maßes für komplexere Internet-Angebote wird vorgestellt. Als Untersuchungsbeispiele wurden exemplarisch deutschsprachige Web-Verzeichnisdienste gewählt. Dabei konnte u.a. gezeigt werden, dass die Wahrscheinlichkeit für Hypertext-Verbindungen auf Verzeichnisdienste mit zunehmender hierarchischer Tiefe der Seiten extrem stark abnimmt und dass die Anzahl der Links auf eine Seite eines Verzeichnisdienstes nicht mit der Anzahl der Links auf die dort genannten Seiten konform geht.

Abstract

This paper provides an overview on the possibilities of the evaluation of the popularity of web sites by analyzing link structure. Especially back links are more and more considered as an important indicator for the authority and even the quality of pages. The most important algorithms for these tasks are presented. Furthermore, this page oriented method is extended for complex sites. The possibilities of web structure mining are explored by analyzing German web catalogues. It could be shown that the probability for a link to point to a page in a catalogue decreases drastically when the page is low in the hierarchy and that the number of links to a page in an internet catalogue does not correlate with the number of back links of the sites mentioned.

1 Einleitung

Die Popularität der Suchmaschine Google und des darin verwendeten Page-Rank Algorithmus, der die Verlinkung der Hypertext-Struktur des Internets als Ergänzung für das Ranking der Ergebnis-Dokumente heranzieht, hat zu einem breiten Interesse an der Bewertung von Internet-Angeboten durch Link-Analysen geführt. Die Anzahl von auf eine Internet-Seite verweisenden Hypertext-Verbindungen (*back links* oder *in links*) gilt als Maß für die Autorität und Popularität dieser Seite. Die unkritische Verwendung der einfachen Maßzahlen wirft jedoch zahlreiche Fragen auf. Dieser Beitrag will einige dieser Probleme im Kontext einer Evaluierung von Web-Sites vorstellen.

Da die Link-Analysen meist im Rahmen von Schlagwort-Suchen genutzt werden, wurden hier bewusst Angebote gewählt, die das Browsing als primäre Suchstrategie nutzen. Das Verfolgen von Hyperlinks ist eine sehr wichtige Strategie für die Orientierung (MARCHIONINI 1995), die häufig unterschätzt wird. Eine Untersuchung von MAHOUI & CUNNINGHAM 2001 zeigt, dass in der untersuchten digitalen Bibliothek lediglich ca. 50% der Benutzer eine Suchanfrage stellen. Vermutlich wünschen viele der übrigen Besucher die Möglichkeit, ihr Informationsbedürfnis durch effiziente Browsing-Angebote zu lösen.

Browsing im Internet basiert auf Verknüpfungen, die entweder vage Ähnlichkeitsbeziehungen wiederspiegeln wie etwa in zweidimensionalen Karten (cf. EIBL & MANDL 2002) oder fest definierte Beziehungen zeigen, die wie etwa in Verzeichnisdiensten (*web directories, internet catalogues*). Die Bedeutung dieser Dienste und ihrer Bewertung steigt, da sie inzwischen auch automatisiert erstellt werden (cf. z.B. FROMMHOLZ 2001). Dazu werden Verfahren der Text-Kategorisierung (MANDL 2001) auf hierarchische Strukturen übertragen.

Die Verzeichnisdienste zeichnen sich vor allem durch ihre hierarchische Struktur aus. Ansätze zu ihrer Bewertung lassen sich demnach auch auf zahlreiche andere hierarchische Web-Angebote übertragen. Hier haben sich automatische Verfahren für die Evaluierung noch wenig etabliert. RITTBERGER 2001 stellt etwa eine umfassende Methodik vor, welche die intellektuelle Bewertung ermöglicht.

2 Web-Mining

Web-Mining nennt sich eine Forschungsrichtung, die auf die Gewinnung von Wissen aus den großen Datenmengen im Internet abzielt. Dabei geht es nicht

um das Retrieval einzelner Seiten mit einem geringen Funktionsumfang, sondern um die Zusammenfassung sehr vieler Daten zu neuem Wissen, das so explizit nicht gespeichert ist. Der Begriff entstand als Analogie zum Data-Mining.

Data-Mining war eine Konsequenz der großen Datenmengen, die mit fallenden Preisen für Massenspeicher in immer mehr Anwendungsbereichen vorlagen: "Now that we have gathered so much data, what do we do with it?" (FAYYAD & UTHURUSAMY 1996:24). Die Definition eines Algorithmus erinnert noch stark an das maschinelle Lernen: "... Any algorithm that enumerates patterns from, or fits models to, data is a *data mining algorithm*" (FAYYAD 1997:5). Allerdings soll Data-Mining den gesamten Prozess der Sammlung und Pflege der Daten angefangen, über ihre Auswertung bis hin zu der Anwendung der Ergebnisse umfassen.

Web-Mining überträgt Data-Mining auf Anwendungen im Internet und passt insbesondere die Systeme zum Sammeln der Daten auf die Gegebenheiten im Web an. In diesem Kontext wurde bereits der Begriff *Web-Intelligence* geschaffen, der auch schon Titel für eine wissenschaftliche Konferenz war[1]. Folgende Teilaspekte von Web-Mining sind nach WALTHER 2001 relevant:

- Web-Content-Mining
- Web-Structure-Mining
- Web-Usage-Mining

Diese Teilaspekte werden im Folgenden kurz erläutert.

2.1 Web-Content-Mining

Content-Mining versucht im weitesten Sinne Inhalte von Web-Seiten auszunutzen und durch das Zusammenfassen von großen Mengen von Inhalten zu neuem Wissen zu gelangen.

Dazu zählt etwa das Extrahieren von mehrsprachigen Parallel-Korpora aus dem Web für die maschinelle Übersetzung oder multilinguales Information Retrieval. Dabei werden anhand heuristischer Regeln, die z.B. auf Web-Servern nach Verzeichnissen mit gleichnamigen Dateien suchen, deren Verzeichnisnamen auf Sprachversionen hinweisen (cf. z.B. NIE ET AL. 2001). Die von Menschen geleistete Übersetzungsarbeit für diese konkreten Texte wird dann in einem Transfer-Prozess ausgenutzt, um durch die Parallelisierung von

[1] http://kis.maebashi-it.ac.jp/wi01.

Texten in verschiedenen Sprachen eine Übersetzung vorzunehmen, die auf den Kontexten basiert, in denen ein Wort vorkommt (cf. MANDL 2001).

Andere Ansätze versuchen anhand von Trends in Publikationen auf sinnvolle Forschungsrichtungen zu schließen (PIROLLI ET AL. 1996).

