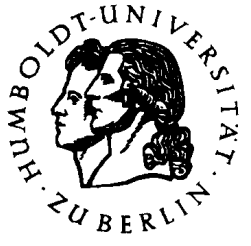
HUMBOLDT-UNIVERSITÄT ZU BERLIN
INSTITUT FÜR BIBLIOTHEKSWISSENSCHAFT



BERLINER HANDREICHUNGEN
ZUR BIBLIOTHEKSWISSENSCHAFT

HEFT 129

ENTWICKLUNG UND TEST EINER LOGFILEBASIERTEN METRIK ZUR ANALYSE VON WEBSITE ENTRIES AM BEISPIEL EINER AKADEMISCHEN UNIVERSITÄTS-WEBSITE

VON
PHILIPP MAYR

ENTWICKLUNG UND TEST EINER LOGFILEBASIERTEN METRIK ZUR ANALYSE VON WEBSITE ENTRIES AM BEISPIEL EINER AKADEMISCHEN UNIVERSITÄTS-WEBSITE

VON
PHILIPP MAYR

Berliner Handreichungen
zur Bibliothekswissenschaft

Begründet von Peter Zahn
Herausgegeben von
Konrad Umlauf
Humboldt-Universität zu Berlin

Heft 129

Mayr, Philipp

Entwicklung und Test einer logfilebasierten Metrik zur Analyse von Website Entries am Beispiel einer akademischen Universitäts-Website / von Philipp Mayr. - Berlin : Institut für Bibliothekswissenschaft der Humboldt-Universität zu Berlin, 2004, 106 S. - (Berliner Handreichungen zur Bibliothekswissenschaft und Bibliothekarsausbildung ; 129)

ISSN 14 38-76 62

Abstract:

Web Logfiles protokollieren Benutzertransaktionen auf Webservern und bieten aufgrund ihres Umfangs, ihrer Eigenschaften und Potenziale ein ausgezeichnetes Untersuchungsfeld für heutige Informations- und Onlineverhaltensstudien. Die empirische, explorative Untersuchung aus den Bereichen Web Mining, Webometrics und Logfileanalyse stellt neue Gesichtspunkte und Analysemöglichkeiten für Logdaten vor. Zu diesem Zweck entwickelt und testet die Arbeit ein quantitatives, nicht-reaktives Messverfahren (Logmetrik „Web Entry Faktoren"), das anhand von einfachen Web Logdaten, Aussagen über die Zugänglichkeit und Sichtbarkeit von hochfrequentierten Einstiegspunkten einer Website ermöglicht. Im Mittelpunkt stehen die drei unterscheidbaren Navigationsarten im Web „Navigation über Suchmaschinen", „Navigation über Backlinks" und „direkte Navigation". Die Untersuchung integriert ein Klassifikationsschema für Webseiten sowie den prominenten externen Parameter PageRank der heute wichtigsten Suchmaschine Google. Untersuchungsgegenstand sind Web Logfiles zweier kompletter Jahrgänge (2000 und 2002) des Webservers des Instituts für Bibliothekswissenschaft an der Humboldt-Universität zu Berlin (http://www.ib.hu-berlin.de/), sowie die 100 am häufigsten genutzten Einstiegsseiten dieser akademischen Universitäts-Website.

Diese Veröffentlichung geht zurück auf eine Magisterarbeit im Magisterstudiengang Bibliothekswissenschaft (Master of Arts, M.A.) an der Humboldt-Universität zu Berlin (Library and Information Science).

Inhalt

1 Einleitung

Web-bezogene Studien sind eine relativ neue und sich rasch ausdifferenzierende interdisziplinäre Forschungsdisziplin (vgl. *Spink*, 2002 [86]). Basis dieser meist empirisch orientierten Forschung ist das bislang ungebremste starke Wachstum an Webusern, Suchmaschinen und Websites. Die Untersuchung der Navigation und Dynamik im Web spielt seit Beginn der Internetforschung eine bedeutsame Rolle (vgl. *Huberman et al.*, 1998 [44], *Lawrence & Giles*, 1999 [58]). Navigation im Web erfolgt zum Großteil über Hyperlinks, die entweder intellektuell erstellt (z.B. Verzeichniseintrag) oder automatisch auf Anfrage generiert werden (z.B. Suchmaschinenergebnis). „Direkte" Navigation koexistiert neben der „linkbasierten" Navigation und kann als ein Gradmesser für die Bekanntheit und Etablierung einer Website bzw. seiner Inhalte gelten.

Web Logfiles protokollieren Benutzer-Transaktionen auf Webservern und bieten aufgrund ihres Umfangs, ihrer Eigenschaften und Potenziale einen ausgezeichneten Untersuchungsgegenstand für heutige Informations- und Onlineverhaltensstudien. Aus Web Logfiles lassen sich neben Informationen zum Benutzungs- bzw. Navigationsverhalten seiner Besucher weitergehende Informationen extrahieren, die allgemeine Rückschlüsse über Zugangswege (accessibility), Sichtbarkeit (visibility) und Verlinkungen (interlinking) von Webinhalten möglich machen. Die folgende empirische, explorative Untersuchung aus den Bereichen Web Mining, Webometrics und Logfileanalyse stellt neue Gesichtspunkte und Analysemöglichkeiten für Logdaten vor. Zu diesem Zweck entwickelt und testet die Arbeit ein quantitatives, nicht-reaktives Messverfahren (Logmetrik „Web Entry Faktoren"), das anhand von einfachen Web Logdaten, Aussagen über Zugänglichkeitsmuster (Navigationspfade und Navigationsarten) einer Website ermöglicht. Eine Idee der Untersuchung ist es die Bedeutung der externen Linksstrukturen für eine Website einschätzen und beziffern zu können. Weiterhin liefern die erhobenen seitenbezogenen Zugangsinformationen eine Grundlage für Optimierungs- und Evaluierungsverfahren bzgl. einer Website. Die Untersuchung integriert ein Klassifikationsschema für Webseiten (*Haas & Grams*, 2000 [41]) sowie den prominenten externen Parameter *PageRank* der heute wichtigsten Suchmaschine Google. Untersuchungsgegenstand sind Web Logfiles zweier kompletter Jahrgänge (2000 und 2002) des Webservers des Instituts für Bibliothekswissenschaft an der Humboldt-Universität zu Berlin (http://www.ib.hu-berlin.de/), sowie die 100 am häufigsten genutzten Einstiegsseiten dieser Website.

2 Bezugsrahmen

Akademische Websites werden für unterschiedliche Zwecke und Benutzergruppen (vgl. *Middleton et al.*, 1999 [66]) weltweit in großer Zahl (vgl. *Lawrence & Giles*, 1999 [58]) erstellt und meist fortlaufend erweitert und verändert (vgl. *Koehler*, 2001 & 2002 [51,52]). Diese Websites dienen mit ihren vielgestaltigen Internetressourcen einem sehr breiten und heterogenen Benutzerkreis. Grundsätzlich lassen sich zu akademischen Websites alle Webserver zählen, die universitäre, ausbildungs- und forschungsrelevante oder andere wissenschaftliche und akademische Inhalte (scholarly content) anbieten und über Standardbrowser zugänglich sind. Darunter zählt vor allem die große Anzahl der Universitätswebsites, wissenschaftliche E-Journals und ein weites Spektrum verwandter Websitetypen wie z.B. Preprint-Server, Konferenzwebsites, wissenschaftliche Diskussions-foren und Mailinglisten, usw., die akademische Inhalte zwecks wissenschaftlichem Informationsaustausch (scholarly communication) zugänglich machen. Untersuchungen zeigen, dass die Inhalte akademischer Webseiten in verschiedene Typen (proto-typology) eingeteilt werden können (vgl. *Cronin et al*., 1998 [23]). Weiterhin finden sich in der Literatur zahlreiche Untersuchungen zu einzelnen Bereichen akademischer Websites.

- Universitätswebsites (*Middleton et al.*, 1999 [66], *Thomas & Willet*, 2000 [113], *Smith & Thelwall*, 2002 [83], *Thelwall*, 2002 [104]),
- E-Journals (*Kim*, 2000 [48]),
- Journal und themenspezifische Websites (*Larson*, 1996 [55], *Hernández-Borges*, 1999 [43]).

Akademische Websites bieten aufgrund ihres Umfangs, der Beschaffenheit der Inhalte, der Vielzahl ihrer internen und externen Verknüpfungen sowie der Intensität der Benutzung ihrer Webangebote einen ausgezeichneten Untersuchungsgegenstand für heutige Informations- und Onlinebehavior-Studien. Vor allem in Studien und Modellen zur Beschreibung von Verlinkungsphänomenen in der informationswissenschaftlichen Fachdiskussion wurden akademische Websites als Untersuchungsobjekt herangezogen (*Rousseau*, 1997 [82], *Ingwersen*, 1998 [46], *Thelwall et al.*, 2001-2002 [93, 109, 95, 98, 101, 104]).
Im Mittelpunkt dieser Untersuchung steht eine umfangreiche deutschsprachige akademische Website eines Universitätsinstituts aus dem Forschungsbereich Bibliothekswissenschaft, Dokumentation und Informationswissenschaft.

Im folgenden werden die Forschungsdisziplinen vorgestellt, die Bezugspunkte zu der hier dargestellten Untersuchung aufweisen.

2.1 Web Mining

> *„The Web mining research is at the cross road of research from several research communities, such as database, information retrieval, and within AI, especially the sub-areas of machine learning and natural language processing." Kosala & Blockeel, 2000 [53]*

Die Basis aller Web Mining Bemühungen besteht darin, dass angenommen wird, das die anfallenden Daten im Web ausreichend strukturiert sind, um mit Algorithmen nach Mustern zu suchen. Unter Web Mining werden allgemein Data Mining-Techniken verstanden, die zum automatischen Auffinden und Extrahieren von nützlichen Informationen aus Webdokumenten und -services dienen (vgl. *Etzioni*, 1996 [31], *Kosala & Blockeel*, 2000 [53], *Cooley et al.*, 1997 [20]). *Kosala & Blockeel* unterteilen folgende typische Aufgabenbereiche des Web Minings.

1. Auffinden von Ressourcen (Webdokumente),
2. Auswahl von Informationen und Pre-Processing,
3. Generalisierung (automatische Identifikation allgemeiner Muster auf Websites),
4. Analyse (Überprüfung und/oder Interpretation der extrahierten Muster).

Weiterhin unterscheiden die Autoren das Forschungsgebiet Web Mining in drei verschiedene Kategorien.

1. Web Content Mining beschreibt die Identifikation von nützlichen Informationen in unstrukturierten Webinhalten, -daten, -dokumenten.

   > *"Web content mining focuses on techniques for searching the web for documents whose content meets web users queries." [8]*

2. Web Structure Mining beschreibt die Erforschung der dem Web zugrundeliegende Linkstruktur.

   > *„Web structure mining could be used to discover authority sites for the subjects (authorities) and overview site for the subjects that point to many authority (hubs)." [53]*

3. Web Usage Mining versucht Muster (user behavior) aus den Daten (secondary data) zu gewinnen, die durch die Webuser auf Webservern generiert werden.

   > *„Web usage mining can be viewed as the extraction of usage patterns from access log data containing the behavior characteristics of users. (Srivastava et al., 2000 [88], vgl. [8])*

Die Autoren heben hervor, dass die einzelnen Kategorien nicht eindeutig voneinander zu trennen sind und die verschiedenen Web Mining Ausprägungen durchaus in unterschiedlichen Kombinationen eingesetzt werden können [53]. Die Web Mining Forschung entwickelt sich nach *Kosala & Blockeel* in die Richtung der Informationsintegration (z.B. web knowledge base) sowie dem Einsatz und der Entwicklung von webbasierten (neuen) Algorithmen für maschinelles Lernen.

2.2 Webometrics

Webometrics (dt. Webometrie, vgl. Cybermetrics[1]) kann seit Mitte der 1990er Jahre als neues Forschungsgebiet in der Informationswissenschaft wahrgenommen werden. Im Mittelpunkt dieser Disziplin, die eine starke methodische Verwandtschaft zur Informetrie[2] und Bibliometrie[3] aufweist, steht das Bestreben neue Regeln, Charakterisierungen und Ergebnisse über das Netzwerk-Phänomen Internet (WWW) als Zitationsnetzwerk (citation network) zu gewinnen (vgl. *Larson*, 1996 [55], *Almind & Ingwersen*, 1997 [2], *Boudourides et al.*, 1999 [12], *Cronin*, 2000 [24]). Im Mittelpunkt stehen quantitative Methoden. Die drei folgenden Zitate verdeutlichen das Einsatzgebiet und die Methoden der Webometrie. Zur wissenschaftlichen Einordnung und Abgrenzung der Webometrie siehe Abbildung 2-1 (vgl. *Björneborn*, 2002 [11]).

[1] Seit 1997 erscheint die elektronische Zeitschrift "Cybermetrics" (siehe http://www.cindoc.csic.es/cybermetrics/), die webometrische bzw. cybermetrische Untersuchungen in englischer Sprache herausgibt. *Björneborn* versteht Cybermetrics als den Forschungsbereich, der sich mit quantitativen Studien zu allen Bereichen des Internets beschäftigt und definiert somit Cybermetrics weiter als Webometrics (vgl. *Björneborn*, 2002 [11]).

[2] "While bibliometrics and scientometrics refer to all quantitative aspects and models of printed media and sciences, informetrics is not limited to media or scientific communication. (…) Neither it is restricted to scientific research. However, it is considered usable for tasks such as issue management, gathering of business intelligence and research evaluation. (…) Informetrics is, thus, an emerging subfield in information sciences, which is based on the combination of advances of information retrieval and quantitative studies of information flows." (*Boudourides et al.*, 1999 [12])

[3] "Bibliometrics is traditionally associated with the quantitative measure of documentary materials and it embraces all studies which seek to quantify the process of written communication, i.e., the application of mathematical methods to books and other media of communication. Bibliometric methods are used especially in studies of properties and behaviour of recorded knowledge, for analysing the structures of scientific and research areas, and for evaluation of research activity and administration of scientific information. Various statistical methods are applied to study, for example,

- *"The idea is to use the traditional bibliometric applications for the web that include the study of communication patterns, the identification of research fronts, historical studies of the development of a discipline or domain, and the evaluation of research activities of countries, institutions or individuals." (Boudourides et al., 1999 [12])*
- *"… the study of quantitative aspects of the construction and use of information resources, structures and technologies on the Web, drawing on bibliometric and informetric methods" (Björneborn, 2002 [11])*
- *"We define webometrics … to be the quantitative study of web-related phenomena. We use this broad definition to encompass research from disciplines outside of Information Science such as Communication Studies, Statistical Physics and Computer Science. The extra topics include web log file analysis, and investigations into the link structure and growth of the Web." (Thelwall, Vaughan, Björneborn, 2004 [111])*

---

the patterns of authorship, publication and literature use, the relationships within scientific domains and research communities and to analyse the structure of specific fields." (*Wormell*, 1999 [120])

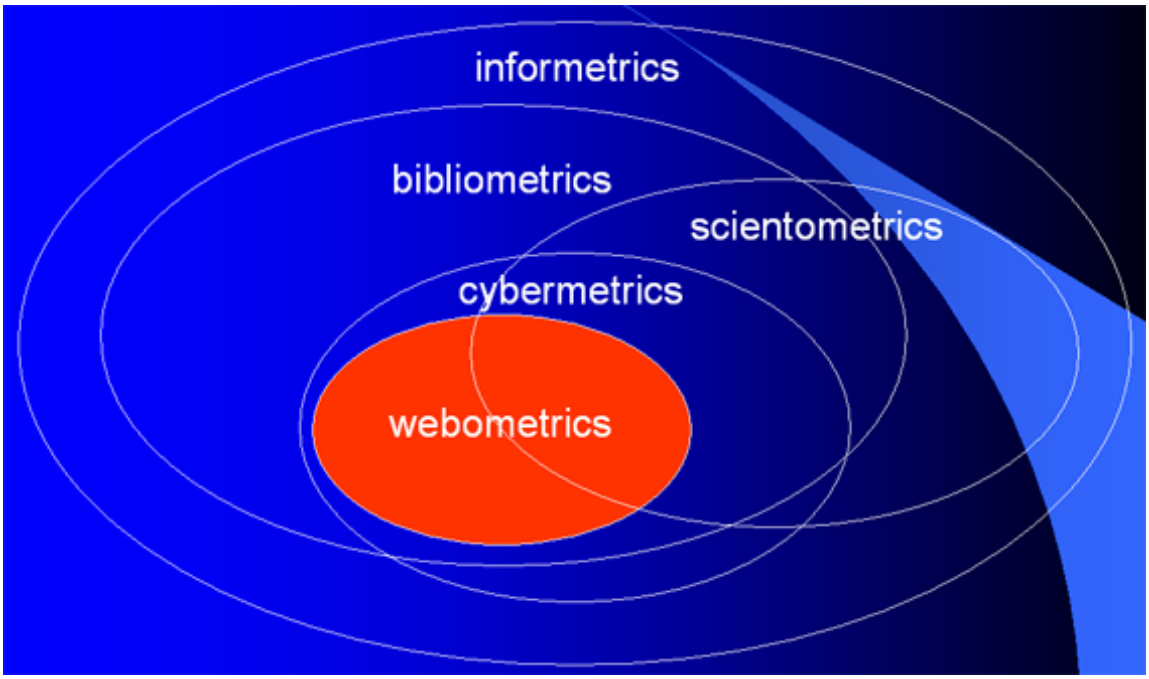


Abbildung 2-1: Einordnung des Forschungsbereich Webometrics nach *Björneborn*

Folgende Teilbereiche finden sich in der webometrischen Fachliteratur.

- Suchmaschinenanalyse (web technology analysis) (*Lawrence & Giles*, 1998 [58], *Bar-Ilan*, 1998 [6], *Snyder & Rosenbaum*, 1999 [85], *Rousseau*, 1999 [81], *Mettrop & Nieuwenhuysen*, 2001 [65], …),
- Inhaltsanalyse (web page contents) (*Cronin et al.*, 1999 [23], *Bar-Ilan*, 2001 [4]),
- Linkanalyse (link structures) (*Rousseau*, 1997 [82], *Ingwersen*, 1998 [46], *Cui*, 1999 [25], *Thelwall,* 2001-2003 [93, 96, 98, 102, 104, 112]),
- Informationsverhalten (user's information behavior, web usage analysis) (*Wolfram*, 2000 [80], *Jansen & Pooch*, 2001 [47], *Ross &*, *Iivonen*, 2001 [45], *Thelwall*, 2001 [105], *Cothey*, 2002 [22], *Hargittai*, 2002 [42]),
- Issue tracking (*Bar-Ilan*, 2000 [7]),
- Knowledge discovery (*Etzioni*, 1996 [31], *Brin & Page*, 1998 [13], *Kleinberg*, 1999 [49]).

Der Großteil der erwähnten webometrischen Linkanalysen bezieht sich auf Trefferlisten-analysen großer kommerzieller Internetsuchmaschinen (z.B. Altavista) [vgl. 82, 58, 55] und spezieller Forschungssuchmaschinen [97]. Verallgemeinernd lässt sich zu den webometrischen Untersuchungen sagen, dass die Ergebnisse häufig durch die Unverlässlichkeit des Internets bzw. World Wide Web[4] eingeschränkt werden [96, 85, 81, 65]. Die dynamische Struktur des Webs bzw. dessen „unsicheren" Eigenschaften wirken sich z.T. negativ auf die wissenschaftliche

Aussagefähigkeit der Ergebnisse aus. Jüngste „webometrische“ Untersuchungen [111, 99, 107, 106] konzentrieren sich zunehmend auf erweiterte Linkanalysen und beziehen externe Parameter in ihre Metriken mit ein.

> *“Webometrics is a new research field now passing through a necessary tentative and exploratory phase. (…) In the years to come, a challenge for researchers in webometrics will be to analyse and synthesise the findings and to further develop theories and methodologies in order to provide a better understanding of the complex topology, functionalities and potentials of the Web.” (Björneborn & Ingwersen, 2001, S.79 [10])*

2.3 Web Logfileanalyse

> *„With the web being such a universally popular medium, accounting forever more of people’s information seeking behaviour, and with every move a person makes on the web being routinely monitored, web logs offer a treasure trove of data. This data is breathtaking in its sheer volume, detail and potential ... Unfortunately the logs turn out to be good on volume and (certain) detail but bad at precision and attribution.” (Nicholas et al., 1999, S.263 [69])*

Web Logfileanalysen finden seit einigen Jahren, insbesondere im kommerziellen Webbereich, immer stärkere Verbreitung und Anwendung. Die Anwendung quantitativer Nutzungsmessung und die zugrundeliegenden nicht-reaktiven Messverfahren spielen heute bei kommerziellen und nichtkommerziellen Webinhalten aus unterschiedlichen Motivationen eine zunehmend wichtige Rolle. Im Mittelpunkt dieser Untersuchungen stehen die Daten, die ohne Wissen der Webuser (nicht-reaktiv) bei ihrer Webbenutzung aufgezeichnet werden. Diese Daten liefern Informationen über die Interaktionen der Webuser mit der Website und lassen darauf schließen wie das Web genutzt wird. Trotzdem haben Logfileanalysen individueller Websites in der informationswissenschaftlichen Forschung bislang nur am Rande eine Rolle gespielt (mit Ausnahme folgender Beiträge *Eschenfelder et al.,* 1997 [30], *Nicholas et al.*, 1999 [67], *Silverstein et al.*, 1999 [83], *Thelwall*, 2001 [105]). Intensive Forschungstätigkeit lässt sich seit einigen Jahren insbesondere in einzelnen Disziplinen (z.B. Web Usage Mining) der Informatik feststellen (siehe Abschnitt „Web Mining“ oben [2.1]).