2.2 Web-Structure Mining

Das Ziel des Structure-Mining besteht darin, Wissen aus der Struktur von Hypertexten zu gewinnen. Struktur kann sich dabei auf die HTML Struktur einzelner Dokumente beziehen, auf die Struktur von Sites und auf die Verlinkungsstruktur des gesamten Internets oder von Ausschnitten des Internet wie etwa den Resultaten von Suchanfragen. Meist sollen die Struktur-Informationen die Suche im Web verbessern (cf. z.B. WEISS ET AL. 1996, BEKAVAC 1999, KRUSCHWITZ 2001). Diese Verfahren beschreibt das nächste Kapitel ausführlich.

2.3 Web-Usage-Mining

Beim Web-Usage-Mining geht es um die Interpretation von sehr vielen Benutzeraktionen im Internet. Daraus lassen sich Folgerungen über die Angemessenheit und Gebrauchstauglichkeit von Angeboten ziehen.

Als Quelle dienen meist die Log-Dateien einzelner Server oder großer Proxy-Server. Ebenso können lokale Bookmark-Dateien, lokale History-Dateien von Web-Clients oder spezifische Client-Log-Dateien genutzt werden.

Die Log-File-Analyse basiert i.d.R. auf dem Server-Log eines Anbieters und kann für die Analyse und Optimierung der Seiten dieses Anbieters eingesetzt werden (cf. z.B. BARTEL 2002). Ein Vergleich mit anderen Sites für ähnliche Anwendungen (wie etwa der Vergleich zweier Online-Geschäfte für Bücher) ist so nicht möglich. Dazu wäre eine globale Perspektive erforderlich, aus der sich die Interaktionen vieler Benutzer mit vielen Internet-Angeboten ersehen lässt. Dies würde am ehesten durch die Proxy-Server großer Internet Service Provider (ISP) erreicht, die allerdings dem Datenschutz unterliegen.

Die Analyse von Nutzungsdaten aus dem Internet gilt auch als eine große Chance für die empirische Forschung zur Mensch-Maschine-Interaktion. Die Evaluierung von Benutzungsschnittstellen ist extrem wichtig, da in einer derart komplexen und noch nicht vollständig erforschten Domäne wie der Mensch-Maschine-Interaktion rein analytische Aussagen zur Qualität einer

Software hinsichtlich der Interaktion kaum getroffen werden können. Um aussagekräftige Ergebnisse zu gewinnen, müssen Benutzertest durchgeführt werden. Dazu gehört die Beobachtung und Befragung einiger repräsentativer Testbenutzer während der Erledigung von Testaufgaben. Alle Interaktionsschritte sollten aufgezeichnet und der Benutzer gefilmt werden. Solche Untersuchungen erfordern den Einsatz von vielen Ressourcen und sind sehr teuer.

Dagegen entstehen die Daten in Log-Files durch tatsächliche reale Aktionen von Internetnutzern. Verfahren für ihre Auswertung gelten daher als „'discount' usability assessment methods" (SULLIVAN 1997:2). Allerdings enthalten die Log-Files nur sehr wenig Information und sind daher keineswegs so aussagekräftig wie Benutzertests. So zeichnet der Server lediglich den Abruf einer Datei auf, die Interaktion mit Web-Seiten geschieht lediglich auf Seiten des Clients, so dass darüber keine Informationen vorliegen. Darüber hinaus sind die Daten der Log-Dateien mit großen Unsicherheiten behaftet. Weder lässt sich ein Benutzer eindeutig identifizieren, noch erreichen alle Seitenaufrufe den Server, da Dateien teils aus dem lokalen Cache oder von Proxy-Servern geladen werden.

3 Analysen der Verlinkungsstruktur

Im Web-Structure-Mining dominieren Analysen der Hypertext-Verknüpfungen (für einen Überblick cf. BAEZA-YATES & RIBEIRO-NETO 1999:380f.). Die Grundideen dieser Analysen der Verlinkungsstruktur stammen aus der Biblio- oder Szionometrie, die das Netzwerk der wissenschaftlichen Zitate analysieren und darauf abzielen, Publikationen aufgrund der Häufigkeit der auf sie verweisenden Zitate zu bewerten. Komplexere Maße betrachten z.B. die Stellung eines Autors im Diskurs-Netzwerk und berücksichtigten über die Zitate hinaus die institutionelle Zugehörigkeit und Ko-Autorenschaft (cf. MUTSCHKE 2001). Andere komplexere Analysen errechnen aus den Häufigkeiten von Zitaten, Maßzahlen für das Renommee von Zeitschriften, Tagungen oder Fachbereichen (cf. z.B. SCHLÖGL 2000).

Die technischen Möglichkeiten der online Verfügbarkeit von wissenschaftlicher Literatur führt dazu, dass bibliometrische Analysen heute Teil von kostenlos zugänglichen digitalen Bibliotheken sind[2].

[2] Entsprechende Kennzahlen sind z.B. in Daffodil (http://www.daffodil.de) und Research-Index (http://citeseer.nj.nec.com/cs) integriert.

3.1 Autoritätsmaße

Die Suchmaschine Google kann als umfangreichste und erfolgreichste Implementierung einer automatischen Analyse der Verlinkungsstruktur gelten. Der in Google wirkende PageRank Algorithmus (PAGE ET AL. 1998) zählt nicht nur die Links auf eine Seite. Zunächst erhalten alle Seiten das gleiche Gewicht als Verteiler. Das bedeutet, dass die Autorität, die eine verweisende Seite vergeben kann, an der Anzahl der ausgehenden Links relativiert wird. Darüber hinaus wird der Einfluss einer Seite auch mit deren Autorität relativiert. Je größer die Autorität einer Seite ist, desto höheres Gewicht haben die von ihr ausgehenden Links.

Der Algorithmus arbeitet iterativ. Zunächst werden alle Seiten mit dem gleichen Autoritätswert initialisiert und dann berechnet der erste Schritt die neue Autorität aller Seiten aus der Verlinkung. Dabei ergeben sich neue Autoritätswerte, so dass alle Werte nun erneut berechnet werden und das Ergebnis die angestrebte Autorität besser wiedergibt. Laut den Autoren konvergiert der Algorithmus nach einer Anzahl von Schritten (cf. PAGE ET AL. 1998), d.h. bei einem weiteren Berechnungsschritt verändern sich die Autoritätswerte kaum mehr. Der Autoritätswert wird für das Berechnen des Rankings der Dokumente nach einer Anfrage benutzt (cf. PAGE ET AL. 1998).

Der PageRank-Algorithmus benutzt sinnvolle Annahmen, um die Autorität zu berechnen. Allerdings beschränkt er sich auf die Ebene der Seite und berechnet z.B. per se nicht die Autorität einer gesamten Site.

Laut PAGE ET AL. 1998 gibt der PageRank-Algorithmus die Wahrscheinlichkeit wieder, mit der ein Surfer auf eine Seite trifft, wenn er Hypertext-Verbindungen verfolgt und nie auf eine bereits besuchte Seite zurückführt. In der Suchmaschine Google wird der PageRank-Wert einer Seite mit der System-Relevanz kombiniert. Die Wahrscheinlichkeit des Treffens auf eine Seite beim Browsen wird also auf die Suche übertragen.

3.2 Autorität bei der Vermittlung

Der PageRank-Algorithmus kann als eine einfachere Version des Kleinberg-Algorithmus betrachtet werden, die aber für wesentlich größere Mengen von Seiten geeignet ist. Dieser zielt ebenfalls auf Autorität ab und berücksichtigt nur die Verbindungsstruktur zwischen einer Menge von Seiten. Kleinberg führt zwei Rollen ein, um die Autorität zu bewerten (KLEINBERG 1998). Er spricht von hubs und authorities und weist jeder Web-Seite ein Gewicht für

beide Rollen zu. Der hub entspricht einem Mittelpunkt, Verteiler oder Informationsvermittler, dessen Aufgabe im Wesentlichen in der Bereitstellung von Verbindungen zu anderen Seiten besteht. Dahinter steht die Vorstellung eines clearinghouses oder in der Wissenschaft der eines guten Überblicksartikels. Die authorities dagegen enthalten die eigentliche Information in unterschiedlicher Qualität.

Im Gegensatz zum PageRank-Algorithmus findet das Verfahren von Kleinberg nur Anwendung auf eine Menge von ca. 5.000 bis 10.000 Seiten, die aus einer Suchanfrage ermittelt werden. Die besten Suchergebnisse eines Suchdienstes werden analysiert und die enthaltenen Verbindungen extrahiert. Die entsprechenden Seiten gelangen bis zu einer bestimmten Tiefe in den Datenbestand. Die Verbindungen innerhalb dieser Menge werden nun iterativ analysiert. Jede Seite besitzt sowohl ein Gewicht als hub als auch als authority, die in jedem Durchlauf modifiziert werden. Die Autorität einer Seite steigt mit der Anzahl der ankommenden Verbindungen. Diese Zahl wird aber an dem hub-Gewicht der Ausgangsseite relativiert. Nur die Links von guten Verteilerseiten wirken sich somit stark auf die Autorität einer Seite und damit auf das authority-Gewicht aus. Ebenso unterliegt das hub-Gewicht einer Veränderung, die von der Autorität der Zielseiten abhängt. Auf je bessere Seiten der Verteiler verweist desto besser ist er und desto stärker steigt sein hub-Gewicht. Ziel ist die Identifikation der Seiten mit der höchsten Autorität innerhalb der Untermenge.

Der Algorithmus birgt die Gefahr der weiten thematischen Entfernung durch die Integration weiterer Seiten neben dem eigentlichen Suchergebnis. Diese können von dem Thema, das mit der Suchanfrage verbunden ist, schon weit entfernt liegen (*topic-drift*).

Weitere Algorithmen wurden von GIBSON ET AL. 1998 und LEMPEL & MORAN 2000 vorgeschlagen. MEGHABGHAB 2002 untersucht die Algorithmen für Linkanalysen aus dem Blickwinkel der Grafentheorie und Matrizenalgebra und leitet einige formale Eigenschaften der Berechnungsmethoden ab.

3.3 Evaluierung

Inwieweit die Autoritätsmaße wie PageRank für die Benutzer Vorteile bringen, müssen Evaluierungen zeigen. Für das Information Retrieval gibt es Hinweise, dass sich RageRank bei großen Kollektionen positiv auf die Retrieval-Qualität auswirkt.

Im Rahmen der Evaluierungsstudie TREC (Text Retrieval Conference, cf. VOORHEES & HARMAN 2001) wird die Leistung von Information Retrieval Systemen hinsichtlich der Fähigkeit gemessen, thematisch relevante Dokumente zu identifizieren. Die Bedeutung von TREC liegt in dem großen Umfang der Daten und der erreichten Vergleichbarkeit zwischen Systemen. TREC stellt sich in seiner zehnjährigen Geschichte auch zahlreicher Kritik und neuen Anforderungen. Kritik richtete sich u.a. gegen die Methodik, welche den Benutzer vernachlässigt. Als eine Konsequenz wurde ein *interactive track* eingeführt, welcher die Unterstützung des Benutzers durch die Benutzungsoberfläche untersucht.

Die neuen Anforderungen ergaben sich zu einem großen Teil aus der Realität im Internet und der dortigen Verfügbarkeit von Information Retrieval Systemen. Als Konsequenz wurden sehr große Kollektionen und kurze Anfragen betont und Mehrsprachigkeit und multimediale Objekte rückten in den Fokus.

Die Bedeutung des Internet führte zur Einführung des Web-Track (cf. HAWKING 2001), bei dem nicht Zeitungstexte die Grundlage bilden sondern Internet-Dokumente. Um das übliche TREC Prozedere beibehalten zu können und den Systemen eine feste Datenmenge und Übungszeit zu bieten, speichern die Veranstalter eine Momentaufnahme eines Teils der im Internet angebotenen Daten. Davon liegen zwei verschieden große Versionen vor[3], welche hinreichend groß sind, um die Wirksamkeit von linkbasierten Verfahren zu testen. Bei der intellektuellen Überprüfung der Ergebnisse der Suchmaschinen achten die Juroren auf thematische Relevanz der Dokumente zu den formulierten Topics. Neben den thematischen Suchen wurde als weitere Aufgabe das Finden von *Homepages* eingeführt, für die als Anfrage z.B. der Name einer Institution vorlag.