Grund für die Zurückhaltung der Informationswissenschaftler beim Thema Logfileanalyse ist zum einen sicherlich der große Umfang der zu analysierenden Daten und der erschwerte Zugang[5], zum anderen die Aussagekraft (vgl. *Thelwall, Vaughan, Björneborn*, 2004 [111]) der Analyse einzelner Webserver sowie die Fehleranfälligkeit dieser Analysen. Vielfach wird aber der Wert der Daten, die sich in den Protokollen befinden schlichtweg unterschätzt. *Nicholas et al.* [69] weisen auf zwei wichtige Charakteristika hin, die im Zusammenhang von serverbasierten Web Loganalysen zu beachten und bei der Interpretation der Ergebnisse im Hinterkopf zu behalten sind. Zum einen sind die Einträge, die sich in den Logfiles befinden, Einträge, die von „virtual users“ stammen, also virtuellen

[4] Das Web ist ein dynamisches Netzwerk aus Informationsbestandteilen, dass von vielen verteilt agierenden Autoren („multi-agent constructed“) aus unterschiedlichsten Motiven und mit wenigen Regeln bzw. Standards zusammengesetzt wird.

[5] Zugang zu den Logdaten hat in der Regel nur der Webadministrator der entsprechenden Website. Die Logdaten unterliegen außerdem datenschutzrechtlichen Bestimmungen, die eine leichtfertige Weitergabe der Daten verhindern.

Benutzern oder anders ausgedrückt von vernetzten Computern. Die Folge ist, dass sich die Aktionen und damit die entstehenden Einträge dieser „virtual users“ in der Regel nicht direkt auf eine bestimmte Person beziehen lassen. Das liegt daran, dass im Logfile in der Regel zur Identifikation von Benutzern lediglich die Internetadresse (IP-Adresse) des Computers protokolliert wird (siehe dazu auch Erklärungen zum Logfile-Feld „HOST“ [3.1.1]). Der zweite Punkt, den *Nicholas et al.* ansprechen, zielt daraufhin ab, dass das World Wide Web die ursprüngliche Aufgabe hatte, Informationen (Dokumente) weltweit und verteilt zugänglich zu machen. Als Online Retrieval System war das WWW demnach nie konzipiert. Folglich haben die Logfiles als primäre Aufgabe die Zugriffe auf diese weitweit verteilten Dokumente aufzuzeichnen. Für weitergehende Interpretationen der Daten aus Logfiles wie z.B. die Identifikation von Suchstrategien oder die Extraktion von Navigationspfaden, etc., sind die Logfiles zunächst einmal nicht konzipiert. Die Daten in Logfiles bieten sowohl für den Website-Besitzer als auch für die informationswissenschaftliche Forschung eine Fülle von Informationen, die zur Zeit allerdings nur in Ansätzen genutzt werden. Heute werden Logfiles hauptsächlich für allgemeine Zugriffsstatistiken und websiteinterne Navigationsanalysen verwendet, eine weitreichendere Analyse und Interpretation der Zusammenhänge der erhobenen Daten in Logfiles bleibt allerdings aufgrund der Eindimensionalität der Betrachtung in den meisten Fällen aus. Insbesondere die Identifikation und Untersuchung von „neuen“ Besuchern stellt eine große Herausforderung dar und bietet weitere Potentiale zur Optimierung der Site. Aus dem Logfile einer Website lassen sich neben Angaben zum allgemeinen Benutzungs- bzw. Navigationsverhalten seiner Besucher weitergehende Informationen extrahieren, die tiefere Rückschlüsse (Muster) über Zugangswege (accessibility) und Sichtbarkeit (visibility) von Webinhalten bzw. einer Website möglich machen. *Nicholas et al.* beschreiben den Forschungsprozess im Umfeld der Logfile-Analysen folgendermaßen:

> *“The research, in fact, turned out to be the type of research where the journey itself proved to be more important than the destination; where the act of attempting to crack the code proved more important than the information produced as a result of having cracked the code. This was because, by the very act of cracking the code, you are questioning the web itself.” (Nicholas et al., 1999, S.263 [69])*

Web Logdaten liefern aufgrund der Detaillierung und dem Umfang der erhobenen Daten eine Reihe weiterer Potenziale. Insbesondere für die informationswissenschaftliche Forschung scheinen folgende Untersuchungsbereiche wertvolle Informationen liefern zu können (vgl. *Thelwall*, 2001 [105]).

- Untersuchung einer u.U. großen und heterogenen Besucherschaft (Internetnutzer weltweit),
- Untersuchung des Informationsverhalten (online information retrieval, web surfing behaviour) von Internetusern,
- Analyse des Informationsaufnahmeverhalten (kognitive Ergonomie) von Internetusern,
- Untersuchung der konkreten Nutzung von verschiedenen Webentitäten (user interaction),
- Hinweise zur Optimierung, Evaluation und Adaption von Webinhalten.

Web Logfileanalysen werden durch unterschiedliche Faktoren bzgl. ihrer Aussagekraft und Genauigkeit eingeschränkt. Das Hauptproblem stellen dabei die methodischen Probleme bei der Loganalyse dar. Methodisch lässt sich beispielsweise nicht in Erfahrung bringen, welche Benutzer-

Transaktionen im Logfile nicht aufgezeichnet werden, weil sie zuvor durch Caching Mechanismen (Browser-Cache oder Proxy-Cache) herausgefiltert werden und daher nicht bis zum Webserver gelangen (vgl. *Cooley et al.*, 1999 [21]). Firewalls spielen bei Loganalysen unter Umständen eine problematische Rolle, weil sie die IP-Adresse der einzelnen Computer im Firmennetz auf eine einheitliche anonyme IP-Adresse verkürzen. Einzelne Websitebesucher lassen sich somit nicht mehr unterscheiden. Die Proxy- sowie die Firewall-Problematik können sich beim Einsatz beider Techniken überlagern (vgl. *Fühles-Ubach*, 2001 [35]). Die Folge sind ungenaue und verzerrte Aussagen bzgl. der Besucherzahlen.

Eine weitere Einschränkung bei der Loganalyse besteht darin, dass die Untersuchung der Logdaten eines singulären Webservers selbstverständlich auch nur Aussagen über diesen einzelnen Webserver zulassen. Allgemeingültige Aussagen bzgl. des Informationsverhaltens im Web werden aufgrund der Volatilität der Daten sowie der Begrenztheit der Stichprobe eines Webservers außerordentlich schwierig. *Nicholas et al.* fordern deshalb in bezug auf Loganalysen eine überlegte Herangehensweise (vgl. folgendes Zitat).

> *"The trouble, of course, is that there is no single measure of consumption and each measure has to be taken with a large dose of statistical salt." … " - Nobody logs off on the web (you have to allow for a suitable time - say, 30 minutes - interval and then assume they are no longer there). - People can be logged on to the web but are not using it (having a coffee break, for instance). - The fact that a page was downloaded does not mean that anyone actually wanted it (the person was on the way to another page or was simply provided with an irrelevant link). - It is almost impossible to relate a transaction to an individual, to a human (numerous people could use the same IP address - the only information fingerprint left behind, and the same person could be using more than one IP address)."*
> *Nicholas et al., 1999, S. 265 [69]*

3 Untersuchung

Im Mittelpunkt dieser explorativen Untersuchung steht die Analyse und Bezifferung von Zugangspfaden anhand quantitativer nicht-reaktiver Nutzungsmessung über eine neue Webmetrik (WEF). Besonderes Augenmerk soll hierbei auf die externen Linkstrukturen (externe Links und Suchmaschinen) gelegt werden, die nachweislich zu einem Websitezugang (Websitenutzung anhand des Logs) geführt haben.

3.1 Grundlagen

Nachfolgend werden die einzelnen Grundlagen und Bestandteile der Untersuchung dargestellt.

3.1.1 Logdaten

Web Logfiles sind strukturiert aufgebaute serverplattformabhängige Protokolldateien, die aus einfachem ASCII bestehen. Der Webserver bzw. die Webserversoftware schreibt dazu für jede HTTP-Transaktion (erfolgreiche und auch gescheiterte) typischerweise eine Zeile mit Informationen, die die Transaktion betreffen. Eine Zeile im Logfile repräsentiert somit eine Anfrage oder „hit“ auf dem Webserver. Der Inhalt der Zeilen basiert auf einer eindeutigen Feldstruktur. Die Anordnung der protokollierten Felder kann allerdings von Webserver zu Webserver variieren. In vielen Fällen lässt sich der Umfang der mitprotokollierten Felder konfigurieren. Protokolliert der Webserver zusätzlich die Felder „REFERER“ und „AGENT“, befindet sich das Logfile im sogenannten „extended format“ (vgl. Beispiel unten). Das untere Beispiel zeigt eine einzige Zeile im Combined Log Format des Apache Webservers. Die einzelnen Felder des Logfiles werden durch das Tab oder Space-Zeichen voneinander getrennt.

```
120.0.0.7 - - [06/Jan/2002:11:14:34 +0100] "GET
/~fern/fernstudium/magister/magister.html HTTP/1.1" 200 12872
"http://www.google.de/search?q=fernstudium&start=20&sa=N" "Mozilla/4.0
(compatible; MSIE 5.5; Windows NT 5.0)"
```

Listing 1: Zeile eines Logfiles im Format „NCSA Combined Log Format“

Im folgenden werden die Einträge der einzelnen Felder anhand der obigen Beispielszeile kurz beschrieben (vgl. Listing 1 oben).

Einträge des Logs vgl. Listing 1 (oben)	**Feldname**	**Beschreibung**
120.0.0.7	host[6]	Im Feld „host“ protokolliert der Webserver die Adresse des Computers, der den HTTP-Request ausführt. Aufgezeichnet wird die numerische IP-Adresse bzw. wenn möglich die Auflösung der IP-Adresse als qualifizierten Domain Name durch den Domain Name Service (DNS).
–	ident	Dieses Feld ist dafür vorgesehen den Benutzer zu identifizieren. Normalerweise wird das Feld leer gelassen und mit einem Bindestrich gefüllt (siehe Beispiel oben)
–	authuser	Dieses Feld ist dafür vorgesehen bestimmte Zugriffe von authentifizierten Benutzern zu kennzeichnen, die z.B. auf ein passwortgeschütztes Dokument zugreifen. Aufgezeichnet wird die userid des authentifizierten Benutzers.
[06/Jan/2002:11:14:34 +0100]	date[7]	Das Feld „date“ protokolliert das Datum und die genaue Zeit für jeden hit im Logfile.
"GET /~fern/magister.html HTTP/1.1"	request	Das Feld „request“ beinhaltet den HTTP-Request des Clients bzw. des Benutzers. Der HTTP-Request wird in Hochkommata gesetzt und besteht aus folgenden Bestandteilen: 1) Kürzel für die HTTP-Methode (GET|POST|HEAD), 2) URL des zugegriffenen Dokuments und 3) dem Namen und Version des Protokoll. Dieses Feld identifiziert eindeutig die zugegriffene Ressource.
200	status	Das Feld „status“ verzeichnet den Status des entsprechenden Requests. Der dreistellige Code gibt an zu welcher Klasse von Requests der Zugriff gehört. Es gibt vier Klassen von Requests: 1. erfolgreiche „success“ (200er Code), 2. weitergeleitete „redirect“ (300er Code), 3. fehlergeschlagene „failure“ (400er Code) und 4. Serverfehler „server error“ (500er Code).
12872	bytes	Dieses Feld gibt die übertragene Datenmenge des HTTP-Requests an. Wenn keine Daten übertragen wurden steht

[6] Als „Host“ im eigentlichen Sinne kann ein beliebiges Computersystem angesehen werden, dass über eine IP-Adresse mit dem Internet verbunden ist. Die dynamische Vergabe von IP-Adressen, wie sie das Dynamic Host Configuration Protocol (DHCP) unterstützt, wird mittlerweile von vielen Firmen und Internet Service Providern (ISP) intensiv (Beispiel AOL) angewendet. Die Folge ist, dass ein bestimmter Nutzer an unterschiedlichen Tagen mit verschiedenen IP-Adressen im Logfile auftaucht und i.d.R. nicht anhand seiner IP-Adresse identifiziert werden kann.

[7] Das Feld date folgt in der Regel dem Format [Tag/Monat/Jahr:Stunde:Minute:Sekunde Zone]. Tag = 2 Ziffern, Monat = 3 Buchstaben, Jahr = 4 Ziffern, Stunde = 2 Ziffern, Minute = 2 Ziffern, Sekunde = 2 Ziffern, Zone = + | - 4 Ziffern (gemessen wird hier die Abweichung von der Greenwich Mean Time (GMT) in Stunden)

		an dieser Stelle ein Bindestrich oder die Ziffer 0.
"http://www.google.de/search?q=fernstudium&start=20&sa=N"	referer[8]	Das Feld „referer“ ist Bestandteil des „extended log format“. Die URL des „referer“-Feldes zeigt die URL an, die der Benutzer zuletzt angefordert hat. Dies gilt sowohl für URLs (z.B. Seiten), die sich innerhalb der Website befinden, als auch für URLs, die sich außerhalb der Website (z.B. auf Seiten fremden Webservern oder Suchmaschinen) befinden. Aus dem „referer“-Feld lässt sich somit die Navigation innerhalb der Website und auf diese identifizieren. Insbesondere wird aus diesem Feld deutlich über welche externen URLs der Request auf die Website erfolgt.
"Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0)"	useragent	Das Feld „user agent“ ist ebenfalls Bestandteil des „extended log format“. Hier werden Informationen zum Browsertyp (Name und Version) und Angaben zum Betriebssystem des zugreifenden Computers gespeichert. Die Roboter der Suchmaschinen tragen in dieses Feld in der Regel ihren Namen ein (z.B. Googlebot).

Tabelle 3-1: Log-Zeile im Format „NCSA Combined Log Format“

3.1.2 Begriffe

Der folgende Abschnitt erwähnt allgemeine Begriffe, die im Zusammenhang mit dem Thema Logfileanalyse häufig fallen und zum Teil unterschiedlich angewendet werden (siehe dazu *Web Characterization Terminology and Definitions Sheet*, 1999 [117]). Außerdem sollen weitere Begrifflichkeiten erläutert werden, die für das Verständnis der Untersuchung in den folgenden Kapiteln elementare Bedeutung haben.

Die Maßeinheit „hit“ gibt die Nutzung einer Website nach der Anzahl der HTTP-Requests an. Als „hit“ würde demnach ohne Unterschied jede einzelne Zeile des Logfiles gezählt werden. *Nicholas et al.* kritisieren die Maßeinheit „hit“ recht deutlich (siehe Zitat unten). Ihre Kritik bezieht sich insbesondere darauf, dass anspruchsvolle Web Loganalysen Unterschiede bei den Daten machen muss. Wenn die Benutzung einer Website analysiert werden soll, ist „hit“ folglich nicht gleich „hit“. Ein „hit“ im Logfile, der auf ein grafisches Element zurückgeht, das automatisch bei einem Zugriff auf eine Webseite mitgeladen wird, lässt sich eben nicht mit dem „hit“ der eigentlich angeforderten Seite gleichsetzen. Genauso verursachen die Robots oder Crawler der Suchmaschinen „hits“ im Logfile, die keine „Besucher-Hits“ im eigentlichen Sinne sind. Die Maßeinheit „hit“ erscheint somit im Zusammenhang von detaillierten Web Loganalysen als zu ungenau und allgemein.

[8] Das „Referer-Feld“ kann durch spezifische Benutzereinstellungen geblockt werden. Der Webserver schreibt in diesen Fällen einen Vermerk in das Log.

“Note the way the word ‘hits’ is bandied about by all and sundry. The fact that the ‘hit’ - one line in a web log - is the crudest and most misleading measure of them all is illustrative of the problem that we face.“ (Nicholas et al., 1999, S.265 [69])

Die Maßeinheit „Page View“, auch „Page Impression“ genannt, bezieht sich auf die Anzahl der zugegriffenen Webseiten. Berücksichtigt werden hier bei der Zählung lediglich die vom Besucher explizit angeforderten Webseiten. Zusätzliche Seiten wie zum Beispiel „Pop-ups“ oder die oben bereits angesprochenen automatisch mitgelieferten Nicht-HTML-Elemente werden bei dieser Methode nicht berücksichtigt. Diese Messmethode zählt die Intensität bzw. die Länge der Websitebesuche und gibt sicherlich ein genaueres Bild der Benutzung. Unkritisch kann allerdings auch diese Maßeinheit nach *Nicholas et al.* nicht betrachtet werden [vgl. 67]. Ein häufig genannter Indikator für die Intensität der Benutzung einer Website, und somit der Popularität und dem Erfolg der Site, ist die Anzahl der Websitebesuche, auch „visit“ oder „session“ genannt. Ein Websitebesuch („visit“) setzt sich aus Transaktionen einzelner Besucher (Visitors) zusammen. Problematisch bei der Bewertung von Visit-Messungen ist die Tatsache, dass bei der Analyse zeitliche Festlegungen getroffen werden müssen. Zum Beispiel wann beginnt ein Websitebesuch und wann endet er. Im Gegensatz zu kommerziellen Datenbankrecherchen bei denen sich der Recherchierende am Host an- und abmeldet, gibt es ein „Logoff“ bei Websitesessions auf öffentlichen Websites nicht. Das bedeutet, dass bei Web Loganalysen für die Bestimmung der „visits“ ein sogenanntes Timeout gesetzt wird. Die Untersuchung von *Oldenburg* zeigt, dass die Anzahl der „visits“ bzw. „sessions“ durch die Länge des gewählten Timeouts bestimmt wird (vgl. *Oldenburg*, 2003 [71]). *Nicholas et al.* weisen daraufhin, dass Probleme bzgl. der Bestimmung der visits in der Konsistenz der vorliegenden Logdaten und der Natur des Webs liegt.

„Visits are in fact an old friend, the search session, which is denoted by someone or something arriving at the site, searching and then departing the site; ironically, as we shall learn, something that is hard to determine in respect to the web.” (Nicholas et al., 1999, S. 146 [67])

Weitere, zumeist kommerziell orientierte Webmetriken wie zum Beispiel „Time Online“, sind ebenfalls in Gebrauch. Die Begriffe, die im Zusammenhang mit dieser Untersuchung verwendet und entwickelt werden (z.B. Entry, Entry Page), finden sich im folgenden Abschnitt. Nachfolgend werden die beiden zentralen Bestandteile der Untersuchung dargestellt.

3.1.3 Entries und Einstiegsseiten

Die Grundlage dieser Untersuchung bildet die Maßeinheit „Entry“. Der Begriff Entry existiert als Maßeinheit bei Loganalysen in dem Sinne nicht, hat aber eine starke Verwandtschaft zu dem geläufigen Begriff „Visit“, der in vielen Loganalysen angewendet wird (*Tauscher & Greenberg*, 1997 [92], *Fühles-Ubach*, 2001 [35], *Cothey*, 2002 [22], *Nicholas et al.*, 1999, 2003 [67, 69, 68]) und in

verschiedenen Loganalyzern[9] implementiert ist (z.B. *WebTrends*, *Analog*). Entry steht hier für den Zugriff auf die erste Webseite (Entry Page), die bei einem Websitebesuch (vgl. Visit) durch den Besucher angefordert wird. Alle Folgezugriffe eines „visits" auf weitere Webseiten werden in dieser Untersuchung als Navigationszugriffe bezeichnet und werden für die spätere Analyse nicht berücksichtigt. Im Unterschied zu „visit" besitzt die Maßeinheit „Entry" keine zeitlich einschränkende Komponente, wie sie bei der Sessionsidentifikation[10] im Web Mining eingesetzt wird (vgl. *Cooley et al.*, 1999 [21], *Fu et al.* [34]). Entry zählt somit ohne Timeout alle Zugriffe auf Startseiten (Entry Pages) einer Website.