PageRank wurde von der Universität Neuchatel in TREC eingesetzt (cf. SAVOY & RASOLOFO 2000). Die Ergebnisse des Web-Track in TREC mit PageRank und anderen Verfahren weisen darauf hin, dass die Berücksichtigung von Hypertext-Verknüpfungen die Ergebnisse des Retrievals verbessern kann. Dies gilt bereits bei den in TREC verwendeten Momentaufnahmen, die natürlich bei weitem nicht das gesamte Internet umfassen. Damit werden weder alle Links auf die in dem Sample enthaltenen Seiten erfasst, noch können alle in dem Sample vorkommenden Verknüpfungen benutzt werden, weil viele von ihnen auf Seiten außerhalb verweisen. Die Verbesserung der Ergebnisse konnte bisher nur für die Suche nach *Homepages* und damit für Suchen nach

[3] Die kleine Momentaufnahme besteht aus 1,7 Millionen Seiten (10 Gigabyte), während der große *snapshot* 18,5 Millionen Seiten (100 Gigabyte) umfasst (HAWKING 2001:1).

einer konkreten Seite nachgewiesen werden, während für die thematisch orientierten Suchen keinerlei Vorteile gemessen wurden (HAWKING 2001:10).

3.4 Nachteile

Die Nachteile von Autoritätsmaßen wie dem PageRank-Algorithmus liegen auf der Hand:

- Web-Seiten werden unabhängig ihres Inhalts und Kontexts bewertet. Ebenso wie Wissenschaften unterschiedliches Zitationsverhalten aufweisen, ist davon auszugehen, dass in verschiedenen Internet-Dokument-Typen unterschiedliche Verlinkungsneigung herrscht.
- Kritiker von bibliometrischen Analysen bemängeln, dass die Dynamik und Pragmatik des wissenschaftlichen Publizierens durch einfache Kennzahlen nicht hinreichend abgebildet wird (cf. z.B. FRÖHLICH 2000).
- Die Algorithmen können manipuliert werden. Dies geschieht sicher in hohem Maße, da ein erhebliches wirtschaftliches Interesse daran besteht, eigene Seiten bei vielen Internet-Suchen an vorderen Ranking-Positionen zu sehen.
- Die Algorithmen beruhen auf plausiblen Annahmen, berücksichtigen aber nicht die Sicht der Benutzer.
- Die Maße beziehen sich auf einzelne Seiten und können nicht direkt auf komplexere Angebote übertragen werden.

4 Entwicklung der Evaluierungsmethode

Um die Argumente für und wider besser abwägen zu können, wird im Folgenden versucht, diese Verfahren auf sogenannte Verzeichnisdienste im Internet (Web-Kataloge) anzuwenden. Die in dieser Studie entwickelte Methode zielt darauf ab, mit den Verfahren der Linkstrukturanalyse zu Erkenntnissen über die Benutzung der Verzeichnisdienste zu gelangen. Dabei muss das sich an Einzelseiten orientierte Maß In-Links auf komplexere Hypertext-Strukturen bezogen werden. Die Struktur der untersuchten Dienste ist durchweg hierarchisch.

4.1 Fragestellungen

Die Untersuchung soll vor allem folgende Aspekte untersuchen:

- Verweisen Links auf Verzeichnisdiensten eher auf Seiten, die in der Hierarchie weit oben stehen oder auf Unterkategorien zu spezifischen Themen?

- Zeichnen sich Unterschiede zwischen Verzeichnisdiensten für die Autorität ähnlicher Themen ab?

Die ethische Problematik des sog. *deep linking* dürfte hierbei kaum eine Rolle spielen. Von diesem erst jüngst wieder in den Schlagzeilen erscheinenden Phänomen spricht man, wenn Hyperlinks nicht auf die Homepage eines Web-Auftritts verweisen, sondern auf tiefer gelegene Seiten. Problematisch erscheinen solche Links, wenn der entsprechende Anbieter ein Geschäftsmodell verfolgt, nach dem er auf der Homepage Werbung platziert und auf den tieferen Seiten einen informationellen Mehrwert bietet (cf. SPINELLO 2001). Die Internetkataloge werben wenn überhaupt auch auf hierarchisch tieferen Seiten.

4.2 Systemarchitektur

Für diese Untersuchung wurde ein spezielles Web-Mining-System in JAVA 1.4 entwickelt, das auf mehreren frei im Internet verfügbaren Komponenten aufbaut. Die Seiten werden mit einem Roboter aus dem Netz übertragen. Dazu wurde ein Opensource-Roboter an die Bedürfnisse der Anwendung angepasst[4]. Die HTML-Seiten werden mit dem Tidy-Parser von W3C analysiert und in ein *Document Object Model (DOM)* Repräsentation überführt[5]. Anhand von Heuristiken wurden Regeln für jeden Verzeichnisdienst entworfen, die beschreiben, wie das System innerhalb der DOM-Repräsentation u.a. Titel, Hierarchie-Level, Unterkategorien und die eigentlichen Verweise auf andere Web-Angebote erkennt.

```
<?xml version="1.0" encoding="UTF-8"?>
<html><!-- ysx:2025869239 --><!-- 020513:0014 --><head>
    <meta content="HTML Tidy, see www.w3.org" name="generator"/>
    <title>Yahoo! Nachschlagewerke&gt;Statistiken</title>
    <base href="http://de.dir.yahoo.com/Nachschlagewerke/Statistiken/"/>
  </head>
  <body bgcolor="ffffff">
    <table cellpadding="0" cellspacing="0" width="100%" border="0">
      <tr>
        <td width="1%">
          <a href="http://de.yahoo.com">
            <img height="40" width="147" border="0" alt="Yahoo"
src="http://eur.i1.yimg.com/eur.yimg.com/i/de/g/yde4.gif"/>
          </a>
        </td>
        <td>
          <table width="100%" cellpadding="0" cellspacing="0" border="0">
            <tr>
```

Abb. 1: Ausschnitt aus der DOM-Repräsentation einer Yahoo-Seite

[4] http://www.matuschek.net/software/jobo/index.html.

[5] http://w3c.org.

Die Ergebnisse wurden in einer relationalen Datenbank abgelegt und mit einer Tabellenkalkulationssoftware ausgewertet.

Um die eingehenden Verbindungen einer Web-Seite zu erhalten, ist theoretisch eine vollständige Analyse des Internets nötig. Dies erfordert einen immensen Aufwand an Software und ist nicht bei jeder Untersuchung zu leisten. Die Crawler von Suchmaschinen durchsuchen das Web ohnehin und einige erlauben auch, die Links auf eine Seite abzufragen. Google und Altavista gehören dazu. In diesem Experiment wurde mit der Adresse der jeweiligen Seite eine gültige Abfrage in der Syntax der beiden Sprachen gestellt und die zurückgelieferte Ergebnisseite nach der entsprechenden Zahl durchsucht. Verbindungen aus dem Verzeichnisdienst selbst wie sie für die Interaktion unvermeidlich sind, wurden nicht eliminiert.