Genaugenommen greift jeder Benutzer, der auf eine Website gelangt, auf ein konkretes Webdokument zu, das über eine URL eindeutig identifiziert wird. Diese erste aufgerufene Webseite (URL) bildet somit immer den Einstieg (Entry oder auch Entry Point, vgl. *Pirolli et al*., 1996 [76]) eines Websitebesuchs und entscheidet unter Umständen über die Länge des Websitebesuchs bzw. über Besuche zu einem späteren Zeitpunkt. Bei kleineren Websites mit wenigen Seiten und Unterstrukturen ist die genaue Kenntnis der Startseiten (Entry Points) der Websitebesuche vergleichsweise unproblematisch, da sich der Besucher aufgrund der Übersichtlichkeit des Webangebots in der Regel leicht zurechtfindet. Umfangreiche Websites, beispielsweise aus dem universitären Bereich aber auch großer Unternehmen, mit vielen tausend Webseiten und verzweigten Unterstrukturen bzw. verschiedenen Unterwebsites, zeichnet im Gegensatz zu kleineren Sites, ein sehr viel komplexeres und vielfältigeres Zugangsverhalten seiner Besucher aus. *Thelwall et al.* (2002) liefern einen Überblick über die durchschnittlichen Größenordnungen von europäischen Universitätswebsites bzgl. der Anzahl der Webseiten [93].

3.1.4 Navigationsarten im Web

Die Navigation im Web spielt seit Beginn der Internetforschung eine große Rolle. Die Hypertextualität des WWW, die es ermöglicht einzelne Internetressourcen manuell oder automatisch miteinander zu verknüpfen, liefert dem Internetnutzer verschiedene Wege um auf Internetseiten zu zugreifen. Methodisch lassen sich grundsätzlich drei verschiedene Wege bzw. Arten unterscheiden, wie Besucher auf eine Webseite bzw. Website navigieren. Im Folgenden sollen die drei alternativen Navigationsarten voneinander abgegrenzt werden (vgl. *Sullivan*, 2002 [91], *Cothey*, 2002 [22], *Oldenburg*, 2003 [71]).

[9] Ein Loganalyzer ist eine Software, die Logdaten auswertet und standardisierte Statistiken des Webuse einer Website ausgibt.

[10] Bei der Sessionidentifikation werden in der Regel Timeouts gesetzt, die es ermöglichen die Zugriffe (Pageimpressions) im Log einzelnen Besuchen (Visits) zuzuordnen. Generiert ein User beispielsweise über 30 Minuten (Timeout) keine Transaktionen mehr im Log, werden die bisherigen Transaktionen zu einem Visit zusammengefasst. Oldenburg [71] zeigt in ihrer Magisterarbeit, dass eine Verkürzung der Timeout-Zeit zwangsläufig eine Erhöhung der Visitzahlen zur Folge hat.

- Navigationsart „Direkt“: Direktes Aufrufen einer Seite

Diese Kategorie beinhaltet das direkte Eingeben einer konkreten URL in die Adresszeile des Browsers bzw. das Aufrufen einer Seite aus einer Bookmark- oder Verlaufsliste[11]. Für diese Kategorie von Websitezugriffen (Entries) ist anzunehmen, dass der Benutzer bewusst und direkt auf eine Webressource zugreift und nicht zufällig eine Seite aufruft. Interessante Hinweise im Zusammenhang mit den „direkten“ Zugriffen auf Webseiten bieten die Untersuchungen zu Wiederbesuchsmustern (revisitation patterns) auf Websites von *Tauscher & Greenberg* (1997, [92]) sowie die Folgeuntersuchung von *Cockburn und McKenzie* aus dem Jahr 2000 [19]. Diese beiden Untersuchungen kommen beide zu sehr hohen Wiederbesuchswerten (1997 zu 61% und 2000 zu 81%). *Gottlieb und Dilevko* (2001, [40]) untersuchen in einer empirischen Nutzerstudien die Klassifikation und Organisation von URLs in elektronischen Bookmarksystemen. Über die Motivation eine Seite zu einer Bookmarkliste hinzuzufügen und unter Umständen regelmäßig wieder zu besuchen, kann an dieser Stelle nur gemutmaßt werden (vgl. *Nicholas et al.*, [67]). Sicherlich spielen moderne Browserfunktionalitäten wie Verlaufsfunktionen oder die Autovervollständigungs-mechanismen bei der Adresseingabe eine große Rolle. Trivialerweise kann außerdem angenommen werden, dass URLs, die sich leicht einprägen oder sehr kurz sind (z.B. Homepages und Startseiten), tendenziell häufiger „direkt“ aufgerufen werden als Webseiten mit sehr langen und komplizierten URLs.
Diese Kategorie von Websitezugriffen bzw. Entries wird im folgenden als „Direkt“ oder verkürzt mit „d_“ bezeichnet.

- Navigationsart „Suchmaschine“: Verfolgen eines Links, der durch eine Suchanfrage (Query) mittels Suchmaschine generiert wurde

Die Kategorie „Suchmaschine“ beinhaltet alle Seitenzugriffe (Entries), die über eine kommerzielle (z.B. google.com) oder nicht-kommerzielle Suchmaschine vermittelt wurden. Hier gilt zu berücksichtigen, dass die Ergebnisse von Suchmaschinen einem stetigen Wandel unterliegen [vgl. 6, 65]. Bei einer Suchmaschinensuche handelt es sich i.d.R. um einen interaktiven Prozess, der aus folgenden Schritten besteht. Als erstes ruft der Benutzer die Homepage der Suchmaschine auf und gibt eine Suchanfrage (Query) in ein Suchformular ein. Diese Query, die aus einem oder mehreren Suchbegriffen (Keywords) bestehen kann, schickt der Benutzer anschließend an die Suchmaschine. Die Suchmaschine interpretiert die Query und liefert daraufhin ein geranktes Suchergebnis zurück, dass aus einer Ergebnisliste mit Links besteht (vgl. *Brin & Page*, 1998 [13]). Aus dieser Ergebnisliste, die von den meisten Suchmaschinen mit maximal 1000 Treffern ausgegeben wird und aus bis zu 100 Trefferseiten besteht (*Mayr*, 2002 [62]), kann der Benutzer per Klick ihn interessierende Treffer (Webseiten) aufrufen. Trivialerweise ist anzunehmen, dass der Großteil der Internetnutzer die Treffer (Links), die die Suchmaschine als Ergebnis auf seine Anfrage präsentiert, nicht kennt. Grundsätzlich ließe sich auch gegenteiliges Verhalten denken. Beispielsweise benutzt der Benutzer immer die gleiche Suchanfrage, um zu einer ihm bekannten Seite zu gelangen. Im allgemeinen kann aber

angenommen werden, dass es sich bei Besuchern einer Website, die über eine Suchmaschinen-Query auf die Site gelangen, um Besucher handelt, die entweder zum ersten Mal oder zumindest nicht regelmäßig auf die Website zugreifen (vgl. *Thelwall*, 2001 [105]).
Diese Kategorie wird im folgenden als „Suchmaschine“ bzw. verkürzt mit „s_“ bezeichnet.

- Navigationsart „Referenz“[12]: Verfolgen eines externen Links (Backlinks)

Die Kategorie „Referenz“ beinhaltet alle Entries, die über externe Links erfolgt sind. Die Navigationsart „Referenz“ beschreibt die Navigation eines Internetnutzers von einer Website zu einer anderen Website über definierte Linkbeziehungen. Die Möglichkeit über interne Links innerhalb einer Website zu navigieren, sich sozusagen von einer Webseite innerhalb einer Website zur anderen zu klicken (Browsing), wird in dieser Untersuchung nicht betrachtet. Externe Links werden in der Fachdiskussion auch als „Sitations“ oder „Backlinks“ bezeichnet (vgl. *Rousseau*, 1997 [82], *Brin & Page*, 1998 [13], *Thelwall*, 2001 [105]). Es kann davon ausgegangen werden, dass Referenzen (Backlinks), die auf eine andere Webseite verweisen und meist manuell erstellt werden, bewusst gesetzt werden. Die Motive andere Websites bzw. Webseiten mit dem eigenen Webangebot zu verlinken sind an verschiedenen Stellen untersucht worden (*Kim*, 2000 [48], *Thelwall & Harries*, 2002 [108], *Cronin et al.*, 1998 [23], *Haas & Grams*, 2000 [41], *Thelwall*, 2003 [107]). Trivialerweise ist anzunehmen, dass ein Großteil der Besucher, die über die Navigationsart „Referenz“ (Backlink) auf eine Website navigieren, anhand der Linkbeschriftung (Kontext des Links) zumindest ahnt oder aus vorherigen Besuchen bereits weiß, was sich hinter dem Link verbirgt. Diese Art der Navigation hebt sich von der Suchmaschinen-Navigation in den Sinne ab, dass der Backlink-Navigation eine größere Zielorientierung unterstellt werden kann. Da sich interne und externe Links optisch nicht zu unterscheiden sind, ist anzunehmen, dass viele Nutzer nicht oder erst nach Aktivierung eines Backlinks merken, dass sie eine Website verlassen haben und gerade eine neue Website besuchen.
Diese Kategorie wird im folgenden als „Referenz“ oder verkürzt mit „r_“ bezeichnet.

3.1.5 Fragestellungen im Zusammenhang von Web Loganalysen

Bevor die zugrundeliegende Idee und Konzeption der Arbeit vorgestellt wird, werden allgemeine Fragestellungen skizziert, die für den Zugang und das Verständnis der Arbeit wichtig erscheinen. Grundeinschränkung heutiger serverseitiger Web Loganalysen ist, dass aus Gründen, wie z.B. der enorm großen Datenmengen, aber auch wegen datenschutz-rechtlicher Einschränkungen, bislang hauptsächlich Untersuchungen singulärer Websites möglich sind. Die Untersuchung einer größeren Anzahl von Websites, wie sie in der Linkforschung (vgl. Webometrics, z.B. *Thelwall*, 2002 [104]) seit

[11] Vgl. *Oldenburg*, 2003 [71]
[12] Hinweis: Die englische Bezeichnung für Literaturangaben „references“ steht in keinem Zusammenhang zu der hier beschriebenen Navigations- bzw. Zugangsart „Referenz“ einer Website. Wenn in dieser Arbeit von Referenzen gesprochen wird, sind damit Hyperreferenzen bzw. Hyperlinks gemeint. Diese Hyperreferenzen sind Webressourcen, die Links auf die untersuchte Website enthalten, sogenannte „Sitations“ oder „Backlinks“.

Jahren angewendet wird, stellt auf Seiten des Web Minings (Web Usage Mining) bislang zu hohe technische, organisatorische und rechtliche Anforderungen an die Analysewerkzeuge.
Folgende Fragestellung lassen im Zusammenhang mit dieser Arbeit aufzählen:

- Welche Potenziale verbergen sich in den Logdaten freizugänglicher Webserver im Hinblick auf detaillierte Messung der Nutzung der verschiedenen Navigationsarten?
- Welche spezifischen Erkenntnisse bzgl. Sichtbarkeit, Zugänglichkeit und Verlinkung von Webinhalten lassen sich aus Web Logdaten gewinnen?
- Lassen sich Gesetzmäßigkeiten für das Informationsverhalten im Web in den Web Logdaten finden und beschreiben?
- Kann über die Feststellung der Nutzungshäufigkeiten beliebiger Websiteinhalte auf die Bedeutung dieser Entitäten geschlossen werden?

Nachdem in den vorhergehenden Abschnitten die Grundlagen der Untersuchung dargestellt wurden, entwickelt der folgende Abschnitt das zentrale Konzept dieser Untersuchung.

3.2 Web Entry Faktoren

Die folgende Untersuchung stellt neue Gesichtspunkte und Analysemöglichkeiten für Web Logdaten[13] vor. Im Mittelpunkt steht zum einen die Analyse der serverseitig erhobenen Daten der Navigationsarten „Suchmaschine“, „Direkt“ und „Referenz“, die im oberen Abschnitt beschrieben wurden. Zum anderen entwickelt und testet die Arbeit eine Metrik, die *Web Entry Faktoren* (*WEF*), die Intensitäten des Gebrauchs der Navigationsarten bis zu einem seitengenauen Niveau berechnet. Zentraler Untersuchungsgegenstand sind Start-, bzw. Einstiegsseiten (Entry Pages) eines singulären akademischen Universitäts-Webservers.

3.2.1 Vorbemerkung

Das entwickelte Konzept *Web Entry Faktoren* (WEF) erinnert namentlich an die 1998 veröffentlichte und breit rezipierte Fallstudie von *Peter Ingwersen* „The calculation of Web Impact Factors“ [46]. Diese Untersuchung hatte das Ziel das bibliometrische Instrument Zitationsanalyse (citation analysis) auf das Gebiet World Wide Web anzuwenden und anhand von extensiven Linkzählungen Impact Faktoren[14] für Websites (WIF) zu berechnen[15]. *Ingwersen* misst den Impact[16] einer Website A an der

[13] Voraussetzung für die Untersuchung sind Logdaten, die sich im Combined oder Extented Logformat (siehe REFERER-Feld) befinden.
[14] Das ISI gibt regelmäßig Impact Faktoren für wissenschaftliche Zeitschriften heraus. “Impact Factors (IF) for scientific journals are published by ISI (the Institute of Scientfic Information) in the annual Journal Citation Reports and commonly used for evaluation purposes, … Generalised IFs for journals covered by ISI services can be reproduced accurately online, including external- and self-citations, by applying the publicly available citation indexes.” (*Ingwersen*, 1998, S. 236 [43])

Anzahl der externen Webseiten (mit extern sind Webseiten gemeint, die sich nicht innerhalb der Website A befinden), die einen Link auf eine Webseite der Website A enthalten. Der WIF einer Website A wird berechnet, indem die Anzahl der externen Links auf die Website A durch die Anzahl der einzelnen Seiten der Website A dividiert wird. Ergebnis der WIF-Berechnung (Linkaggregation) ist ein Maß für den durchschnittlichen Web Impact einer Webseite der Website A.

"Ingwersen (1998) introduced the Web Impact Factor (WIF) for this, in one version measuring the 'impact' of a Web space through the ratio of external pages containing a link to any page in the target space divided by the number of pages in that target space, which could be a Web site or even a national or international domain." (Thelwall, 2002 [100])

Mike Thelwall hat weitere webbezogene Linkmetriken (Web Use Factor (WUF) und Web Connectivity Factor (WCF)) für akademische (Universitäts-)Websites vorgestellt (*Thelwall*, 2003, [106]), die sich an *Ingwersens* WIF orientieren bzw. ihn erweitern. Der „Web Use Factor" (WUF) misst beispielsweise die Verlinkungen, die von einer Universitäts-Website zu einer zweiten Website bestehen. Somit werden Aussagen über das Ausmaß der Verlinkungen zwischen einzelnen Websites möglich (Website Interlinking). Diese Verlinkungen können nach *Thelwall* Hinweise über die Intensität des Webuse einer Institution bzw. seiner Angehörigen geben. *Thelwall* dividiert am Beispiel des WUF die Anzahl der Links durch die Anzahl der Angehörigen einer Institution.

„In this context it would be expected that a university with academics that were effectively using the Web and active in informal scholarly communication or collaboration of various sorts would generate a reasonable WUF score. A low value could not be directly tied to any one deficiency, however, since the variety of uses of Web links is so great." (Thelwall, 2003 [106])

Während sich die oben dargestellten Linkmetriken hauptsächlich auf die Zählung und Aggregation von externen Linkstrukturen akademischer Websites beziehen, basiert das hier zu entwickelnde Konzept „Web Entry Faktoren" (WEF) auf der Messung der konkreten Nutzung von Linkstrukturen anhand der Nutzungseinträge im Logfile. Die Messung der Intensität der Nutzungsvorgänge der externen Linkstrukturen einer Website über die Logmetrik WEF liefert daher einen sehr praktischen und realen Nutzungswert. Es lässt sich zusammenfassen, dass Linkmetriken (z.B. WIF) im allgemeinen Aussagen über die Existenz von Linkstrukturen auf einem relativ abstrakten Niveau möglich machen, während die Logmetrik WEF Aussagen über die reale Nutzung von Links aus der Perspektive eines Knotens in der Linkstruktur des Webs ermöglicht.

---

[15] *"We have tested whether similar online estimations of selected national, sector, and institutional impact factors for the World Wide Web (Web-IFs) are feasible and reliable."* (*Ingwersen*, 1998, S. 236 [43])

[16] Englischer Begriff für Tragweite bzw. Einfluss.

3.2.2 Idee

Die Hauptidee dieser Untersuchung besteht darin, anhand einfacher Zugriffsdaten, grundlegende Hinweise über das Informationsverhalten (Onlineverhalten) von Webusern zu erhalten. Folglich spielen die verschiedenen Navigationsarten über die Webuser auf die akademische Website navigieren, eine bedeutsame Rolle bei der Loganalyse. Die Untersuchung der verschiedenen Navigationsarten anhand des Logs liefert detaillierte Informationen über die Nutzung vorhandener externer Linkstrukturen einer Site (vgl. aktuelle Forschungstätigkeit im Bereich webometrischer Linkanalyse). Idee war es, insbesondere die externen Links in ihrer Bedeutung für eine Website einschätzen zu können. Da die Messung der Nutzung von externen Links auf die gesamte Website gesehen, sehr abstrakte Informationen generiert, ist es sinnvoll die Nutzungsmessung für kleinere Entitäten der Website vorzunehmen. Als Entitätsklassen bieten sich Verzeichnisse und insbesondere einzelne Webseiten an (vgl. *Thelwall*, 2002 [100, 102]). Weitere grundlegende Idee dieser Arbeit ist es, ein einfaches Verfahren zu entwickeln, mit dem einzelne Webseiten aufgrund der Intensität und Eigenschaft ihrer Nutzung (Entries) innerhalb der Website beschrieben und eingeordnet werden können. Weiterhin liefern die entstandenen detaillierten seitenbezogenen Nutzungsinformationen eine Grundlage für Optimierungsbemühungen bzw. Evaluierungs-verfahren bzgl. der Website.
Ziel der Untersuchung ist es, insbesondere Ansatzpunkte in den Forschungsgebieten Webometrics (Cybermetrics) und Web Mining zu suchen und wenn möglich für das Loganalyse-Verfahren zu nutzen. Weiterhin versucht die Untersuchung am Beispiel einer beliebigen Website allgemeine Gesetzmäßigkeiten der Nutzung und der Zugänglichkeit (accessibility) von akademischen Websites zu beschreiben. Die Exploration und Erweiterung des Bereichs Web Loganalyse, insbesondere vor dem Hintergrund heutiger informations-wissenschaftlicher Forschung, hat bei der Konzeption der Arbeit ein gewichtige Rolle gespielt.

3.2.3 Aufbau der Web Entry Faktoren

Die Web Entry Faktoren (WEF) setzen sich aus folgenden Bestandteilen zusammen. Grundlegendes Konzept der WEF liegt in der Unterscheidung der drei abgrenzbaren Navigationsarten „Suchmaschine“, „Direkt“ und „Referenz“ für jede untersuchte Einheit (Entität) einer Website. Für jede einzelne Webentität (Site, Directory, Page) der Untersuchungsstichprobe werden die Entry-Zugriffe der drei verschiedenen Navigationsarten unterschieden und einzeln gezählt. Jede untersuchte Entität erhält folglich drei Einstiegs- bzw. Entry-Werte, die sich aus den absoluten Zahlen der aufsummierten Entries (siehe Entries total) für jede URL und Navigationsart ergeben (siehe Tabelle 3-2).
Beispiel (vgl. Tabelle 3-2): Die 10.000 Entries der Beispielsseite „beispiel.htm“ setzen sich aus den Entry-Werten der drei Navigationsarten „Suchmaschine“ (2.000 Entries), „Direkt“ (7.000 Entries) und „Referenz“ (1.000 Entries) zusammen. Aus den drei Entry-Werten pro Entität und dem Wert der Gesamtzugriffe (Entries total) lassen sich drei Verhältnis-Werte bilden. Diese Verhältnis-Werte stehen für die Web Entry Faktoren (siehe wef s, wef d, wef r, Werte in der Klammer).

Webentität	URL	entries s (wef s)	entries d (wef d)	entries r (wef r)	entries total
Page	/beispiel.htm	2.000 (0,20)	7.000 (0,70)	1.000 (0,10)	10.000
Directory	/inf/studium/*	-	-	-	-
Site	/*	-	-	-	-

Tabelle 3-2: Entry-Werte und WEF-Faktoren für unterschiedliche Webentitäten

Die drei Verhältnis-Werte wef s, wef d, und wef r werden folgendermaßen berechnet.

- „Suchmaschinen-WEF“ (WEF der Suchmaschinen-Entries)
  wef s = entries s / entries total
- „Direkt-WEF“ (WEF der direkten Entries)
  wef d = entries d / entries total
- „Referenz-WEF“ (WEF der Referenz-Entries)
  wef r = entries r / entries total

Die Summe der einzelnen WEF’s pro Entität ergibt immer 1 bzw. 100 %.

wef s + wef d + wef r = 1

Analog zu der Bestimmung der Webseiten WEF’s (Page) lassen sich wie bereits angedeutet auch WEF-Werte für andere Website-Entitäten (Directory und Site) berechnen (vgl. „Advanced Web Document Models”, *Thelwall & Harries*, 2003 [108]).

- *Page WEF*: jede einzelne URL bzw. Webseite (HTML-Datei) kann über die drei WEF’s auf seine Zugänglichkeit (accessibility) untersucht werden.
- *Directory WEF*: einzelne Directories einer Website lassen sich ebenfalls über eine URL bzw. erweiterte URL identifizieren und bezüglich ihrer WEF’s untersuchen. Directories bzw. die HTML-Dateien, die sich in einem Directory befinden, werden über trunkierte URLs (siehe Trunkierungszeichen /*, Tabelle 3-2) extrahiert.[17]
- *Site WEF*: das gleiche Verfahren wie bei den Directory WEF’s funktioniert auch auf die gesamte Website gesehen. Über die Trunkierung aller Directories und deren HTML-Dateien werden Accessibility-Aussagen über die gesamte Website möglich.

[17] *Mike Thelwall* beschreibt in seinen Advanced Web Document Models die Identifikation des Web Documents Directory folgendermaßen: “*All HTML files in the same directory are treated as a*

Die Abbildung 3-1 verdeutlicht die Web-Entitäten Site, Directory und Page. Eine Website kann demnach aus unterscheidbaren Webseiten (pages) und Verzeichnissen (directories) bestehen. In einem Directory können sich wiederum Unterverzeichnisse und Webseiten befinden.

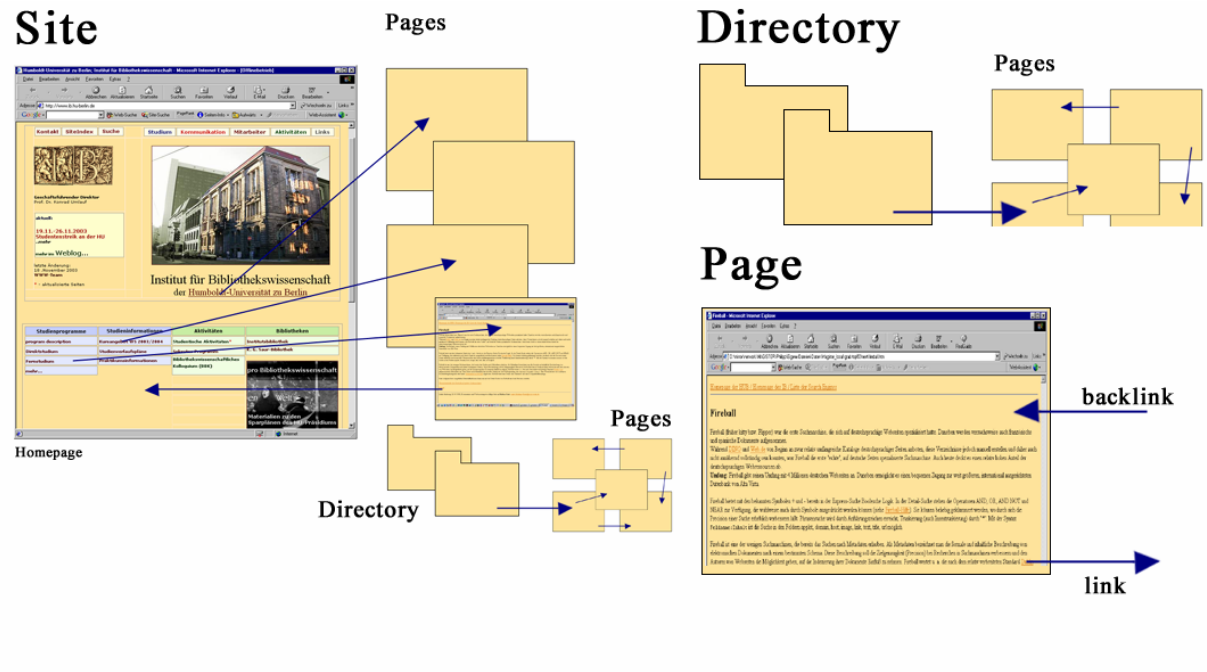


Abbildung 3-1: Web-Entitäten Site, Directory, Page

Denkbar sind weitere Differenzierungen der Entries bzgl. der Navigationsarten „Suchmaschine", „Direkt" und „Referenz" (erweiterte WEF's). Es wäre beispielsweise möglich die Entries der Navigationsart „Suchmaschine" (Suchmaschinen-WEF's) detaillierter z.B. nach dem Namen und/oder der Domain einer Suchmaschine zu unterscheiden. So werden erweiterte WEF-Werte wie (*„WEF_google"*, *„WEF_yahoo"* bzw. *„WEF_google.de"* bzw. *„WEF_google.uk"*) möglich. Analog zu der Differenzierung der „Suchmaschinen-WEF's" wäre es möglich, die „Referenz-WEF's" weiter zu unterteilen. Beispielsweise wären Log-Auswertungen bzgl. der „Referenzen" einzelner Domains (Land, Universität, etc.) möglich.

3.2.4 Konzeption der Untersuchung

Untersuchungsgegenstand sind Webserver Logfiles zweier kompletter Jahrgänge (2000 und 2002) der Website des Instituts für Bibliothekswissenschaft an der Humboldt-Universität zu Berlin (http://www.ib.hu-berlin.de/). Bei der untersuchten Website handelt es sich um eine mehrere hundert Webseiten umfassende universitäre Institutswebsite aus den Wissenschaftsbereichen Informationswissenschaft, Dokumentation und Bibliothekswissenschaft. Innerhalb der Domain „ib.hu-

document. All target URLs are automatically shortened to the position of the last slash, …" (*Thelwall*, 2002)

berlin.de“[18] befindet sich ein breites Spektrum an unterschiedlichsten wissenschaftlichen und nichtwissenschaftlichen Webangeboten, die inhaltlich von Informationsseiten zum Studium am Institut über persönliche Homepages der Mitarbeiter und Studenten bis zu Datenbankangeboten und Volltextartikeln reichen. Die Untersuchung der Logdaten konzentriert sich hauptsächlich auf die Extraktion und Analyse der definierten Zugriffsarten auf eine Auswahl 100 hochfrequentierter Start-, bzw. Einstiegsseiten (Entry Pages) dieser Site, sowie die Berechnung der WEF's dieser Auswahl (vgl. Untersuchung der 100 am häufigsten verlinkten Webseiten britischer Universitätswebsites, *Thelwall*, 2002 [104]). Die Website Entry Faktoren (WEF) basieren auf der Annahme, dass die Anzahl und die Eigenschaften der Website Entries (in folgenden auch als „Site Entries“ bezeichnet) nützliche Indikatoren bei der Bewertung einzelner Seiten bzw. Entitäten innerhalb einer Website liefern. Basis für die WEF's sind die Werte der Site Entries (siehe Abschnitt zu Entries oben). Weiterhin wird die Auswahl der zu untersuchenden Webseiten mit Hilfe einer Klassifikation für Webseiten klassifiziert (siehe *Haas & Grams*, 2000 [41]). Die Klassifikation der Webseiten soll allgemeine Aussagen über die Zusammensetzung und Nutzung der wichtigsten Einstiegsseiten bzw. einzelner Seitencluster dieser umfangreichen Website ermöglichen. Zuletzt soll die Untersuchung zeigen, ob externe Faktoren in Zusammenhang mit der quantitativen Nutzungshäufigkeit von Webseiten gebracht werden können. Zu diesem Zweck wird der PageRank-Wert der Suchmaschine Google (siehe Google's Toolbar, toolbar.google.com) mit in die Analyse der Untersuchungsdaten einbezogen (vgl. Untersuchung *Thelwall*, 2002, „Can Google's PageRank be used to find the most important academic Web pages?“ [99]).
Im Mittelpunkt dieser Untersuchung stehen die webseitenbezogenen Page WEF's der Website.

3.2.5 Triviale Annahmen und Potenziale der Metrik

Was bedeuten hohe bzw. niedrige Entry-Werte bei den drei WEF's für die einzelnen Webseiten? Lassen sich anhand der WEF-Werte spezifische Seitentypen identifizieren? Lassen sich Trends und Probleme in der Zugänglichkeit einer Website anhand des Logs mit Hilfe der WEF-Metrik ausmachen? Folgende triviale Annahmen der Verteilung der Entry-Werte verdeutlichen die Potenziale des Verfahrens WEF.

- Hat eine Seite einen sehr hohen Suchmaschinen-Wert (wef s), dann ist anzunehmen, dass die Seite auf Suchmaschinen-Trefferlisten zu bestimmten Anfragen (z.B. Query 1-n) sehr gut positioniert ist. Damit hat die Seite zu einem bestimmten Thema eine hohe Suchmaschinen-Sichtbarkeit (vgl. dazu *Lawrence*, 2001 [57]). Die Keywords der protokollierten Suchmaschinen-Queries lassen in Gegensatz zu den übrigen Navigationsarten relativ genaue Rückschlüsse auf das Informationsbedürfnis des Besuchers (vgl. *Thelwall*, 2001 [105]) zu. Hohe Suchmaschinen-WEF's bedeuten aber auch, dass die beiden anderen Navigationsarten unterrepräsentiert sind. Die Gründe hierfür können vielfältiger Art sein.

[18] Anmerkung: zu den Besonderheiten der Website zählt, dass sich die Struktur der Website seit 1997 kaum verändert hat.

- Hat eine Seite einen sehr hohen Referenz-Wert (wef r), dann ist anzunehmen, dass die Seite von intensiv genutzten Webseiten („referrers“ oder „hubs“) verlinkt ist. Die Verteilung der Nutzungszahlen könnte in dem Fall auf eine bestimmte Art von Webseite hindeuten z.B. „Authority“-Seiten[19] (vgl. dazu das linkbasierte Konzept „hubs & authorities“ von *Kleinberg*, 1999 [49]). Beinhaltet eine häufig referenzierte Seite selber viele „Outlinks“ (Links, die auf eine externe Webressource z.B. authority verweisen), wird die Seite eventuell als „hub page“ genutzt. Eine „hub page“ definiert *Kleinberg* als eine Webseite die Links auf „Authorities“ innerhalb des Webs oder einer Community enthält. Hohe Referenz-Werte einer Webseite deuten aber zumindest daraufhin, dass die Webseite verlinkt ist und die Links genutzt werden. Der WEF-Algorithmus liefert somit die Webseiten, die aufgrund der Verteilung der Entries „Nutzungs-Authorities“ innerhalb der analysierten Website darstellen.

- Hat eine Webseite einen hohen Wert „direkter“ Entries (wef d) dann liegt nahe, dass die Seite einen hohen Bekanntheits- und Etablierungsgrad im Web oder einer Community hat. Eine Seite mit hohen direkten Werten hat sich demzufolge als Startseite etabliert bzw. wurde durch diverse Maßnahmen bekannt gemacht. Grundsätzlich ließe sich daraus auch folgern, dass Webseiten mit hohen direkten Werten ebenfalls auf Authorities und Hub-Pages schließen lassen. Dieses Zugangsverhalten trifft i.d.R. insbesondere auf Homepages, Projektstartseiten und andere Startseiten innerhalb der Site zu.

Wie bereits an einigen Stellen angedeutet, liefert die Logmetrik WEF eine Reihe von Möglichkeiten, die mit heutigen herkömmlichen Loganalyse-Tools nicht denkbar sind. Insbesondere die Möglichkeit anhand der detaillierten WEF-Werte die Zugänglichkeit (accessibility bzw. visibility) zu messen und damit Hinweise auf die Bedeutungen einzelner Webseiten zu erhalten, birgt Potenziale für künftige Web Loganalysen. Die Darstellung der Anwendungsmöglichkeiten für die Metrik ist hiermit sicherlich noch nicht abgeschlossen. Weitere Anwendungen und Erweiterungen des Konzepts WEF sind denkbar.
Grundlage der folgenden Untersuchung bilden die serverseitig gespeicherten Zugriffsdaten (Logdaten) der Zeiträume Januar bis Dezember 2002 und 2000. Die Vergleichsdaten aus dem Jahr 2000 ermöglichen es Entwicklungen und Tendenzen des Webuse der untersuchten Website zu skizzieren.

Im folgenden Teil werden die Methoden dieser Untersuchung vorgestellt.

---

[19] *“Hyperlinks encode a considerable amount of latent human judgment, and we claim that this type of judgment is precisely what is needed to formulate a notion of authority. (…) We observe that a certain natural type of equilibrium exists between hubs and authorities in the graph defined by the link structure, and we exploit this to develop an algorithm that identifies both types of pages simultaneously.” (Kleinberg, 1999 [49])*

4 Methoden

Im folgenden sollen nun die einzelnen Schritte und Methoden des Loganalyseverfahrens WEF beschrieben werden (vgl. Abb. 4-1).

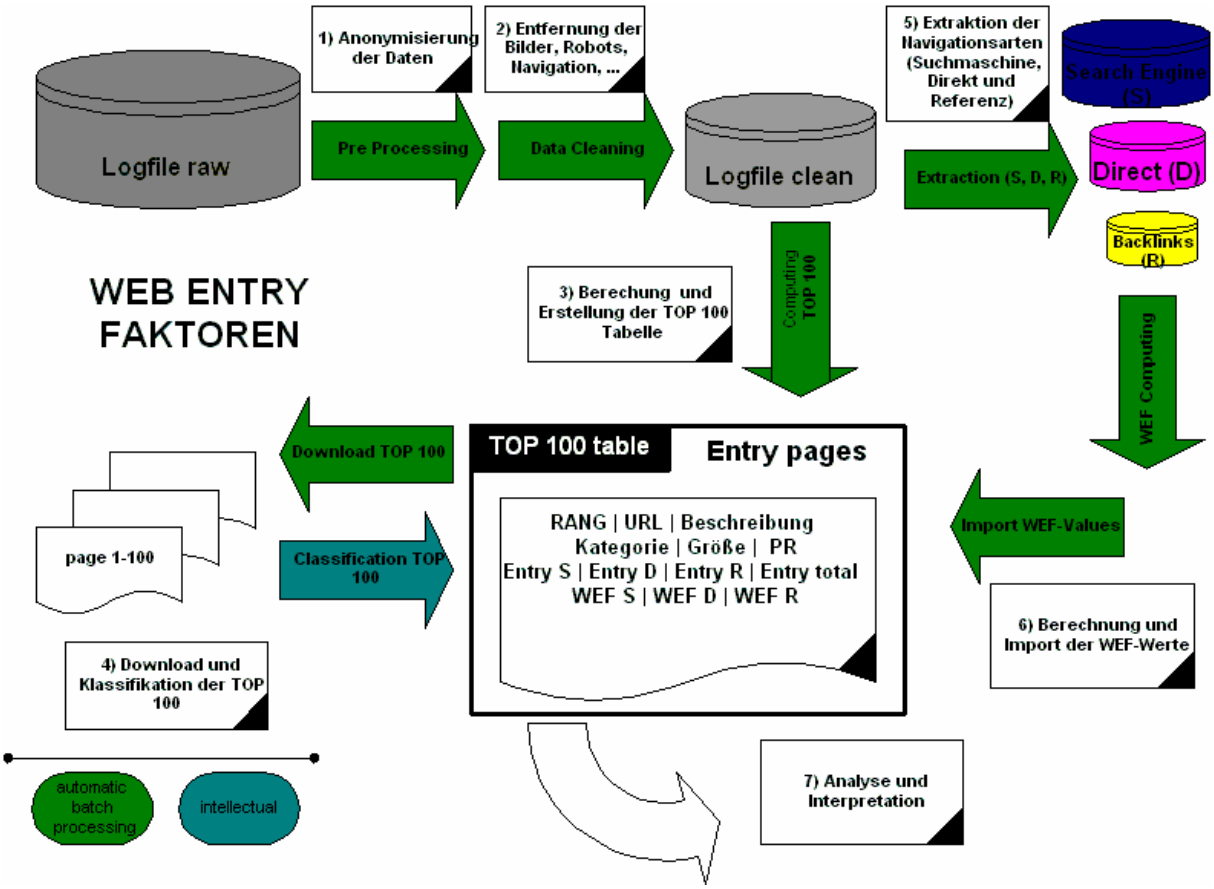


Abbildung 4-1: Loganalyseverfahren Web Entry Faktoren

4.1 Pre-Processing

Bevor die Analyse der Logdaten beginnt, wird in einem ersten Schritt die Anonymisierung der Logdaten vorgenommen. Zu diesem Zweck ersetzt ein Perlscript[20] alle IP-Adressen und Domainnamen des Logfile-Feldes HOST. Das Script fragt für jeden Zeileneintrag im Log den Domain-Name-Service (DNS) des Webservers ab und ersetzt die IP-Adresse mit der Top Level Domain (TLD) des zugreifenden Rechners (z.B. DE, COM, NET, usw.). Rechnerzugriffe aus den IP-Bereichen der Humboldt-Universität (HUB) bzw. dem Institut für Bibliothekswissenschaft (IBdHUB) werden gesondert identifiziert und als zwei verschiedene Besuchergruppen behandelt.

[20] Das Ersetzungsscript zur Anonymisierung der Logdaten wurde von Herrn Michael Heinz erstellt und freundlicher Weise für diese Untersuchung zur Verfügung gestellt.

Der folgende Schritt des Analyse-Verfahrens WEF ist das Säubern der rohen, ungefilterten Logdaten. Dieser Prozess wird im Webmining allgemein „data cleaning“ (vgl. *Cooley, et al.*, [21]) genannt. Das ungefilterte Logfile enthält in chronologischer Abfolge alle vom Webserver aufgezeichneten Transaktionen. Unter diesen Transaktionseinträgen befinden sich unter anderem fehlgeschlagene Zugriffe, Zugriffe auf Elemente, die vom Besucher nicht explizit angefordert werden und Roboter-Zugriffe. Diese erwähnten Zugriffsarten werden beim „data cleaning“ entfernt, da sie für die meisten Logfileanalysen nicht relevante Informationen[21] darstellen. Folglich werden alle Zugriffe auf grafische Elemente und Bilder (z.B. Dateinamensuffix gif oder jpg) und andere Nicht-HTML-Elemente (z.B. Dateinamen-Suffix css), die beim Aufrufen einer Seite automatisch mitgeladen werden und somit jeweils einen Eintrag verursachen, aus dem zu untersuchenden Logfile entfernt. Weiterhin werden alle fehlerhaften Zugriffe (z.B. Fehlercode 404, Redirections), weitere irrelevante Logfileeinträge und insbesondere die „maschinellen“ Zugriffe der Suchmaschinen-Roboter (z.B. „Googlebot“) entfernt (vgl. *Huberman et al.,* 1998 [44], *Nicholas et al.*, 1999 [67], *Oldenburg*, 2003 [71]). Beim Nichtentfernen der oben erwähnten Zugriffe besteht die Gefahr, dass sich die folgenden statistischen Auswertungen verfälschen (data quality, vgl. *Kosala & Blockeel*, 2000 [53]). Angemerkt sei, dass „data cleaning“-Algorithmen, insbesondere für Web Logfiles, eine gewisse Unschärfe beinhalten, da in einigen Fällen die Zugriffsinformationen im Log nicht darauf schließen lassen, zu welcher „Sorte“ von Zugriffen die Zugriffszeile gehört. Insbesondere die Entfernung der „maschinellen“ Zugriffe, und hier der „unstandardisierten“ Roboter oder Crawler, stellt ein Problem bei der Bereinigung der Logfiles dar. Die Zugriffe von selbstgeschriebenen oder -konfigurierten Robotern lassen sich z.T. nur schwer aus dem Logfile entfernen, weil allgemeine Suchmaschinenregeln[22] nicht eingehalten werden.

Für die Untersuchung wird aus methodischen Gründen nur der erste Zugriff (vgl. Einstieg bzw. Entry) auf die Website, also eine bestimmte URL gezählt. Alle nachfolgenden Webseitenzugriffe eines Websitebesuchs, die hier als „Navigationszugriffe“ (bzw. Navigation) bezeichnet werden, werden ebenfalls aus dem Logfile entfernt. Unter Navigation werden die Zugriffe verstanden, die über interne Links innerhalb der Websitedomain ib.hu-berlin.de („internal links“, vgl. [117]) erfolgt sind. Kurzum, alle Websitebesuche werden durch das Pre-Processing der hier beschriebenen Untersuchung auf die erste aufgerufene Seite (Einstiegsseite bzw. Entry Page) reduziert (siehe Tabelle 4-1).