Experimente im Web-Mining leiden noch unter anderen Unwägbarkeiten und Schwierigkeiten, die zu einer gewissen Vagheit bei den Ergebnissen führen Bricht etwa während der langen Zeit des Retrievals die Netzverbindung kurzzeitig ab und überschreitet die Ausfallzeit einen bestimmten timeout, dann werden möglicherweise einige Seiten nicht erfasst. Dies gilt auch für die Abfrage der In-Links bei Google und Altavista, die durchaus auch kurzzeitig nicht erreichbar sind.

5 Ergebnisse der Evaluierung deutschsprachiger Internet-Verzeichnisdienste

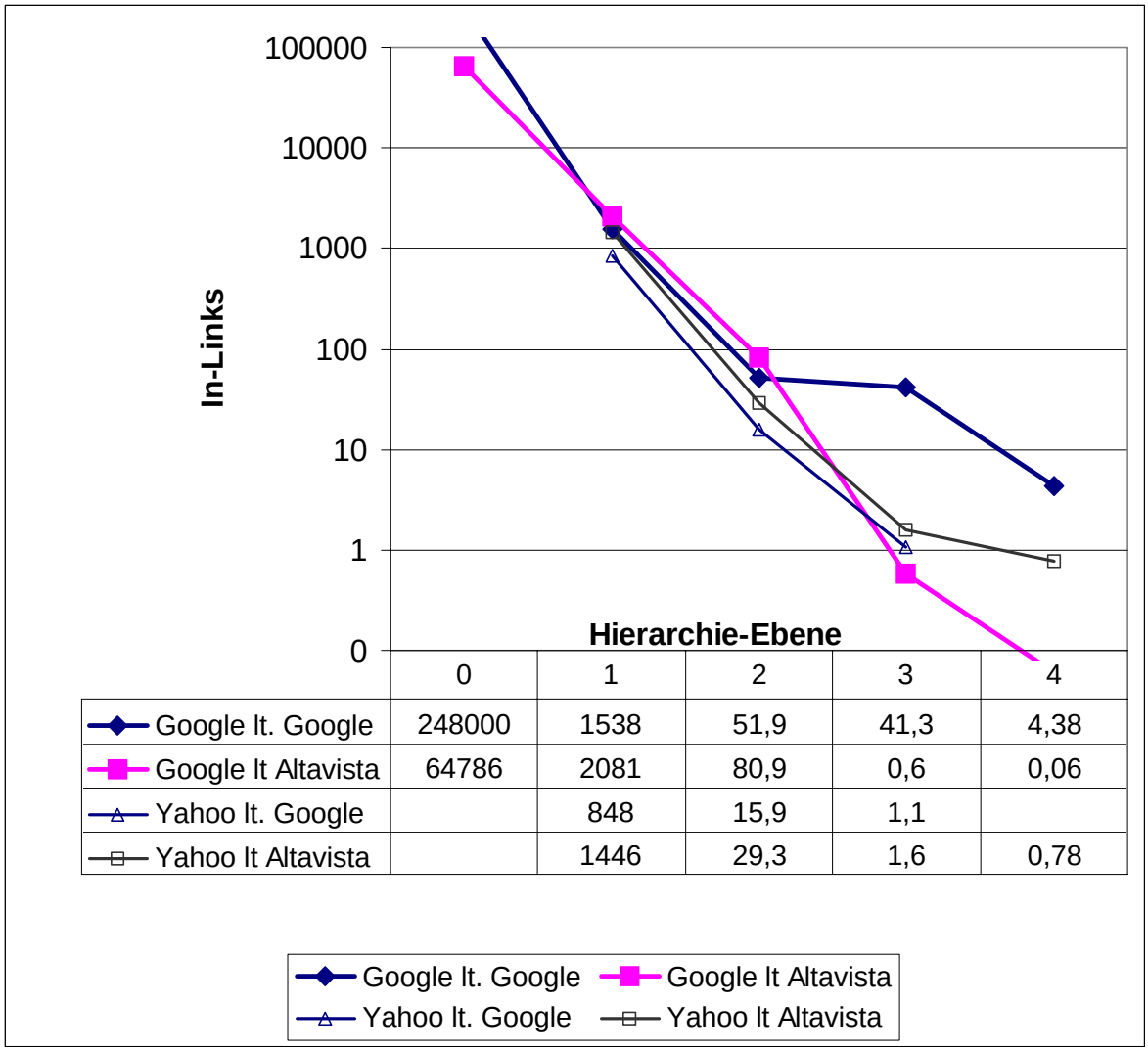
Die Bewertung von Verzeichnisdiensten anhand von Linkanalysen bietet einen Einblick in die Benutzung von Verzeichnisdiensten allgemein. Die Analyse zeigt an einigen Parametern, wann Verbindungen auf Seiten in Verzeichnisdiensten gelegt werden.

Die Untersuchung konnte vor allem beweisen, dass die Zahl der Verbindungen auf einen Verzeichnisdienst sehr stark abnimmt, je tiefer die Seite in der hierarchischen Struktur steht. Dazu wurde der Dienst Google auf mehreren Ebenen untersucht. Trägt man die Abhängigkeit in einer logarithmischen Skala ein, dann ergibt sich ein fast linearer Zusammenhang (siehe Abb. 2). Die Ebene Null entspricht dabei der Einstiegsseite[6], während die folgende Adresse ein Beispiel für eine Seite der vierten Ebene darstellt: Computer/Software/

[6] http://directory.google.com/Top/World/Deutsch/.

Spezialsoftware/Sport[7]. Analysiert wurden ca. 3000 Seiten des Internet-Katalogs von Google und ca. 500 Seiten der deutschen Version von Yahoo.

Abb. 2: Zahl der durchschnittlichen In-Links pro Ebene



	0	1	2	3	4
Google lt. Google	248000	1538	51,9	41,3	4,38
Google lt Altavista	64786	2081	80,9	0,6	0,06
Yahoo lt. Google		848	15,9	1,1	
Yahoo lt Altavista		1446	29,3	1,6	0,78

Die extrem starke Abnahme der eingehenden Links von meist mehr als den Faktor zehn weisen sowohl die Seiten von Google als auch die von Yahoo auf. Ähnliche Beziehungen ergaben sich auch bei anderen Analysen eingehender Links von ADAMIC & HUBERMANN 2001. Darin konnten die Autoren zeigen, dass die Zahl der Seiten mit vielen eingehenden Verbindungen sehr stark abnimmt bis hin zu Seiten ohne In-Links. Auch dabei ergab sich auf einer logarithmischen Skala ein linearer Zusammenhang.

[7] http://directory.google.com/Top/World/Deutsch/Computer/Software/Spezialsoftware/Sport.

Obwohl die Ergebnisse von Google und Altavista für die Abfrage der In-Links oft sehr stark voneinander abweichen, gilt der oben vorgestellte Zusammenhang bei beiden Suchmaschinen.

Die Abhängigkeit, die sich hier für eine Untermenge der Google- und Yahoo-Seiten ergibt, gilt möglicherweise auch für andere hierarchisch aufgebaute Seiten.

Die Autoren von Internet-Seiten verweisen also grundsätzlich eher auf generelle Seiten in Verzeichnisdiensten und weniger auf Seiten zu spezifischen Themen. Offensichtlich wollen sie dem Benutzer die Suche nach für ihn interessanten Abgeboten durch Browsing selbst überlassen. Möglicherweise bewerten die Autoren Seiten der Verzeichniseinträge als gut geeignet für thematisch breite Informationsbedarfe oder für Informationsbedarfe, die noch vage sind und bei denen der Benutzer sich zunächst einen Überblick verschaffen will. Eine Interpretation, dass die redaktionelle Arbeit für die Dienste nicht gewürdigt wird, geht sicher zu weit.

Bei der Anzahl der Einträge pro Seite und der Anzahl der Unterkategorien (mit Querverweisen) ergaben sich keine interessanten Aspekte.

5.1 Vergleich von Verzeichnisdiensten

Ein Vergleich von Verzeichnisdiensten aufgrund von In-Links muss berücksichtigen, dass auf den unteren Ebenen grundsätzlich weniger Verbindungen zu erwarten sind. Im Folgenden werden die Startseiten mehrerer Dienste verglichen. Dabei ist allerdings ist die Unsicherheit größer, die sich aus der Abfrage zweier unterschiedlicher Suchmaschinen nach den eingehenden Verbindungen ergibt. Folgende Tabelle zeigt die eingehenden Links für einige deutschsprachige Verzeichnisdienste bzw. die deutschsprachigen Unterkategorien global angelegter Verzeichnisdienste.

Verzeichnis-dienst	Allesklar	Altavista	Dino	dmoz.org	Google	Web.de
Lt. Google	2390	374	3610	19000	14700	22
Lt. Altavista	51581	1905	32999	1886	27110	14946

Tabelle: Links zu den Startseiten einiger deutscher Verzeichnisdienste

Die Unterschiede zwischen der Anzahl von ankommenden Verbindungen zwischen Google und Altavista sind teilweise sehr hoch, so dass die Aussage-

kraft dieser Zahlen nicht sehr groß ist. Die geringen Werte von Web.de und Altavista bei Google erklären sich daher, dass weitaus mehr Links auf die Suchseiten zeigen als auf die Verzeichnis-Seiten.

Eine weitere Vergleichsmöglichkeit besteht in der Analyse der Seiten zu ähnlichen Themen. Eine Automatisierung ist hier schwierig, da nicht in allen Fällen die gleichen Beschreibungen verwendet werden. Dazu wurden zwei Seiten auf Ebene drei und vier ausgewählt, die in Google in dieser Ebene die höchsten Werte erreichten. Dies waren eine Seite zu Geisteswissenschaften/ Geschichte und eine zu Internet/WWW/Firmen. Diese Seiten ragen aus den sonst auf dieser tiefen Ebene wenig referenzierten Seiten deutlich heraus. In Yahoo hatte eine vergleichbare Seite zur Geschichte auch sehr viele In-Links, ohne dass hier eine vollständige Analyse aller Seiten durchgeführt worden wäre. Dagegen wurde auf eine ähnliche Seite mit Firmen für das WWW sehr viel weniger häufig verlinkt. Dies kann als Hinweis darauf gelten, dass thematisch ähnliche Seiten verschiedener Verzeichnisdienste sehr unterschiedliche Qualität aufweisen.

5.2 In-Links von referenzierten Seiten

Ein wichtiger Maßstab für die Qualität von Verzeichnisdiensten kann neben der Anzahl von Verbindungen, die auf die Seiten des Dienstes verweisen, die Anzahl der Verbindungen für die darin enthaltenen Angebote sein. Korreliert also die Zahl der Verbindungen auf eine Seite eines Verzeichnisdienstes mit der Autorität der Seiten, auf die der Dienst verweist, gemessen an den Verweisen auf diese Seiten? Finden die Redakteure der Dienste solche Web-Angebote, die auch laut Linkanalyse eine hohe Autorität besitzen? Oder unterscheidet sich das menschliche Urteil von den automatisch ermittelten Zahlen?

Um dies näher untersuchen zu können, wurde bei der oben bereits erwähnten Stichprobe von ca. 400 Seiten aus Yahoo neben den Links auf diese Seiten auch die dort verzeichneten Seiten untersucht. Für jeden Eintrag in diesen Yahoo Seiten wurde eine Anfrage sowohl an Google als auch an Altavista geschickt, um die Anzahl der *back links* dieser Angebote zu erhalten. Für jede Yahoo-Seite wurde der Durchschnitt jeweils für die Google- und die Altavista-Ergebnisse gebildet.

Zunächst zeigt sich, dass eine sehr große Streuung besteht. Bei einem Mittelwert von 426 eingehenden Verbindungen weisen die Ergebnisse von Google eine Standardabweichung von 1.346 Verbindungen auf. Altavista liefert als

Mittelwert 79 und mit 209 ebenfalls eine sehr hohe Standardabweichung. Die Autorität und damit die vermutete Qualität der referenzierten Seiten wäre demnach sehr unterschiedlich.

Interessanterweise ergibt sich für die Stichprobe keinerlei Korrelation zwischen den *back links* der Verzeichnisdienst-Seite und den *back links* der dort genannten Web-Angebote. Der Betrag der berechneten Korrelation liegt unter 0,1. Angesichts der obigen Ergebnisse, nach denen auf in der Hierarchie tiefer liegende Seiten kaum verwiesen wird, sollte man aber Seiten auf unterschiedlichen Ebenen getrennt betrachten. Möglicherweise überlagert der in Abbildung 2 skizzierte Effekt der starken Abnahme der Wahrscheinlichkeit für eingehende Verbindungen die gesuchte Korrelation. Daher wurden die Korrelationen auch für die Seiten nur einer Ebene berechnet, allerdings ergab sich auch hier keine nennenswerte Korrelation.

Die menschliche und die automatische aus den Verbindungen abgeleiteten Qualitätsurteile stimmen demnach nicht überein.