[21] Welche Einträge letztlich für die Auswertung als relevant oder irrelevant bewertet werden, hängt stark vom Untersuchungsfokus ab. *Nicholas et al.* (1999) [67] sprechen in diesem Zusammenhang von der Problematik des „over & under reporting“, wobei sie beispielsweise den internen Traffic durch Mitarbeiter als „over reporting“ bezeichnen.

[22] Die Suchmaschinenbetreiber und das W3C haben sich auf einen Quasi-Standard geeinigt, der vorschreibt, dass die Crawler oder Robots der Suchmaschinenbetreiber bei ihren regelmäßigen Websitebesuchen die Konventionen der Datei „robots.txt“ beachten. In dieser Datei spezifiziert der Websitebetreuer welche Suchmaschinen auf seiner Website indexieren dürfen und welche Verzeichnisse bzw. Dateien die Robots nicht berücksichtigen sollen. In der Regel halten sich die Robots an diese Konventionen und fordern bei jedem Besuch zuerst die Datei robots.txt an. „Seriöse“ Robots geben ihren Namen im Useragent-Feld des Logfiles an.

Log vor Pre-Processing	Log nach Pre-Processing	Einstiegsseite (Entry Page)
start.html ->[23] second.htm	start.html ~~-> second.htm~~	start.html
index.htm -> studium.htm -> suche.htm -> home.htm	index.htm ~~-> studium.htm →suche.htm -> home.htm~~	index.htm
home.htm -> lit.htm -> cv.htm -> home.htm	home.htm ~~-> lit.htm -> cv.htm -> home.htm~~	home.htm

Tabelle 4-1: Beispiele Pre-Processing (Entfernung der Navigation)

Der „Data Cleaning"-Prozess ist beendet, wenn das Logfile nur noch aus Einträgen besteht, die auf die erste zugegriffene Seite (siehe Einstiegsseiten, rechte Spalte, Tabelle 4-1) eines Websitebesuchs zurückgehen.
Bemerkungen: Nach dem Pre-Processing sind Messungen bzgl. der Länge von Websitebesuchen aufgrund der Entfernung der Navigation nicht mehr möglich. Die Zugriffe der Mitarbeiter und Studenten des Instituts wurden bewusst nicht aus dem Logfile entfernt.

Die folgenden Schritte der Untersuchung bestehen aus einer mehrstufigen Abfolge von Arbeitsabläufen, die sowohl aus automatisierten als auch intellektuell manuellen Schritten bestehen.

4.2 Berechnung der Top 100

Nach dem „Data Cleaning" verbleiben lediglich die Zugriffe der Websitebesuche im Logfile, die durch das Pre-Processing als Einstiegszugriffe identifiziert wurden. Aus den verbleibenden Zugriffsinformationen wird über einen zweiten Schritt die Menge der Webseiten errechnet, die am häufigsten zugegriffen wurden. Dies erfolgt über das Auszählen aller Entry-Zugriffe für jede Webseite (URL). Ergebnis dieses Schritts ist eine sortierte Rangliste der Seiten, die im Untersuchungszeitraum am häufigsten als erste Webseite aufgerufen wurden. Die Berechnung der Top 100 kann grundsätzlich auch mit jedem Standard-Loganalyzer erfolgen, vorausgesetzt der Loganalyzer analysiert die zuvor gefilterten Daten. Die Länge der Rangliste wird für diese Untersuchung auf die 100 am stärksten frequentierten Einstiegsseiten des Webservers begrenzt (siehe komplette Liste der Top 100 Webseiten im Anhang [10.1]). Die Begrenzung der untersuchten Webseiten auf 100 hat keinen methodischen Hintergrund, sondern ist in erster Linie dem zeitlichen Aufwand bei der intellektuellen Klassifikation der einzelnen Seiten geschuldet (siehe folgender Schritt [4.3]). Vorherige Arbeiten aus den Bereichen der Bibliometrie und Webometrie (*Garfield*, 1990 [36], *Thelwall*, 2002 [104]) haben sich ebenfalls für die ersten 100 Entitäten (Top 100) als Auswahl eines Untersuchungsfeld beschränkt.

[23] Das Pfeilsymbol -> steht für die Nutzung eines internen Links (internal link).

Anschließend werden die errechneten Webseiten der Top 100 URLs für die weitere Untersuchung lokal geladen und für die weitere Analyse unter dem Originaldateinamen gespeichert. Für das Laden und Speichern der Top 100 wurde die Software Teleport Pro[24] verwendet.

4.3 Klassifikation der Top 100

Im folgenden Schritt steht die Analyse und Klassifizierung der Inhalte der Top 100 Webseiten im Mittelpunkt. Die Untersuchung konzentriert sich auf unterschiedliche Kategorieklassen, die sich zum einen auf den Inhalt und zum anderen auf die Größe der Webseiten beziehen. Grundsätzlich wäre es an dieser Stelle auch interessant gewesen weitere Parameter der einzelnen Webseiten zu erheben (z.B. Alter der Seite, Anzahl der Links auf der Seite (interne und externe), Anzahl der Worte auf der Seite, usw.).

Die Klassifizierung bzw. Indexierung von Webinhalten stellt innerhalb der Dokumentation ein relativ neues Anwendungsgebiet der Publikationsanalyse dar. Aufgrund der Neuheit dieser Disziplin sowie der ständig in Veränderung befindlichen Vielfalt von Web-Publikationstypen und digitalen Genres (vgl. *Dillon & Gushrowski*, 2000 [26]), lässt sich für die Einordnung von Webseiten noch nicht auf etablierte Klassifikationsinstrumente zurückgreifen, wie es sie für die traditionellen Publikationen seit langem gibt. Der Metadaten-Standard der Dublin Core Metadata Initiative (www.dublincore.org) stellt mit seinen fünfzehn Elementklassen eine Ausnahme dar, wobei der Fokus der Initiative eher auf der Erschließung der Internetressourcen zu Recherchezwecken[25] liegt.

Um weiterreichende Aussagen über die wichtigsten 100 Einstiegsseiten einer akademischen Website treffen zu können, ist es erforderlich die einzelnen Webseiten Kategorien bzw. Inhaltsklassen zu zuordnen. Erste Webseiten-Typologien, die allerdings mit sehr unterschiedlichen Methoden entstanden sind, liefern die Kategorienklassen verschiedener Untersuchungen (*Pirolli et al.*, 1996 [76], *Cronin et al.*, 1998 [23], *Middleton et al.*, 1999 [66], *Haas & Grams*, 2000 [41], *Bar-Ilan*, 2000 [7], *Dublin Core*[26] [27]). Die Grundlage für die Bildung der inhaltlichen Webseiten-Klassen dieser Untersuchung bildet das Klassifikationsschema von *Haas & Grams* [41]. *Haas & Grams* stellen in ihrem Konzept sieben Klassen von Webseiten vor (siehe Tab. 4-2), die sie in einer umfangreichen Inhaltsanalyse von Webseiten im Jahr 1998 entwickelt haben. Die sieben auf *Haas & Grams* zurückgehenden Klassen, beinhalten eine gewisse Unschärfe, auf die die Autoren hinweisen. Beispielsweise kann die Inhaltsklasse „Text“ als sehr breit angelegte Klasse angesehen werden. Zur Text-Klasse lassen sich viel mehr Seitentypen zuordnen als zu den übrigen spezifischeren Klassen. *Haas & Grams* schreiben, dass es während ihrer Untersuchung grundsätzlich möglich war eine Webseite mehreren Klassen zuzuordnen (Anmerkung: klassisches Indexer-Problem). Außerdem geben die beiden Autoren zu bedenken, dass sie die Klassen und Unterklassen, lediglich aus einer

[24] Teleport Pro ist ein freiverfügbarer Webgrabber.

[25] *"The Dublin Core Metadata Initiative is nearing the fifth anniversary of what has become the broadest international, interdisciplinary effort in resource description on the Internet. It is the leading initiative for improving resource discovery on the Web. ..." Weibel, 2000 [118]*

[26] Das achte Dublin Core Element „DC.Type“ sieht die Zuordnung einer Webressource zu einer umfangreichen, von der DCMI erarbeiteten, Liste von DC-Types (Klassen & Genres) vor.

zufälligen bzw. pseudozufälligen Stichprobe erarbeitet haben (vgl. *Bar-Ilan*, 2000 [5]). Daher kann und will ihre Klassifikation keinerlei Ansprüche auf ein vollständiges allgemeingültiges Klassifikationsschema erheben.

„We do not claim that these classes and subclasses are exhaustive; they were developed from our sample pages and we have no way of knowing what types we have missed. Given the rapid development of the Web, however, potential lack of coverage or newly created or previously unseen types will always exist, and any classification procedure must be able to handle them in a useful way." (Haas & Grams, 2000, S. 191, [41])

Jede der Top 100 Einstiegsseiten wird einer der unterstehende Seitenkategorien nach *Haas & Grams* zu geordnet. Dieser Schritt der Untersuchung wurde vom Autor durch Autopsie jeder einzelnen Webseite durchgeführt. Dazu wurde jede Webseite in den Browser geladen und einer der sieben Seitenkategorien zugeordnet.

Seitenkategorien	Beschreibung	Beispiele für Unterklassen
Organizational (Orga)	Seiten, die Hilfestellungen in bezug auf Websitestruktur und Seitennavigation bieten.	Indexseiten, Sitemap, Inhaltsverzeichnisse
Documentation (Docu)	Seiten, die Beschreibungen und Erklärungen zu einem spezifischen Thema enthalten und als Referenz fungieren.	FAQ, Tutorial, How-to, Beschreibungen
Text (Text)	Seiten, die beliebige Texte von Personen, Gruppen etc. enthalten.	Artikel, Paper, Forschungsbericht, Vertrag, Bibliographie, Lebenslauf, Biographie, Copyright, …
Home Page (Home)	Seiten, die eine Organisation oder Person einführend darstellen und Links zu weiterführenden Seiten enthalten.	Berufliche bzw. private Home Page einer Person, Organisationshomepage,
Multimedia	Seiten, die nichttextliche Dokumente wie z.B. Musik, Videos oder Bilder enthalten.	Sound, Videos, Bilder, Grafiken, Spiele, interaktive Seiten
Tools	Seiten, die es ermöglichen dem Besucher eine Aufgabe online durchzuführen.	Webanwendungen (z.B. Suchmaschine), Formulare (Bestellung, Email, Kommentar)
Database Entry (DB_Entry)	Seiten, die den Einstieg zu einer Datenbank darstellen bzw. strukturierte Information enthalten.	Bibliografische Daten, Katalogdaten

Tabelle 4-2: Kategorien für Webseiten nach Haas & Grams

Als weiterer Schritt gilt es, die 100 Webseiten bzgl. ihrer Dateigröße zu klassifizieren. Eine Einteilung in drei Klassen (große, mittlere und kleine Seiten) wird dazu vorgenommen. Gemessen wird die Dateigröße der untersuchten Seiten in Byte. Der HTML-Code der Seite wird dazu nicht entfernt. Die Bestimmung der drei Größenklassen erfolgt über folgendes Verfahren:

- Logarithmisierung der Dateigrößenangaben aller untersuchter Webseiten $lg(p_i)$,
- Sortierung der logarithmisierten Größenwerte,
- Subtraktion des kleinsten Größenwertes (min) vom größten (max),
- Division des Ergebnisses der Subtraktion (max ($lg\ p_i$) – min ($lg\ p_i$) durch die Anzahl der Größenklassen (3).

Die Division durch die Anzahl der Größenklassen (hier drei) ergibt den Schrankenquotient q.

$$q = ((\max (\lg p_i) - \min (\lg p_i)) / 3$$

Die Subtraktion des Schrankenquotienten q vom größten Größenwert (max) ergibt die erste Schranke s1. Alle Größenwerte der Webseiten, die größer sind als die Schranke s1, werden der Seitenklasse „Groß" zugeordnet. Der zweite Schrankenwert s2 wird durch Subtraktion des Schrankenquotienten q von s1 errechnet. Alle Größenwerte der Webseiten, die kleiner s1 und größer gleich s2 sind, werden der Seitenklasse „mittel" zuordnet. Alle Größenwerte, die kleiner als s2 sind, gehören der Seitenklasse „klein" an.

$$s1 = \max (\lg p_i) - q$$
$$s2 = s1 - q$$

für alle $p_i >= s1$ gilt, p_i = Element von Seitenklasse Groß (lg)
für alle $p_i < s1$ und $>= s2$ gilt, p_i = Element von Seitenklasse Mittel (av)
für alle $p_i < s2$ gilt, p_i = Element von Seitenklasse Klein (sm)

Die Klassifikationsmerkmale (Inhaltstyp und Größe) der einzelnen Seiten werden in der Liste der Top 100 vermerkt (siehe Anhang [10.1]). Die Begutachtung und Klassifizierung der einzelnen Seiten konzentriert sich auf verschiedene Aspekte wie die Beschreibungen, die der Seite zu entnehmen sind z.B. Titel, Überschrift, Textinhalt, Links auf der Seite sowie die Kontexte der Links. Die Begutachtung der Top 100 und die anschließende Klassifizierung wurden manuell durch Autopsie jeder einzelnen Webseite vorgenommen.[27]

Als zusätzlicher externer Parameter bzw. Vergleichswert wird für jede analysierte Webseite der PageRank[28] des Suchmaschinenbetreibers Google zusätzlich in die Liste der Top-Seiten mitaufgenommen. Jede Webseite wird dazu mit einem Browser geladen, der über die Google Toolbar verfügt und den Toolbar-PageRank (toolbar.google.com) ausgibt[29]. Die PageRank-Werte über die Google Toolbar können Werte von null bis zehn annehmen. Hohe PageRank-Werte deuten daraufhin, dass Webseiten gemäß Google über viele Links (Backlinks) oder Links von Webseiten mit hohem PageRank verfügen [vgl. 72, 13]. Als Vergleichswert hätte es sich weiterhin angeboten die Linkangaben einzelner Suchmaschinen für die untersuchten Webseiten mit in die „Top 100-Liste" aufzunehmen[30].

---

[27] Erste Vorschläge zur automatischen Klassifikation von Webseiten liegen aus dem Bereich des Web Mining vor (*Fu et al.*, [34]).

[28] Der PageRank einer Webseite hängt sowohl davon ab, wie viele Seiten auf sie zeigen, als auch vom PageRank dieser anderen Seiten. Der PageRank einer Seite ist hoch, wenn sie von anderen Seiten mit einem hohen PageRank verlinkt wird. Eine Webseite mit wenigen Links trägt mehr zum PageRank einer Seite bei als eine Webseite mit vielen Links [vgl. 72, 13].

[29] Hinweis: der Wert des PageRank-Wertes über die Google-Toolbar ist nicht gleich dem PageRank-Wert, der über den Originalalgorithmus errechnet wird (vgl. *Thelwall*, 2003 [99]).

[30] Die meisten Suchmaschinen ermöglichen es die Anzahl der Links, die auf eine bestimmte Webseite verweisen über eine simple Suchmaschinenquery (z.B. link:http://www.beispiel.de/beispiel.htm) abzufragen.

4.4 Extraktion der Navigationsarten

Ein entscheidender Schritt des Verfahrens WEF besteht in der Extraktion der drei verschiedenen Navigationsarten (vgl. Entry-Kategorien) aus dem gefilterten Log. Für diesen Extraktionsvorgang spielt die Information eines speziellen Felds des Logfiles eine besondere Rolle. Die Einträge in diesem Feld geben darüber Auskunft über welche Navigationsart der Zugriff auf die protokollierte Seite erfolgt ist. Im sogenannten Referer-Feld (HTTP_REFERER, siehe [3.1.1]) protokolliert der Webserver die URL der Seite, von der die zugegriffene Seite angefragt (geklickt) wurde [vgl. 105]. Das Logfile, vorausgesetzt es ist im Logformat *Extended Common Logfile Format (ECLF)*, ermöglicht es die drei Navigations-arten bzw. Entry-Arten mithilfe einfacher Bedingungen und regulärer Ausdrücke zu extrahieren (siehe Beispiele unten). Die Identifikation der drei Entry-Arten aus dem Logfile wird nachfolgend dargestellt.

1. Suchmaschinen-Entries (Queries)

Besonders aufschlussreich sind die Entries, die im Referer-Feld die URL einer Suchmaschinen-Query enthalten. Im Logfile lässt sich dazu die URL der Suchmaschinen-Trefferliste samt Benutzer-Query identifizieren und analysieren. Diese Information aus dem Referer-Feld gibt Aufschluss über den Namen der Suchmaschine, die Query und in einigen Fällen ungefähre Angaben über die Position der Seite in der Trefferausgabe der Suchmaschine. Die Query, die aus der URL der Suchanfrage ersichtlich ist und extrahiert werden kann, gibt Aufschluss über das Informationsbedürfnis des Suchenden anhand der verwendeten Keywords- bzw. Phrasen. Die Extraktion der Suchmaschinen-Entries (Queries) erfolgt über einen Filter, der alle Entries aus dem gefilterten Log extrahiert, die die URL einer Suchmaschine (z.B. http://www.google.de/*) und ggf. zusätzlich die Zeichen „?“ und „=“ in der URL enthalten.

LOGZEILE	net - - [06/Jan/2002:11:14:34 +0100] "GET /~fern/fernstudium/magister/magister.html HTTP/1.1" 200 12872 "http://www.google.de/search?q=fernstudium&start=20&sa=N" "Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0)"
Logeintrag Referer-Feld	"http://www.google.de/search?q=fernstudium&start=20&sa=N"
Navigationsart bzw. Entry-Kategorie	Suchmaschine
Extraktionskriterium	URL der Suchmaschine in Kombination zu den Zeichnen „?“, „=“

Tabelle 4-3: Beispiel Logzeile – Navigationsart „Suchmaschine“

2. „direkte“ Entries

„Direkte“ Entries, die durch direktes Aufrufen der Seite entstanden sind (vgl. „direkte“ Navigation), lassen sich am fehlenden Eintrag im Referrer-Feld (HTTP_REFERER) des Logfiles identifizieren. Ein

Zugriff ohne Referer-Eintrag, entspricht im Logfile dem Eintrag "-" (siehe Beispiel unten). Zu den „direkten“ Zugriffen ist weiterhin anzumerken, dass Suchmaschinen-roboter i.d.R. ebenfalls „direkt“ zugreifen. Grundsätzlich kann daher nicht ausgeschlossen werden, dass sich auch nach dem Pre-Processing (data cleaning) einzelne Roboterrequests unter den „direkten“ Zugriffen befinden.

Logzeile	net - - [04/Jan/2002:15:09:11 +0100] "GET /~fern/fernstudium/magister/magister.html HTTP/1.0" 200 12872 "-" "Mozilla/4.73 [de]C-CCK-MCD DT (Win98; U)"
Logeintrag Referer-Feld	"-"
Navigationsart bzw. Entry-Kategorie	Direkt
Extraktionskriterium	"-" im Referer-Feld

Tabelle 4-4: Beispiel Logzeile – Navigationsart „Direkt“

3. Entries über Referenzen (Backlinks)

Die dritte Möglichkeit auf eine Website zu gelangen, erfolgt über einen externen Link (Backlink). Diese Zugriffsart kann ebenfalls im Logfile identifiziert werden. Der Webserver schreibt dazu, wie bei den Suchmaschinen-Entries, die URL in das Referer-Feld, über die der Zugriff auf die protokollierte Seite erfolgt ist (siehe Beispiel unten). Für diese Untersuchung werden alle URLs als Referenzen bzw. Backlinks gewertet, die Links auf die IB-Site enthalten und nicht aus dem Domainbereich des Webservers ib.hu-berlin.de stammen. Directoy-Einträge, wie das „Open Directory“ (dmoz.org) oder vergleichbare andere Onlinekataloge (z.B. Yahoo, dir.yahoo.com) werden ebenfalls zu den „Referenzen“ der IB-Site[31] gezählt. Alle Entries, die zuvor als Suchmaschinen- und Direkte Entries identifiziert worden sind, werden aus dem ungefilterten Log entfernt. Die verbleibenden Entries im Log werden als Referenzen gewertet (vgl. Extraktionskriterium, Tabelle 4-5).

Logzeile	net - - [14/Jan/2002:12:52:14 +0100] "GET /~fern/fernstudium/magister/magister.html HTTP/1.1" 200 12872 "http://www.inf-wiss.uni-konstanz.de/FG/IV/mitarbeiter.html" "Mozilla/4.0 (compatible; MSIE 5.0; Windows 2000) Opera 5.12 [de]"
Logeintrag Referer-Feld	"http://www.inf-wiss.uni-konstanz.de/FG/IV/mitarbeiter.html"
Navigationsart bzw. Entry-Kategorie	Referenz (Backlink)
Extraktionskriterium	Referenzen = Log_gefiltert - (Direkte Entries + Suchmaschinen-Entries)

Tabelle 4-5: Beispiel Logzeile – Navigationsart „Referenz“ (Backlinks)

[31] Die Website des Instituts für Bibliothekswissenschaft wird nachfolgend verkürzt IB-Site genannt.

Alle Zugriffe der drei verschiedenen Navigationsarten werden nach der Extraktion für die weitere Untersuchung in separaten Dateien gespeichert. Das folgende Listing zeigt die Darstellung des Extraktionsalgorithmus der drei Navigationsarten in der Pseudocodedarstellung.

```
#KOMMENTAR

#L* = gefiltertes Log
#z = Zeile im Log
#Entry_s,d,r sind Extraktionskriterien bzgl. der drei Navigationsarten
#(Suchmaschine, Direkt und Referenz)

# Extraktion der Navigationsarten
For z = 1 to n
  If (z = Entry_s) then
    Write z to Entries_Suchmaschine
  Endif
  If (z = Entry_d) then
    Write z to Entries_Direkt
  Endif
  If (z = Entry_r) then
    Write z to Entries_Referenz
  Endif
Endfor
```

Listing 2: Pseudocode des Extraktionsalgorithmus

Der folgende Abschnitt beschreibt den Aufbau der Web Entry Faktoren, die im Mittelpunkt dieser Arbeit stehen.

4.5 Berechnung der Web Entry Faktoren[32]

Die Zusammensetzung und Berechnung der Web Entry Faktoren schließt direkt an den vorhergehenden Schritt der Extraktion der Zugriffe nach den verschiedenen Navigationsarten an. Die Berechnung der drei WEF's für beliebige Webentitäten liefert folgender Algorithmus.

```
# count Entries
For each URL = 1 to n
       count Entries_Suchmaschine
       count Entries_Direkt
       count Entries_Referenz
       sum Entries_Direkt + Entries_Suchmaschine + Entries_Referenz =
       Entries_total
Endfor

# compute WEF
For each URL = 1 to n
       Entries_Suchmaschine / Entries_total = wef_Suchmaschine
       Entries_Direkt / Entries_total = wef_Direkt
       Entries_Referenz / Entries_total = wef_Referenz
Endfor
```

Listing 3: Pseudocode des WEF-Algorithmus

[32] Diese Untersuchung und insbesondere der Ergebnisteil konzentriert sich auf die detaillierte Analyse der Entität Webseite (Page).

Jede untersuchte URL erhält drei Entry-Werte (Suchmaschine, Direkt und Referenz). Die Summe der drei Werte ergibt die Gesamtanzahl an Entries für die URL. Die drei WEF's werden für jede URL der Top 100 errechnet, indem jeder der drei Entry-Werte einer URL durch die Gesamtanzahl an Entries für diese URL dividiert wird.
Im Anschluss an die Berechnung der WEF-Werte stehen folgende Einträge in der Liste der Top 100 Startseiten für die weitere Analyse zur Verfügung.
Die Liste der Top 100 Webseiten (siehe „Top 100-Liste" [10.1]) setzt sich aus folgenden Einträgen (Spalten) zusammen:

- *Rang*: Nummer des Rangs der Seite, gemessen an der Anzahl der absoluten Entries (Entries total)
- *URL*: URL der Seite innerhalb der Domain www.ib.hu-berlin.de/
- *Beschreibung*: kurze Beschreibung des Inhalts der Seite
- *Kategorie*: Zuordnung der Seite zu einer Inhaltskategorien (nach *Haas & Grams* [41])
- *Größe*: Größe der Seite in Byte, eingeteilt in die drei Größenkategorien (klein, mittel, groß)
- *PR*: PageRank-Wert der Seite (vgl. Google Toolbar)
- *Entries S*: Absolute Anzahl der Entries über Suchmaschinen (Queries) für die Seite
- *Entries D*: Absolute Anzahl der Entries über „direkte" Eingaben (z.B. Bookmarks etc.) für die Seite
- *Entries R*: Absolute Anzahl der Entries über Referenzen (Backlinks) für die Seite
- *WEF SE*: „Suchmaschinen-WEF" (wef s = Entries S / Entries total)
- *WEF D*: „Direkt-WEF" (wef d = Entries D / Entries total)
- *WEF R*: „Referenz-WEF (wef r = Entries R / Entries total)
- *Entries total*: absolute Anzahl der Entries für die Seite

Jede Seite der Top 100 erhält die oben beschriebenen Einträge. Die Spalten „Beschreibung", „Kategorie", „Größe" und „PR" werden manuell durch Klassifikation bzw. gesonderte Eintragung in die Liste eingefügt. Die übrigen Werte (Entries und WEF's) liefert der WEF-Algorithmus (vgl. Pseudocode, Listing 3 oben).
Insbesondere die Identifikation der Ausreißer-Seiten scheint hier besonders interessant und lohnend[33]. Die Ausreißer-Seiten lassen sich relativ leicht durch Filtern der WEF-Faktoren aus der „Top 100-Liste" extrahieren. Das Filtern einer Liste von WEF-Faktoren kann über die Errechnung der Quartile erfolgen.
Beispiel: Über das obere Quartile (75 %-Quartile) lassen sich beispielsweise die 25 höchsten Werte (z.B. WEF's oder Entries) der 100 Webseiten ausgeben. Analog dazu lassen sich die kleinsten Werte, das untere Quartil (25%-Quartil), der Median und der größte Wert einer Matrix bestimmen. Weiterhin sind kompliziertere kombinierte (z.B. Filter 1 kombiniert mit Filter 2) Filterprozeduren möglich.

[33] Anmerkung: Ausreißer-Seiten sind in diesem Zusammenhang Seiten innerhalb der Top 100, die sich aufgrund ihrer Nutzungswerte deutlich von den anderen Seiten dieser Auswahl abheben.

4.6 Analyse der Queries und Backlinks

Die Analyse der Suchmaschinen-Queries (Queries) und Referenzen (Backlinks) stellt einen gesonderten Bereich dieser Untersuchung dar. Nachfolgend sollen einzelne Bereiche skizziert werden, die ein detailliertes Bild der Suchmaschinen-Entries bzw. -Queries und der Referenzen (Backlinks) liefern (vgl. *Thelwall,* 2001 [105]).

4.6.1 Queries

Zum einen lassen sich aus den Zugriffen der Navigationsart „Suchmaschine“ die Suchmaschinen identifizieren, die den meisten Traffic (gemessen an den Entries) für eine Website bringen. Zum anderen bieten die im Logfile vorhandenen Daten der protokollierten Userqueries ein weites Feld zusätzlicher Analysen des Online- bzw. Suchverhalten der Webuser einer Site. Nachfolgend können lediglich einzelne Hinweise gegeben werden, welche Möglichkeiten für die Analyse der Suchmaschinen-Queries denkbar sind. Die Extraktion der Queries aus den Suchmaschinen-URLs ist vergleichbar einfach, zumal die großen Suchmaschinenbetreiber einheitlich beim Aufbau der URL-Syntax (Query-Strings) vorgehen. Beispielweise macht der Suchmaschinenbetreiber Google die Queries innerhalb der URL über das Kürzel (q=) kenntlich.
Die URL-Parameter und Query-Strings der Suchmaschinen-Entries liefern folgende Informationen:

- Name und Top Level Domain der Suchmaschine (Beispiele: google.de, google.ca, google.uk),
- Zusammensetzung der Queries, z.B. Anzahl der Keywords (Beispiele: q=ascii+code, q=%22codex+amiatinus%22),
- Position einer Seite in einer Treffermenge zu einer spezifischen Query (Beispiele: q=Manuskript&hl=de&start=10, q=schaltungssammlung&hl=de&start=50),
- Einsatz von Operatoren bzw. Suchmaschinenfeatures (z.B. And, Near, Similar, Cache, http://www.google.com/search?q=cache:j_Z9geGHjKs:www.ib.hu-berlin.de/~wumsta/umlit90.html+&hl=de).

Diese oben skizzierten Informationen der Suchmaschinen-Entries lassen folgende Berechnungen bzw. Analysen zu:

- Top Queries bzw. Top Keywords pro Entität (Rangliste),
- Anzahl der unterschiedlichen Queries bzw. Keywords pro Entität,
- Inhaltliche Einordnung der Queries (Kategorisierung der Queries vgl. *Thelwall*, 2001 [105]),
- Unterschiedlicher Aufbau der Queries bzgl. der verschiedenen Suchmaschinen,
- Unterschiede der Queries bzgl. der Besucher, Domains, Länder[34].

[34] Die Analyse der Suchmaschinen-Queries unterschiedlicher Besuchergruppen (z.B. Länder, Domains) würde anhand der Logdaten möglich (vgl. die Untersuchung zu Web Suchstrategien finnischer und amerikanischer Onlinesuchende von *Iivonen & White*, 2001 [45]).

4.6.2 Backlinks

Die Logeinträge der Referenzen (Backlinks) der Site liefern den Namen und die Top Level Domain (TLD) der referenzierenden Sites bzw. URL's. Diese Angaben lassen folgende Analysen zu (vgl. *Cui*, 1999 [25]).

- Absolute Anzahl der verschiedenen Sites bzw. URL's, die auf die untersuchte Site referenzieren und Traffic generieren,
- Verteilung der Domains der referenzierenden Sites bzw. URL's nach TLD,
- Fachliche Einordnung, Klassifizierung der referenzierenden Sites (z.B. akademisch, kommerziell, Community, Directory, Universität, Partner, …),[35]
- Anteil der Referenzen, die einer Zitation einer Publikation entsprechen.

[35] Vgl. *Thelwall*, 2001 [105]

5 Ergebnisse

Das folgende Kapitel präsentiert die wichtigsten Ergebnisse der Untersuchung und diskutiert einzelne Punkte an. Es werden vor allem Ergebnisse der Analysen der Logdaten der Website www.ib.hu-berlin.de/ aus dem Jahr 2002 dargestellt. An einigen Stellen werden die Daten aus dem Jahr 2000 zu Vergleichszwecken herangezogen. Der Ergebnisteil unterteilt sich in folgende Bereiche.

5.1 Ergebnisse des Pre-Processing

Der Ergebnisteil beginnt mit der Darstellung der Pre-Processing-Ergebnisse (vgl. Data Cleaning [4.1]). Im Jahr 2000 verbleiben für den Zeitraum Januar bis Dezember von ursprünglich 4.715.037 Zeilen (Log roh) nur 449.727 Zeilen im ungefilterten Logfile (9, 5 %, siehe Log gefiltert). Im Jahr 2002 reduziert der Data Cleaning-Prozeß das ursprüngliche Logfile (6.739.902 Zeilen) auf 820.678 Zeilen, somit verbleiben 12,2 % der Daten zur Analyse. Das ungefilterte Logfile des Jahres 2002 verzeichnet 91,2 % mehr Zugriffe gegenüber dem Jahr 2000 (vgl. Tab. 5-1).

Zeitraum	**Log roh**	**Log gefiltert**
Jan.-Dez. 2000	4.715.037 Zeilen	449.727 Zeilen
Jan.-Dez. 2002	6.739.902 Zeilen	820.678 Zeilen

Tabelle 5-1: Anzahl der Zeilen vor und nach dem Pre-Processing (2002, 2000)

Die untere Abbildung (Abb. 5-1) stellt die Anteile der Logfile-Einträge für das Jahr 2002 dar, die durch das Pre-Processing (Data Cleaning) aus dem ungefilterten Logfile entfernt wurden. 42% der Zugriffe im Logfile 2002 (2.444.676 Zeilen) gehen auf Bilder, 32% auf Roboterzugriffe (1.887.701 Zeilen), 9% auf Codes und Fehler (543.639 Zeilen) und 7% auf Nicht-HTML-Inhalte (428.969 Zeilen). 10% der Zugriffe (614.588 Zeilen) gehen auf interne Navigationszugriffe auf HTML-Seiten zurück.

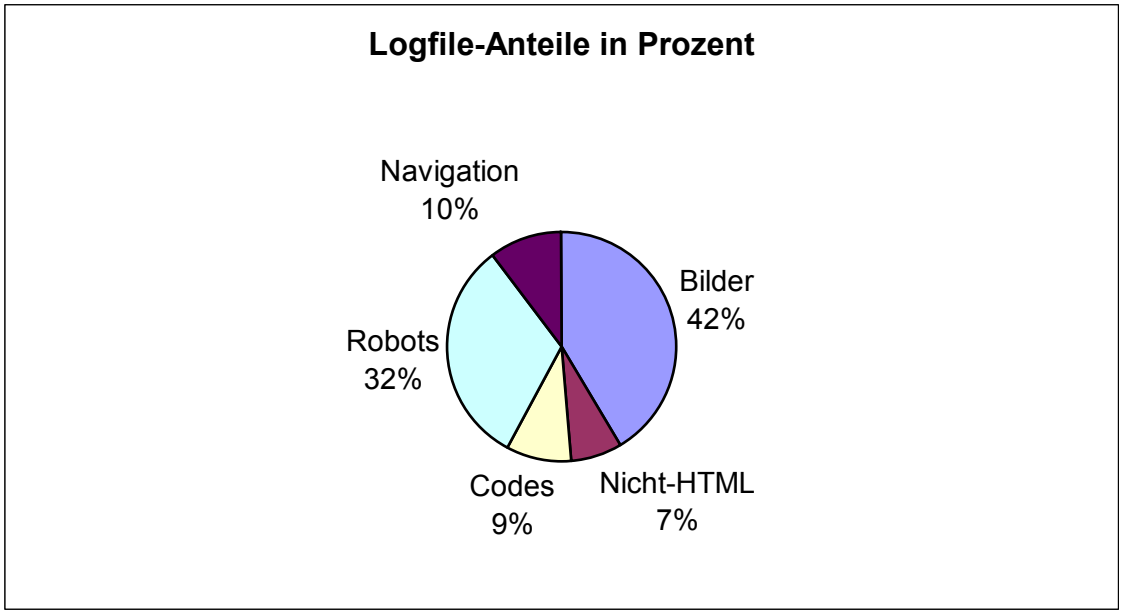


Abbildung 5-1: Logfile-Anteile in Prozent (2002)

Beim Vergleich der Pre-Processing-Ergebnisse des Jahres 2002 mit dem Referenzdatensatz aus dem Jahr 2000, fällt auf, dass im Logfile von 2000 deutlich weniger Zugriffe auf Bilder[36] (gif, jpeg, etc.) und Nicht-HTML-Elemente (css, pdf, xml) zu verzeichnen sind (34 % Bilder und 2 % Nicht-HTML im Jahr 2000 im Vergleich zu 42 % und 7 % im Jahr 2002). Die Anzahl der Navigationszugriffe ist mit 10 % (2002) bzw. 9 % (2000) der Zeilen in beiden Jahren in etwa gleich geblieben. Auffällig ist der deutlich höhere Roboter-Anteil (32 % im Jahr 2002, 41 % im Jahr 2000) im Logfile aus dem Jahr 2000.

5.2 Extraktion der Navigationsarten

Die Extraktion der drei unterschiedlichen Zugangs- bzw. Navigationsarten aus dem gefilterten Logfile ergibt für beide Untersuchungszeiträume, dass der überwiegende Anteil der Websitezugänge (im Folgenden als „Entries“ bzw. „Siteentries“ bezeichnet) aus Zugriffen besteht, die über Suchmaschinen bzw. Suchmaschinen-Queries erfolgt sind. Insgesamt (vgl. Web-Entität Site) gehen im Jahr 2002 59 % der Entries (485.963) auf Suchmaschinen, 34 % (276.502) auf direkte Eingaben (Bookmark, Verlauf, etc.) und lediglich 7 % (58.213) auf Referenzen (externe Links, sog. „Backlinks“) zurück.
Wichtig ist in diesem Zusammenhang zu erwähnen, dass die relativ hohe Anzahl an Entries (siehe Zeilenangaben in der unteren Tabelle) darauf zurückzuführen ist, dass bei der Zählung der Entries kein Timeout[37] gesetzt wurde. Jede Zeile des gefilterten Logs entspricht einem Pageview (Entry), der einer der drei Navigationsarten zugeordnet werden kann. Die in der unteren Tabelle aufgeführten Zeilenangaben entsprechen somit der absoluten Anzahl an Pageviews bzw. Entries (nicht Visits) für die drei unterschiedlichen Navigationsarten.

[36] vgl. *Nicholas et al.* , 1999 [67]

Zeitraum	Suchmaschine	Direkt[38]	Referenz
Jan.-Dez. 2000	253.044 Zeilen (56%)	146.685 Zeilen (33%)	49.998 Zeilen (11%)
Jan.-Dez. 2002	485.963 Zeilen (59%)	276.502 Zeilen (34%)	58.213 Zeilen (7%)

Tabelle 5-2: Anzahl der Zeilen der drei Navigationsarten (Anteile in Prozent) (2002, 2000)

Die Ergebnisse der Extraktion der drei Navigationsarten für den Vergleichsdatensatz (2000) zeigen, dass sich der Anteil der Suchmaschinen von 56% im Jahr 2000 auf 59% im Jahr 2002 gesteigert hat. Weiterhin hat die Navigationsart „Referenz“ in beiden Jahren deutlich die geringsten Entry-Werte erhalten. Im gleichen Zeitraum hat diese Navigationsart sogar an Bedeutung verloren (minus 4%). Die „direkte“ Navigation ist in ihren Anteilen um 1% gestiegen. Eine weitere Untersuchung der Logdaten des IB-Servers am Institut für Bibliothekswissenschaft aus dem Jahr 2003 zeigt für die Jahre 1999 und 2001 eine tendenziell ähnliche Verteilung (*Oldenburg*, 2003 [71]).

5.3 Website Traffic

Der folgende Abschnitt stellt einzelne Ergebnisse dar, die sich auf den Traffic bzw. Webuse des analysierten Webservers www.ib.hu-berlin.de beziehen. Die zugrundeliegende Maßeinheit ist wieder die Anzahl der Siteentries bzw. Entries.
Der 12-Monatsverlauf der beiden Untersuchungszeiträume zeigt für das Jahr 2002 ein sehr geringes Wachstum bzgl. der Anzahl der Siteentries. Die Anzahl der Siteentries des Untersuchungszeitraum 2000 ist im Vergleich dazu deutlich stärker gewachsen (vgl. Trendlinien der Abb. 5-2).

[37] Timeouts werden i.d.R. eingesetzt um einzelne Websitebesuche (visits) voneinander abzugrenzen. Hierbei haben sich ca. 30 Minuten als Timeoutgrenzen etabliert.
[38] Es ist anzunehmen, dass die Werte für die „direkten“ Entries (siehe Spalte Direkt) aufgrund schwer zu identifizierender Roboterzugriffe tendenziell zu hoch sind.

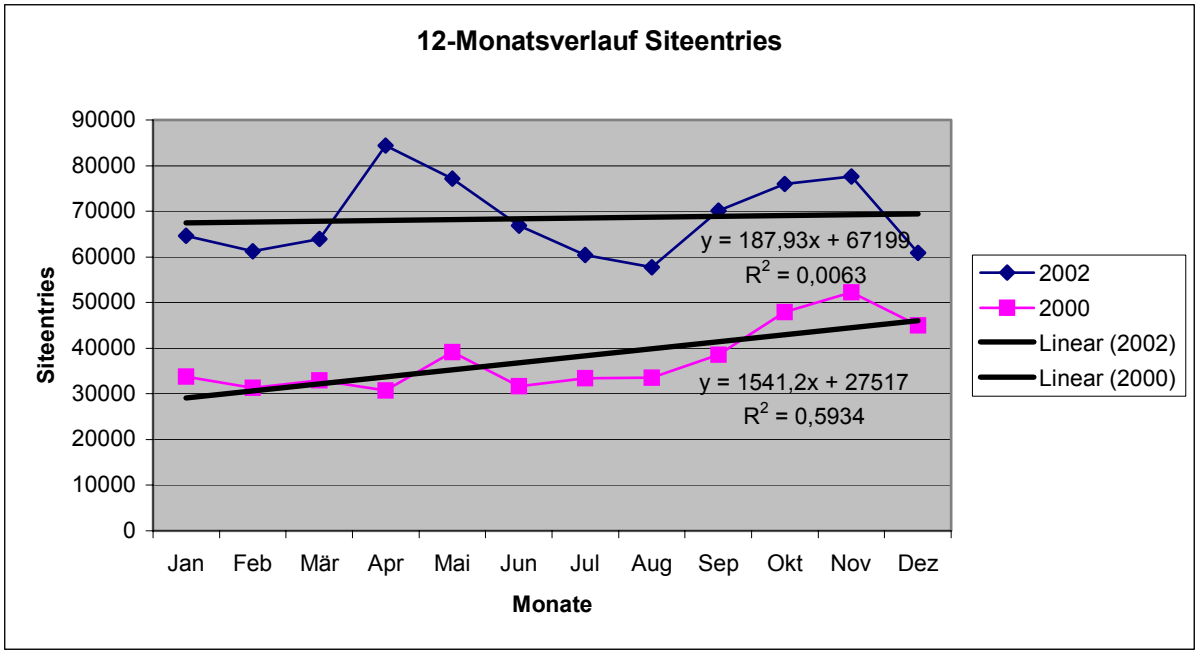


Abbildung 5-2: 12-Monatsverlauf Siteentries (2002, 2000)

Weiterhin werden die zuvor extrahierten Navigationsarten (Suchmaschine, direkte Navigation und Referenz) getrennt voneinander dargestellt. Die untere Abbildung (Abb. 5-3) stellt den Verlauf der Siteentries bzgl. der drei Navigationsarten für den Untersuchungszeitraum Januar bis Dezember 2002 dar.