Von allen untersuchten Parametern ergab sich lediglich für die Anzahl von Unterkategorien einer Seite eine positive Korrelation mit einem Betrag von mehr als 0,5. Demnach steigt die Wahrscheinlichkeit für Verbindungen auf eine Seite eines Verzeichnisdienstes leicht an, wenn die Seite viele Unterkategorien enthält. Dagegen wirkt sich die Anzahl der enthaltenen Verweise auf externe Angebote weder positiv noch negativ aus. Dies konnte so nicht unbedingt erwartet werden, da die Leistung der Dienste gerade in der Bewertung externer Links liegt.

Für denjenigen, der einen Link auf eine Dienst-Seite legt, scheint der Aspekt des Browsing bei der Benutzung im Vordergrund zu stehen. Möglicherweise sollen dem Benutzer, der den Link verfolgt, noch viele Optionen geboten werden. Zum einen wird also auf Seiten verwiesen, die selbst eher viele weitere Kategorien enthalten, aus denen ausgewählt werden kann. Zum anderen wird auch sehr viel häufiger auf Seiten verwiesen, die hoch in der Hierarchie stehen und von denen aus der Benutzer noch mehrere Ebenen weiter verfolgen kann. Damit weisen die beiden Ergebnisse aus der Linkanalyse in die gleiche Richtung.

Die Gründe für das Setzen von Links auf Seiten eines Verzeichnisdienstes umfassen sicher noch viele weitere Aspekte. Und das Erstellen von Hypertext-Verbindungen kann nur ein Anhaltspunkt für die Benutzung von solchen Diensten bieten. Trotzdem lassen sich mit der Linkanalyse interessante An-

haltspunkte gewinnen. Weitergehende Evaluierungen müssten u.a. folgende Aspekte beachten, um zu genaueren Aussagen zu kommen:

- Wie stark überschneiden sich die Einträge in den Diensten? Ist der Unterschied der gemessenen Autorität der Verzeichnisseiten dadurch gerechtfertigt?
- Finden sich die thematisch ähnlichen Seiten auf der gleiche Ebene? Ansonsten muss berücksichtigt werden, dass eine tiefere Einbettung eine geringe Wahrscheinlichkeit für In-Links mit sich bringt.
- Inwieweit unterscheidet sich die Neigung, Verbindungen zu setzen über Themen hinweg? Ebenso wie die Wissenschaften unterschiedliches Zitierverhalten offenbaren, so könnten auch verschiedene Nutzergruppen unterschiedlich häufig Links setzen. Dies könnte über mehrere Verzeichnisdienste hinweg untersucht werden.

5.3 Unterschiede zwischen den Diensten

Allesklar und Web.de linken nicht direkt auf die Einträge, sondern auf eine interne Datenstruktur, die auf die Site des Eintrags verweist. Dies spielt für den Benutzer kaum eine Rolle, allerdings nutzen die beiden Verzeichnisdienste Allesklar und Web.de auch kryptische, mit Zahlen kodierte URLs für ihre Seiten. Dies verringert die Benutzerfreundlichkeit für verweisende Autoren und damit die Zahl der potenziellen In-Links. Auch Altavista kommt nicht ohne für den Benutzer kryptische Zeichen in der URL aus, da die Entwickler hier Parameter einsetzen. Entgegen den anderen untersuchten Diensten zeigt Altavista die Einträge vor den Unterkategorien an. Für Benutzer, die sich an andere Verzeichnisdienste gewöhnt haben, ist dies ungewohnt und kann sogar dazu führen, dass die Kategorien gar nicht wahrgenommen werden. Dino lässt dagegen nur Einträge auf der untersten Ebene zu, so dass Einträge von externen Seiten und die Kategorien getrennt bleiben. Bei einer Analyse der Einträge und Unterkategorien muss dies berücksichtigt werden.

6 Ausblick

Die Link-Analyse im Internet wird in diesem Beitrag exemplarisch auf Verzeichnisdienste angewandt, um zu zeigen, wie die seiten-bezogenen Kennzahlen auf größere Einheiten wie Sites übertragen werden können. Die Untersuchung zeigt einige interessante Ergebnisse für die Analyse des Web-Verzeichnisses ebenso wie für die themenspezifischen Analysen. Die Wahrscheinlichkeit eines Links auf Seiten eines Internet-Katalogs nimmt mit der

hierarchischen Tiefe dieser Seiten sehr stark ab. Die Benutzung der Kataloge beschränkt sich für das Setzen von Links auf sehr generelle Seiten.
Vergleicht man die Anzahl der Links auf eine Seite in einem Katalog, so korreliert sie nicht mit der Anzahl der Links auf die darin aufgeführten Sites. Das intellektuelle Qualitätsurteil der Katalogredakteure geht also nicht konform mit der Linkanalyse. Die Anzahl der eingehenden Verbindungen reicht für die Bewertung von Web-Angeboten nicht aus.

Eine mögliche Erweiterung dieser Analyse liegt in der Verfeinerung der Methode. So könnten etwa die Schnittmenge zwischen Verzeichnisdiensten in Bezug auf die referenzierten Seiten gemessen werden. Darauf aufbauend ließe sich untersuchen, ob die gleichen Web-Seiten auch weitgehend in gleiche thematische Kategorien eingeordnet werden.

Eine mögliche Erweiterung liegt in der Einbeziehung internationaler Verzeichnisdienste. Möglicherweise zeigen sich dabei kulturspezifische Unterschiede in der hierarchischen Organisationsstruktur. Eine Anwendung der hier vorgestellten Methodik kann in der Suche innerhalb von Meta-Verzeichnisdiensten liegen. Durch die Link-Analyse können die Seiten unterschiedlicher Dienste zu einer Thematik auf ihre Autorität untersucht werden und die besten dann präsentiert werden.

7 Literaturverzeichnis

ADAMIC, Lada; HUBERMANN, Bernardo (2001): The Web's Hidden Order. Communications of the ACM 44(9). S. 55-59.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier (Hrsg.)(1999): Modern Information Retrieval. Addison-Wesley.

BARTEL, Torsten (2002): Verbesserung der Usability von WebSites auf der Basis von Web Styleguides, Usability Testing und Logfile-Analysen. Magisterarbeit Universität Hildesheim.

BEKAVAC, Bernard (1999): Suche und Orientierung im WWW. Verbesserung bisheriger Verfahren durch Einbindung hypertextspezifischer Informationen. Konstanz: Universitätsverlag [Schriften zur Informationswissenschaft Bd. 37].

CONSTANTOPOULOS, Panos; SOLVBERG, Ingeborg (Hrsg.): Proc 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001).