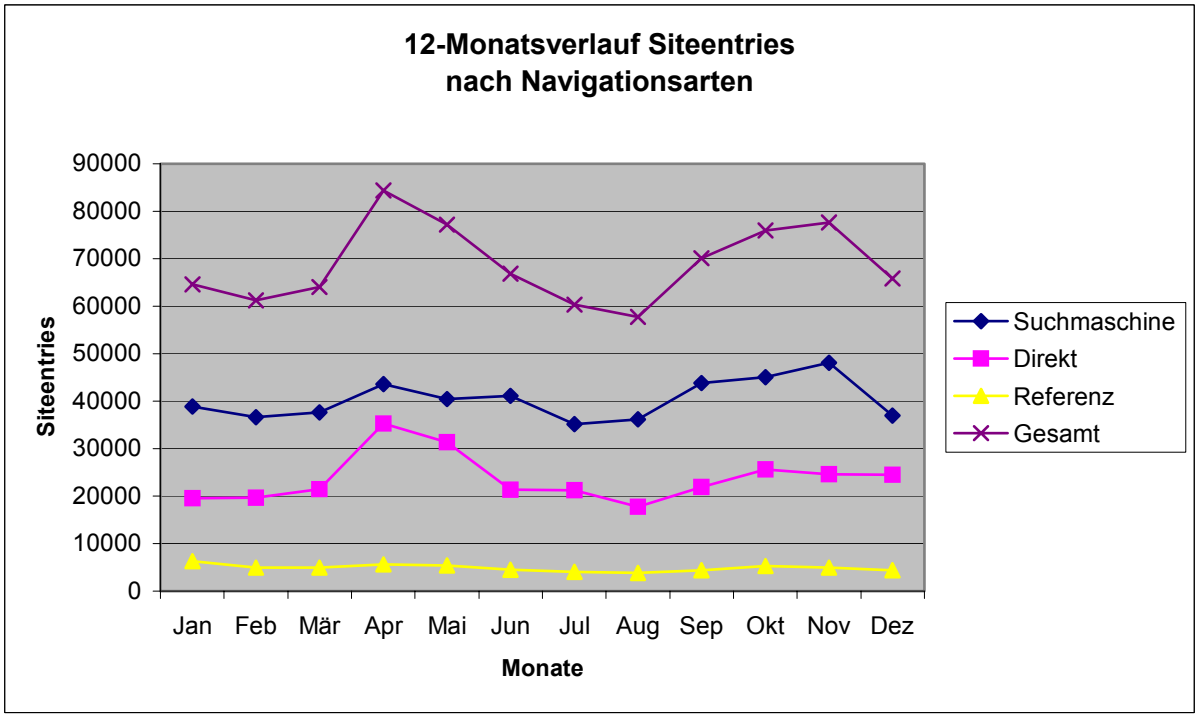


Abbildung 5-3: 12-Monatsverlauf der Navigationsarten (2002)

Die Abbildungen (Abb. 5-2, 5-3 und 5-4) zeigen deutlich, dass die Siteentries jeweils unmittelbar vor Semesterbeginn im April und Oktober sprunghaft ansteigen. Aus Abbildung 5-3 wird sichtbar, dass die beiden Anstiege auf die Navigationsarten „Suchmaschine“ und insbesondere „direkte“ Navigation zurückzuführen sind (siehe Abb. 5-3, Monate April und September). Der Anteil der „Referenzen“ bleibt über das Jahr hinweg konstant niedrig. Kleinere Ausschläge zu Beginn der beiden Semester können bei genauerer Betrachtung aber auch bei den „Referenzen“ wahrgenommen werden.

Die Abbildung 5-4 zeigt die Siteentries über Rechner aus dem Domainbereich (First Level Domain) der Humboldt-Universität (vgl. HUB) bzw. dem Institut für Bibliothekswissenschaft der Humboldt-Universität (siehe IBdHUB). Beide Besuchergruppen[39] zeigen, über das Jahr (2002) gesehen, einen tendenziell ähnlichen Verlauf (Anstieg in der Semesterzeit). Die IBdHUB-Gruppe liegt dabei deutlich über der HUB-Gruppe und verzeichnet stärkere Ausschläge. Es zeigt sich ebenfalls ein Anstieg der Siteentries beider Besuchergruppen in der Semesterzeit (siehe untere Abbildung). Die Peaks der Zugriffe befinden sich jeweils zwei bis vier Wochen nach Semesterbeginn. Besonders auffällig ist der rasante der Anstieg der Siteentries der IBdHUB-Gruppe zum Beginn des Wintersemesters[40]. Die geringsten Siteentries für die Besuchergruppe HUB und IBdHUB lassen sich jeweils in der vorlesungsfreien Zeit einen Monat vor dem Semesterbeginn (März und September) verzeichnen. Der Webuse (gemessen an den Siteentries) für alle Besucher und für die Universitätsangehörigen ist damit abhängig von der Semesterzeit.

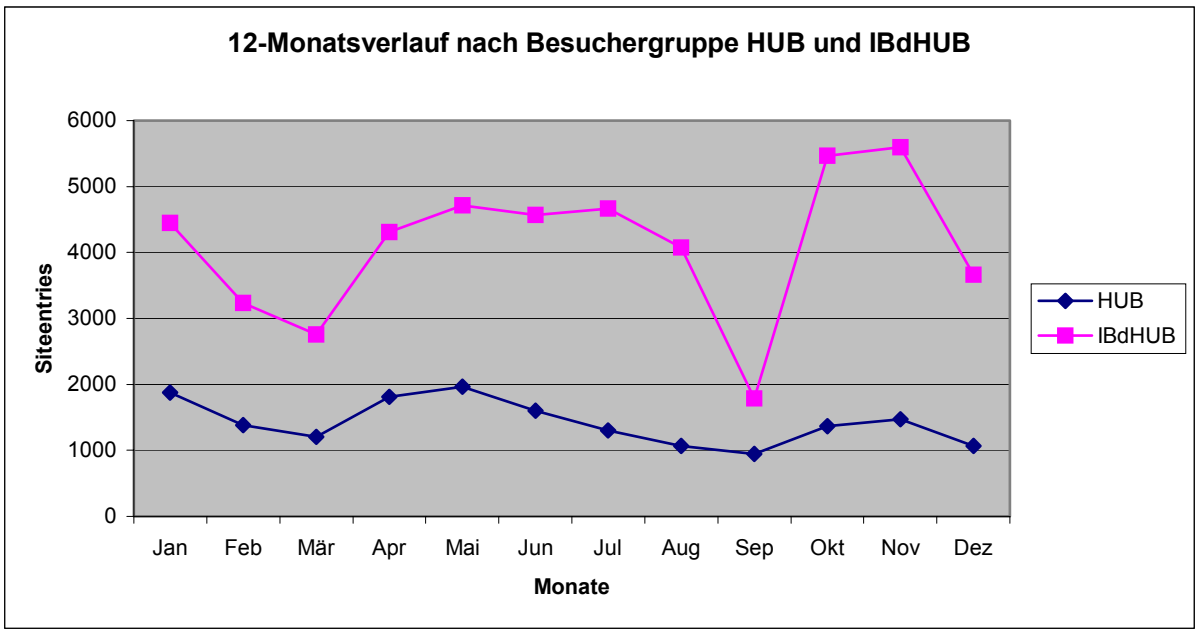


Abbildung 5-4: 12-Monatsverlauf Siteentries über Rechner der HU und des IB (2002)

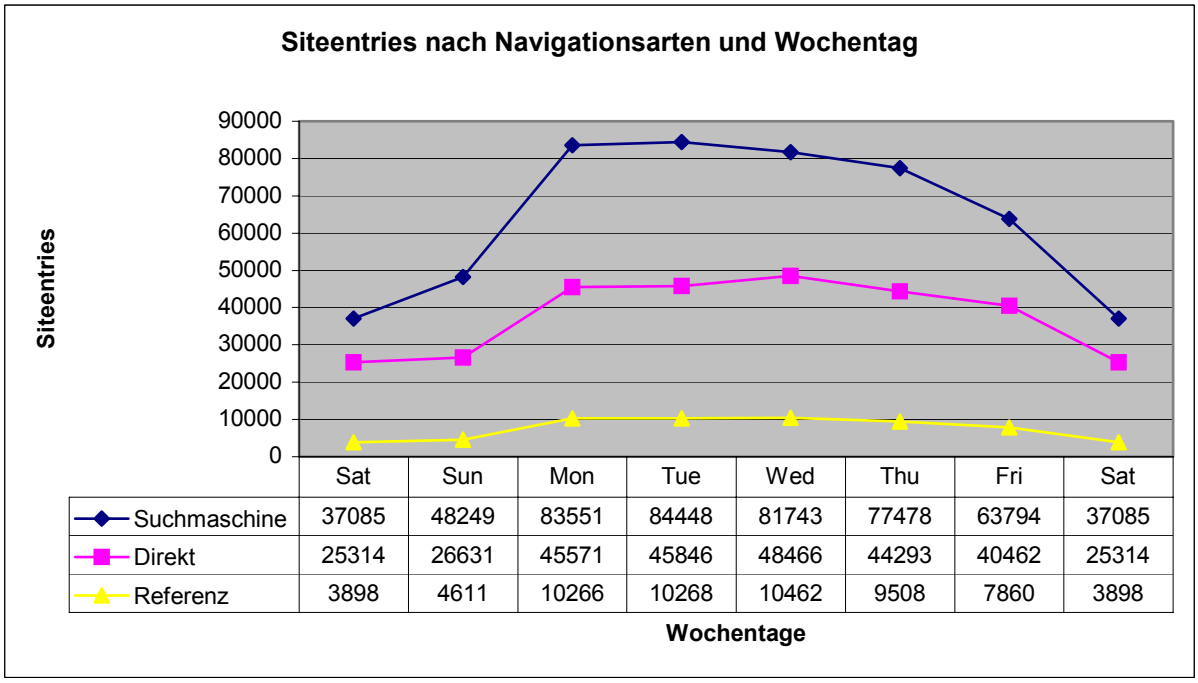
[39] Zu den beiden Besuchergruppen sind vereinfachend alle Studenten und Mitarbeiter der jeweiligen Institutionen zu zählen, die über Computer aus dem Domainbereich ihrer Institution auf die IB-Site zugreifen [vgl. 4.1]. Die folgenden Abbildungen stellen die Verteilung der Siteentries bezogen auf die beiden Besuchergruppen HUB und IBdHUB dar. HUB steht hier für Personen aus dem Domainbereich der Institution Humboldt-Universität. IBdHUB steht für alle Personen aus dem Domainbereich des Instituts für Bibliothekswissenschaft der Humboldt-Universität. Angemerkt sei, das die Humboldt-Universität sehr viel mehr Computer (inkl. externe Einwahl über das Humboldt-Rechenzentrum) und dementsprechend mehr Personen besitzt als das vergleichsweise kleine Institut für Bibliotheks-wissenschaft.

5.4 Allgemeine Nutzungszahlen

Die Nutzung der IB-Site bezogen auf den Wochentag bzw. die Uhrzeit zeigt wenig überraschende Details.

Am Wochenende verzeichnet das Logfile die geringste Anzahl an Siteentries. Bei der Betrachtung der beiden Wochenendtage Samstag und Sonntag fällt auf, dass der Sonntag für alle drei Navigationsarten deutlich vor dem Samstag liegt. Insbesondere wird am Sonntag häufiger über die Navigationsart „Suchmaschine“ auf die IB-Site navigiert. Die übrigen Wochentage Montag, Dienstag und Mittwoch liegen bzgl. der Siteentries sehr nahe zusammen. Der Donnerstag und der Freitag zeigen weniger Siteentries als die drei vorherigen Wochentage (vgl. Abb. 5-5).

Die Entwicklung der Siteentries bezogen auf die Uhrzeit zeigt Abbildung 5-6. Die Siteentries über die Navigationsart Suchmaschine liegen deutlich über den der anderen Navigationsarten (Ausnahme: „direkte“ Siteentries[41] in den frühen Morgenstunden 2.00-6.00 Uhr). Die Anzahl der Siteentries auf die IB-Site steigt von 6.00 Uhr bis 14.00 Uhr bis auf eine kleine Stagnation um die Mittagszeit (12.00-13.00) stetig an. Nach 14.00 Uhr nehmen die Siteentries aller drei Navigationsarten kontinuierlich ab.



	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Suchmaschine	37085	48249	83551	84448	81743	77478	63794	37085
Direkt	25314	26631	45571	45846	48466	44293	40462	25314
Referenz	3898	4611	10266	10268	10462	9508	7860	3898

Abbildung 5-5: Siteentries nach Navigationsarten und Wochentag (2002)

[40] Hinweis: im Wintersemester schreiben sich traditionell mehr Studenten ein als im Sommersemester. Diese Tatsache scheint sich im Logfile widerzuspiegeln.

[41] Bei den „direkten“ Siteentries in der Zeit von 2.00-6.00 Uhr handelt es sich mit großer Wahrscheinlichkeit um automatisierte Zugriffe von Robotern. Diese Roboterzugriffe lassen sich aus methodischen Gründen nur sehr schwer aus dem Logfile entfernen.

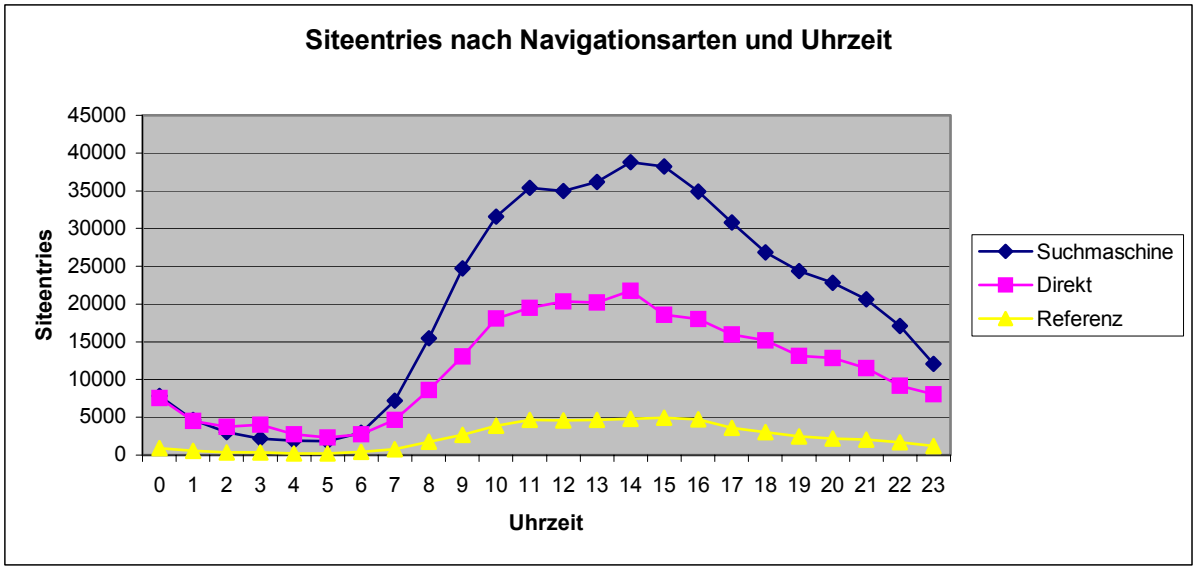


Abbildung 5-6: Siteentries nach Navigationsarten und Uhrzeit (2002)

5.5 Navigation der Besucher

Bei der Betrachtung der Siteentries über die drei Navigationsarten für die Websitebesucher der Humboldt-Universität (HUB), des Instituts für Bibliothekswissenschaft (IBdHUB), der „kommerziellen" Besucher (COM, Besucher mit Top Level Domains .com) und internationale „akademische" Besucher (EDU, Besucher mit Top Level Domains .edu) fällt auf, dass die vier Besuchertypen[42] auf sehr unterschiedliche Weise auf die Website zugreifen. Die Besucher über Rechner des Instituts für Bibliothekswissenschaft, starten den überwiegenden Anteil ihrer Websitebesuche (95%) auf die IB-Site „direkt". Die Navigationsarten „Suchmaschine" und „Referenz" spielen für die Studenten und Mitarbeiter des Instituts mit 2% und 3% nur eine minimale Rolle. Es anderes Bild bzgl. der Navigationsanteile entsteht bei einem genaueren Blick auf die Besucher über Rechner der Humboldt-Universität (HUB). Mit 46% ist ebenfalls die „direkte" Navigation die vorherrschende Navigationsart. An zweiter Stelle befinden sich mit 36% die „Referenzen" (externer Link). Die Suche über Suchmaschinen spielt eine untergeordnete Rolle (18%). Ein anderes Navigationsverhalten zeigen die beiden Besuchergruppen COM und EDU. Diese beiden Gruppen navigieren zu 60% bzw. 64% über die Navigationsart Suchmaschine auf die Website, damit liegen sie knapp über dem Gesamtdurchschnitt (59%). Die direkte Navigation spielt mit ca. 30% die zweit wichtigste Rolle. Die dritte Navigationsart „Referenz" hat für „kommerzielle" Besucher kaum Bedeutung (5%), während „akademische" Besucher viel häufiger (12%) über Backlinks auf die IB-Site navigieren.

[42] Unter Domain wird in diesem Zusammenhang die Top Level Domain des zugreifenden Hostrechners bezeichnet. Diese Rechneradresse befindet sich im Logfile als IP-Adresse, bereits aufgelöster Domainname oder in verkürzter Form. Aus datenschutzrechtlichen Gründen wurden die kompletten IP-Adressen bzw. Host-Domainnamen der Logfileeinträge vor der Untersuchung durch ein Ersetzungsskript auf die Kurzform des jeweiligen Topdomainnamens z.B. de, com, etc. verkürzt (siehe [4.1]). Gesondert wurden die IP-Adressen der Humboldt-Universität (siehe HUB) bzw. des Instituts für Bibliothekswissenschaft der Humboldt-Universität (siehe IBdHUB) behandelt.