EIBL, Maximilian; MANDL, Thomas (2002): Including User Strategies in the Evaluation of Graphic Design Interfaces for Browsing Documents. In: SKALA, Vaclav (Hrsg.): 10th Intl Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG 2002). Pilsen. S. 163-169.
http://wscg.zcu.cz/wscg2002/Papers_2002/B89.pdf

FAYYAD, Usama (1997): Editorial. In: Data Mining & Knowledge Discovery 1(1). S. 5-10.

FAYYAD, Usama; UTHURUSAMY, Ramasamy (1996): Data Mining and Knowledge Discovery in Databases. In: Communications of the ACM 39 (11) S. 24.

FRÖHLICH, Gerhard (2000): Online Informationsvorenthaltung als Strategem wissenschaftlicher Kommunikation. In: ZIMMERMANN, Harald; SCHRAMM, Volker (Hrsg.): Knowledge Management und Kommunikationssysteme: Workflow Management, Multimedia, Knowledge Transfer. Proc 6th Intl Symposium für Informationswissenschaft. (ISI ´98). Prag. S. 535-549.

FROMMHOLZ, Ingo (2001): Categorizing Web Documents in Hierarchical Catalogues. In: Proc 23rd Colloquium on Information Retrieval Research. Darmstadt. http://www.darmstadt.gmd.de/~frommhol/frommholz_ecir01.pdf

GIBSON, David; KLEINBERG, Jon; RAGHAVAN, Prabhakar (1998): Inferring Web Communities from Link Topology. In: Proc. 9th ACM Conf on Hypertext and Hypermedia. http://citeseer.nj.nec.com/gibson98inferring.html

HAWKING, David (2001): Overview of the TREC-9 Web Track. In: VOORHEES & HARMAN 2001.

KLEINBERG, Jon (1998): Authoritative Sources in a Hyperlinked Environment. In: Proc 9th ACM-SIAM Symposium on Discrete Algorithms.

KNORZ, Gerhard; KUHLEN, Rainer (Hrsg.): Informationskompetenz – Basiskompetenz in der Informationsgesellschaft. Proc 7. Int. Symposium für Informationswissenschaft. (ISI 2000). Darmstadt.

KRUSCHWITZ, Udo (2001): Exploiting Structure for Intelligent Web Search. In: Proc Hawai Intl Conf on System Sciences. IEEE.

LEMPEL, R.; MORAN, S. (2000): The Statistical Approach for Link-Structure Analysis (SALSA) and the TKC Effect. In: Proc 9th Intl WWW Conference. http://www9.org/w9cdrom/175/175.html

MAHOUI, Malika; CUNNINGHAM, Sally (2001): Search Behavior in a Research-Oriented Digital Library. In: CONSTANTOPOULOS & SOLVBERG 2001 S. 13-24.

MANDL, Thomas (2001): Tolerantes Information Retrieval: Neuronale Netze zur Erhöhung der Adaptivität und Flexibilität bei der Informationssuche. Konstanz: Universitätsverlag [Schriften zur Informationswissenschaft Bd. 39].

MARCHIONINI, Gary (1995): Information Seeking in Electronic Environments.

MEGHABGHAB, George (2002): Discovering authorities and hubs in different topological web graph structures. In: Information Processing and Management 38. S. 111-140.

MUTSCHKE, Peter (2001): Enhancing Information Retrieval in Federated Bibliographic Data Sources Using Author Network Based Stratagems. In: CONSTANTOPOULOS & SOLVBERG 2001. S. 287-299.

NIE, Jian-Yun,; SIMARD, Michael; FOSTER, George (2001). Multilingual information retrieval based on parallel texts from the web. In: PETERS, Carol (Hrsg.): Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Information Evaluation Forum (CLEF 2000) Lisbon, Portugal, 21.-22. Sept. 2000. Springer [LNCS 2069] S. 188-201.

PAGE, Larry; BRIN, Sergey; MOTWANI, R.; WINOGRAD; T. (1998): The PageRank Citation Ranking: Bringing Order to the Web. http://citeseer.nj.nec.com/page98pagerank.html

PIROLLI, Peter; PITKOW, James; RAO, Ramana (1996): Silk from a Sow's Ear: Extracting Usable Structures from the Web. In: Proc Conference on Human Factors and

Computing Systems.
http://www.acm.org/pubs/articles/proceedings/238386/p118-pirolli/118-pirolli.html

RITTBERGER, Marc (2001): Quality Measuring with respect to electronic information markets and particulary online databases. In: KENT, A. (Hrsg.): New York, NY: Marcel Dekker, 31, Chapter: 68, S. 274-295.
http://www.inf-wiss.uni-konstanz.de/People/MR/pubs/elis_rittberger.pdf

SAVOY, Jacques; RASOLOFO, Yves (2000): Report on the TREC-9 Experiment: Link-based Retrieval and Distributed Collections. In: VOORHEES & HARMAN 2000. S. 579.

SCHLÖGL, Christian (2000): Informationskompetenz am Beispiel einer szionometrischen Untersuchung zum Informationsmanagement. In: KNORZ & KUHLEN 2000: Informationskompetenz - Basiskompetenz in der Informationsgesellschaft. Proc 7. Int. Symposium für Informationswissenschaft. (ISI 2000). Darmstadt. S. 89-112.

SPINELLO, Richard (2001): An Ethical Evaluation of Web-Site Linking. In: SPINELLO, Richard; TAVANI, Herman (Hrsg.): Readings in CyberEthics. Sudbury, MA et al.: Jones and Bartlett. S. 295-308.

SULLIVAN, Terry (1997): Reading Reader Reaction: A Proposal for Inferential Analysis of Web Server Log Files. In: Proc 3rd Conf on Human Factors and the Web.
http://www.pantos.org/ts/papers/rrr.html

VOORHEES, Ellen; HARMAN, Donna (2000)(Hrsg.): The Ninth Text REtrieval Conference (TREC 9). NIST Special Publication 500-249.
http://trec.nist.gov/pubs/trec9/t9_proceedings.html

VOORHEES, Ellen; HARMAN, Donna (2001)(Hrsg.): The Tenth Text REtrieval Conference (TREC 10). NIST Special Publication 500-250.
http://trec.nist.gov/pubs/trec10/t10_proceedings.html

WALTHER, Ralf (2001): Web Mining. In: Informatik Spektrum 24(1). S. 16-18.

WEISS, R.; VELEZ, B.; SHELDON, M.; MANPREMPRE, C.; SZILAGYI, P.; DUDA, A.; GIFFORD D. (1996): HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering. In : Proc Seventh ACM Conference on Hypertext.