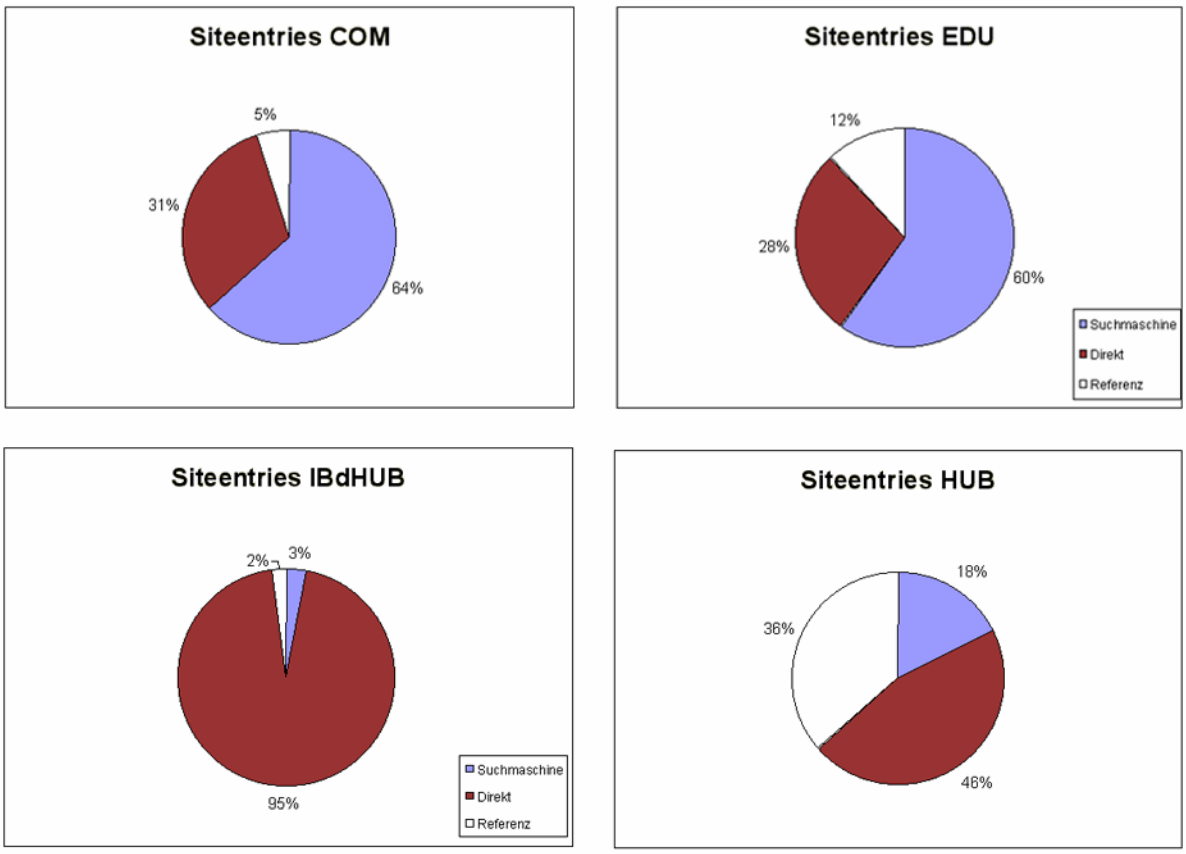


Abbildung 5-7: Navigationsarten der Besuchergruppen (2002)

Die HUB- und IBdHUB-Gruppe unterscheiden sich bezogen auf die Zugangsart deutlich. Insbesondere der hohe Anteil der Referenzen als Navigationsart bei der HUB-Gruppe überrascht. Das Ergebnis deutet daraufhin, dass Links/Referenzen („Backlinks“) auf die IB-Site als Navigationsart insbesondere bei Universitätsangehörigen außerhalb des Instituts (siehe auch EDU) wahrgenommen werden. Der hohe Anteil der „direkten“ Navigation auf die IB-Site von Seiten der Universitätsangehörigen spricht dafür, dass Teile des Angebots der Site innerhalb der Universität bekannt („gebookmarkt“) sind und regelmäßig „direkt“ aufgerufen werden. Außerdem navigieren Universitätsangehörige viel häufiger über Suchmaschinen auf die IB-Site als Institutsangehörige. Dieses Ergebnis deutet daraufhin, dass das Angebot der Website des IB den Universitätsangehörigen viel weniger bekannt ist, als den Institutsangehörigen, die vorwiegend direkt navigieren. Außerdem ist davon auszugehen, dass die HUB-Besuchergruppe aus einer sehr viel größeren und heterogenerer Besucherschaft besteht als die vergleichsweise kleine IB-Gruppe. Die Site-Suche, die die Inhalte aller Universitätswebserver über einen gemeinsamen Index recherchierbar macht, spielt im Zusammenhang mit dem Anteil an Suchmaschinenbesuchen für die HUB-Gruppe eine gewisse Rolle. Der Vergleichsdatensatz für das Jahr 2000 zeigt in etwa ähnliche Werte für die Anteile der

Wenn nachfolgend von Besuchergruppen die Rede ist, sind TLD-Gruppen gemeint, die durch die ip-basierten Ersetzungen entstehen.

Navigationsarten der beiden Besuchergruppen (HUB und IBdHUB). Die HUB-Gruppe navigiert dabei im Jahr 2000 noch häufiger über Referenzen und weniger über Suchmaschinen auf die IB-Site.
Die untere Tabelle und Abbildung (Tab. 5-3 und Abb. 5-7) zeigt die Verteilung der Top Level Domains bzgl. der Nutzung der IB-Site (Anzahl der Siteentries). Zur Übersicht werden die drei Navigationsformen in der Tabellenansicht separat dargestellt. Im Untersuchungszeitraum 2002 haben die „net-Domains" über alle drei Navigationsarten und somit auch insgesamt die meisten Siteentries generiert. Die „net-Domain" steht in der Regel für Internetprovider (z.B. AOL oder T-Online), die den Zugang für private und geschäftliche Internetzugänge liefern. Die höchste Anzahl an Siteentries geht somit auf Websitebesuche zurück, die über Internetprovider zustande gekommen sind. An zweiter Stelle stehen insgesamt gesehen (siehe Spalte „Total") die „de-Domains". Die Nichtauflösbaren Host-Domain-Adressen (unbekannt[43]), stehen insgesamt an dritter Stelle. Es folgen die „com-Domains" und an fünfter Stelle bereits die Domainnamen der Institutsangehörigen (IBdHUB). Die „direkten" Entries der Host-Domain „IBdHUB" verdeutlichen hierbei den im oberen Abschnitt beschriebenen Navigationstrend für Institutsangehörige. Auffällig ist weiterhin, dass die Universitätsangehörigen (HUB, siehe Rang 9) bzgl. des Webuse zwar hinter „kleinen" Ländern wie Ungarn, Österreich und der Schweiz liegen, aber deutlich vor der „Edu-TLD". Die Hauptsprache der Website (im Fall der IB-Site deutsch) spiegelt sich sichtbar in der Besucherschaft (Domainnamen) wieder. So navigieren „kleine" deutschsprachige Länder (bezogen auf die Anzahl der Internetnutzer) wie Österreich und die Schweiz (siehe Tabelle unten Rang 7 und 8) deutlich häufiger auf die IB-Site als „große" Länder wie Frankreich und England (siehe Tabelle unten Rang 14, 16), die andere Sprachen sprechen. Länder wie Ungarn, Italien und die Niederlande (siehe Tabelle unten Rang 6, 10, 11), die eine gewisse Affinität und Nähe zur deutschen Sprache haben, liegen trotz ihrer sehr viel kleineren Anzahl an Internetnutzern vor „großen Internetländern" wie Kanada und Japan (siehe Tabelle unten Rang 15, 19). Überraschend ist, dass die „Education-Domains" („edu"), zu der akademische Institutionen weltweit und insbesondere die amerikanischen Universitäten gehören, bzgl. der Webnutzung der IB-Site eine weniger wichtige Rolle spielen (siehe Tabelle unten Rang 12).

Rang	Total	Suchmaschine	Direkt	Referenz
1	net	net	net	net
2	de	de	IBdHUB	de
3	unbekannt	unbekannt	de	**UNBEKANNT**
4	com	com	hu	HUB
5	IBdHUB	at	unbekannt	com
6	hu	ch	com	at
7	at	HUB	at	IBdHUB
8	ch	it	HUB	ch
9	HUB	nl	ch	pl
10	it	pl	it	it

[43] Für die IP-Adressen, die der Ersetzungsalgorithmus nicht auflösen kann, wird die Kurzform „unbekannt" eingesetzt.

11	nl	edu	nl	edu
12	edu	fr	edu	nl
13	pl	IBdHUB	ca	fr
14	fr	uk	fr	jp
15	ca	be	jp	uk
16	uk	hu	pl	es
17	be	ca	be	ca
18	es	es	uk	hu
19	jp	lu	org	be
20	lu	jp	se	cz

Tabelle 5-3: Top 20 der Top Level Domains für die drei Navigationsarten (2002)

Die Abbildung 5-7 zeigt die Anzahl der absoluten Siteentries für die 20 wichtigsten Top Level Domains seiner Besucher.

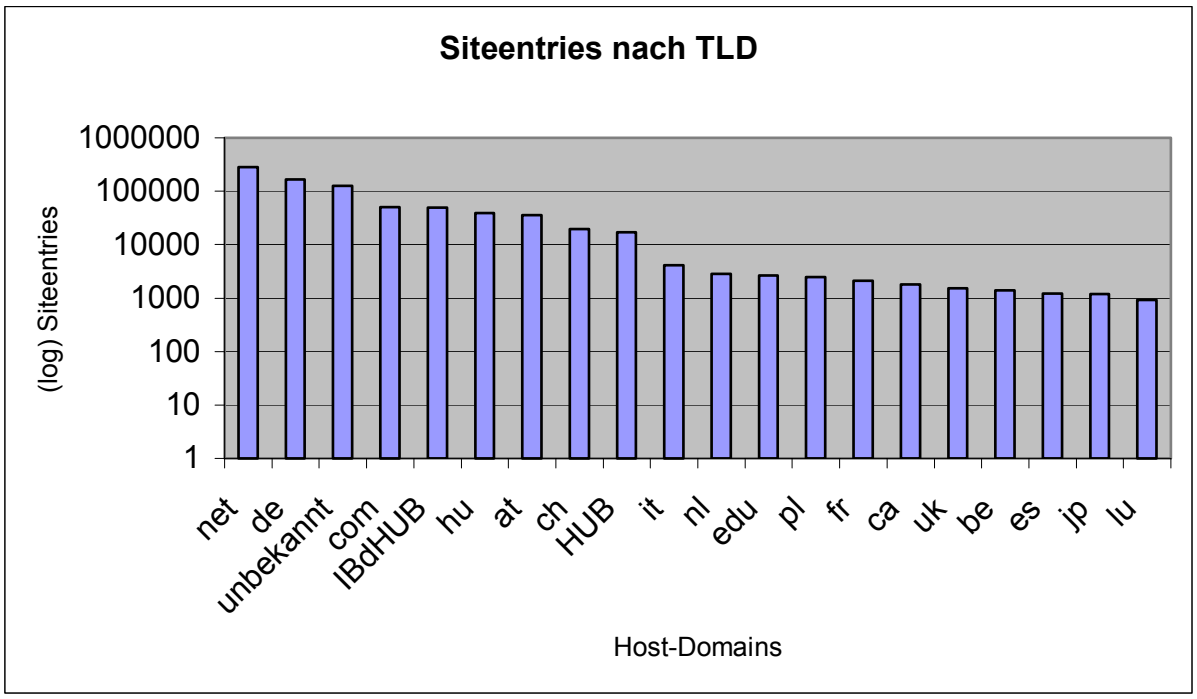


Abbildung: 5-8: Siteentries nach Top Level Domains (Top 20)[44]

5.6 Die „Top 100-Liste“

Ein Hauptergebnis der Untersuchung sind die Ergebnisse der Berechnung der Top 100 Webseiten. Die Liste besteht aus 100 Webseiten, die im Untersuchungszeitraum die meisten Entries verzeichnen konnten.[45] Die Tabelle 5-4 zeigt die ersten fünf Einträge der „Top 100-Liste“ (siehe Anhang [10.1]) für das Jahr 2002.

[44] Hinweis: logarithmischer Maßstab (siehe log Siteentries)

[45] Lediglich eine einzelne Seite der Top 100 (2002) konnte zum Untersuchungszeitpunkt Sommer 2003 nicht mehr auf dem Webserver des IB lokalisiert werden. Die Seite mit der URL /~sbuett/pm/strat_pm2.html enthält dementsprechend keine Klassifizierung und wird in den Statistiken, die sich auf die definierten Klassen beziehen, nicht berücksichtigt.

Rang	URL	Beschreibung	Kate-gorie	Size	PR	Entry S	Entry D	Entry R	WE F S	WE F D	WE F R	Entry total
1	/	Homepage des Instituts für Bibliotheks-wissenschaft (IB Homepage)	Home	av	6	6345	54369	6558	0,09	0,81	0,10	67272
2	/~mh/gedv/ascii.htm	Referenz der ASCII-Code Kodierung	Docu	av	4	19248	2399	187	0,88	0,11	0,01	21834
3	/~mh/projekte/metaopac/	Startseite des „Meta-Opac Berlin-Brandenburg“	DB Entry	av	5	2952	2490	8677	0,21	0,18	0,61	14119
4	/~is/computerkurs/ms-dos.html	Computer-tutorial zum Thema “MS-DOS”	Docu	av	3	10710	1745	43	0,86	0,14	0,00	12498
5	/~rfunk/lv/scripts/bwl/bwl.html	Vorlesungsscript zum Thema „Betriebswirtschaftslehre“	Text	lg	4	7530	1656	719	0,76	0,17	0,07	9905

Tabelle 5-4: Ausschnitt Listeneinträge der „Top 100-Liste“ (2002)

Insgesamt entfallen im Jahr 2002 auf die Top 100 Webseiten insgesamt 382.462 Entries, das entspricht in etwa 46,6% aller Siteentries auf die gesamte Website. Im Jahr 2002 konnten Zugriffe auf insgesamt 11.853 verschiedene Webseiten (Format HTML) im Log festgestellt werden. Damit erhalten die „wichtigsten“ 100 Seiten, dass entspricht ca. 8 Promille der gesamten IB-Website, knapp die Hälfte aller Siteentries. Das bedeutet das jede Seite der Top 100 durchschnittlich 10 Entries pro Tag erhält[46]. Im Jahr 2000 gab es lediglich Zugriffe auf 6.246 verschiedene Webseiten. Damit hat sich das Webangebot des IB-Webservers im Zeitraum Januar 2000 bis Dezember 2002 fast verdoppelt. Die Verteilung der Navigationsarten auf die Top 100 Webseiten entspricht ungefähr den Anteilen der Gesamtsite. Der Anteil der Navigationsart Referenz liegt bei den Top 100 mit 10% etwas höher (3%) als der Anteil der „Referenzen“ der Gesamtsite.
Die Verteilung der absoluten Entries (Entry total) der Top 100 Webseiten nähert sich in beiden Jahren einem klassischen Power Law an (siehe Abbildung 5-9, Rsq (2002) = 0,961; Rsq (2000) = 0,847).

[46] Beispiel: Die IB-Homepage als wichtigste Einstiegsseite erhält auf das gesamte Jahr 2002 gesehen 184 Entries pro Tag. Für die Webseite mit dem Rang 100 wäre es lediglich 1 Entry pro Tag.

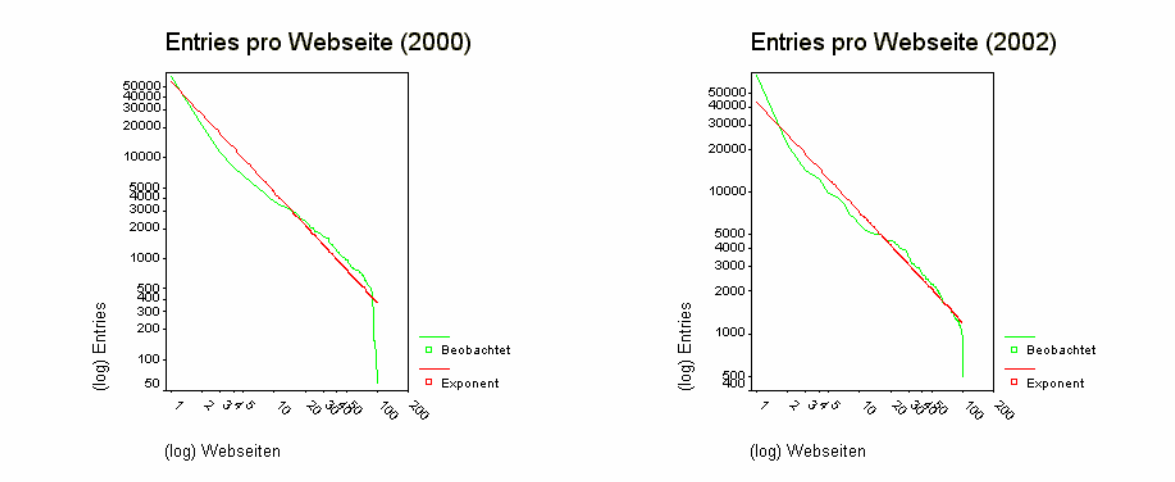


Abbildung 5-9: Verteilung der Siteentries für die Top 100 Webseiten (2000, 2002)[47]

Die Analyse der einzelnen Kategorien bzw. Spalten der „Top 100"-Liste des Untersuchungs-zeitraums Januar bis Dezember 2002 ergibt folgendes Bild (vgl. Tabelle 5-5)[48]. Der Großteil der Webseiten aus der „Top 100"-Liste gehört zu der Kategorie „Text". Die 58 „Text-Seiten" generieren den größten Teil der Siteentries. Insgesamt entfallen auf die „Text-Seiten" 161.548 Entries, das entspricht einem Durchschnitt von 2.785 Entries pro „Text-Seite". Der zweithäufigste Webseitentyp trägt die Bezeichnung Orga (Kurzform für Organizational). Die 21 Seiten erreichen im Schnitt mit 2.470 Entries vergleichbar viele Entries wie die Seiten der Text-Kategorie. Die Ergebnisse der drei übrigen Inhaltskategorien Home, Docu und DB_Entry (Kurzformen für Homepage, Documentation und Database Entry), die in der „Top 100"-Liste allerdings schwächer vertreten sind, generieren im Schnitt deutlich mehr Entries. „Dokumentations-Webseiten" (siehe Inhaltstyp Docu) liegen dabei mit 10.499 Entries vor den „Homepage-Webseiten" (siehe Inhaltstyp Home), die im Schnitt 10.233 Entries (inkl. IB Homepage) erreichen. Die sechs „DB_Entry-Webseiten" der IB-Site erhalten im Untersuchungszeitraum 2002 mit durchschnittlich 3.684 Entries deutlich mehr Entries als die textbasierten Kategorien „Text" und „Orga". Die beiden weiteren Inhaltskategorien Multimedia und Tool, die *Haas & Grams* in ihrer Websitekategorisierung (*Haas & Grams*, 2000 [41]) erarbeitet haben, finden sich nicht in der „Top 100"-Liste der IB-Site.

[47] Hinweis: doppelt logarithmischer Maßstab (siehe log Webseiten und log Entries total)

[48] Die Daten der einzelnen Kategorien werden z.T. dadurch beeinflusst, dass sich in den einzelnen Klassen die hohen Werte der IB-Homepage befinden (siehe Top 1). Da diese Seite aufgrund ihrer Position innerhalb der Site bzw. der Anzahl der Entries eine Ausnahme darstellt, werden in der untenstehende Statistiktabelle für die betroffenen Klassen die Werte der einzelnen Klassen separat in Klammern abzüglich der Werte der IB-Homepage aufgeführt.

Kategorie	Anzahl der Seiten	Total Entries	∅ Entries pro Webseite
Kategorien nach *Haas & Grams* [41]			
Text	58	161.548	2.785
Orga	21	51.875	2.470
Home	10 ohne IB Home	102.338 (35.066)	10.233 (3.506)
Docu	4	41.997	10.499
DB_Entry	6	22.108	3.684
Seitengröße[49]			
groß (lg)	42	124.627	2.967
mittel (av)	46 ohne IB Home	236.225 (168.953)	5.135 (3.672)
klein (sm)	11	19.014	1.728
PageRank-Wert			
PR 6	2 ohne IB Home	69.954 (2.682)	34.977 (2.682)
PR 5	19	71.124	3.743
PR 4	47	167.951	3.573
PR 3	29	67.981	2.344
PR 2	2	2.856	1.428
Navigationsart			
Suchmaschine	100	224.720	2.247
Direkt	100	120.586	1.205
Referenz	100	37.297	372

Tabelle 5-5: Statistik der „Top 100-Liste“ (2002)

Die Ergebnisse bzgl. der Seitengrößen der einzelnen Webseiten ergeben folgendes Bild. Von den 99 klassifizierten Seiten stammen 88% aus der Gruppe der mittleren (46 Seiten) bis großen (42 Seiten) Webseiten. Die großen Webseiten stellen dabei eine besonders homogene Gruppe dar. Von den 42 großen Webseiten sind lediglich vier Seiten nicht aus der Kategorie „Text“. Die kleinen Webseiten mit 1 bis 5180 Byte Seitengröße (inkl. HTML-Code) kommen mit lediglich elf Seiten deutlich seltener in der „Top 100“-Liste vor und erhalten durchschnittlich die geringsten Entries (1.728) pro Seite.
Bei der Betrachtung der Verteilung der PageRank-Werte fällt auf, dass die Seiten mit den höchsten PageRank-Werten (PR 6) im Durchschnitt auch die höchste Anzahl an Siteentries pro Webseite erhalten (vgl. Abb. 5-10). Es besteht somit ein sichtbar positiver Zusammenhang (nichtparametrische Korrelation Spearman = 1,0) zwischen dem externen Parameter PageRank des Suchmaschinen-herstellers Google und den durchschnittlichen Siteentries pro Webseite, die das Logfile des Webservers ausweist. Die Abbildung 5-10 zeigt deutlich, dass mit abnehmenden PageRank-Werten (PR 6, 5, 4, 3, 2) die Anzahl der durchschnittlichen Siteentries drastisch sinken (Hinweis: logarithmischer Maßstab, y-Achse). Auffällig ist weiterhin, dass beinahe die Hälfte der Webseiten (47 Seiten) einen PageRank-Wert von 4 erhält (Median der PageRank-Werte = 4).

[49] Die Seitengrößen der untersuchten Webseiten setzen sich aus den drei Kategorien groß (lg = > 43329 byte), mittel (av = > 5182 < 43329 byte) und klein (sm = > 1 < 5182 byte) zusammen (vgl. [4.3]).

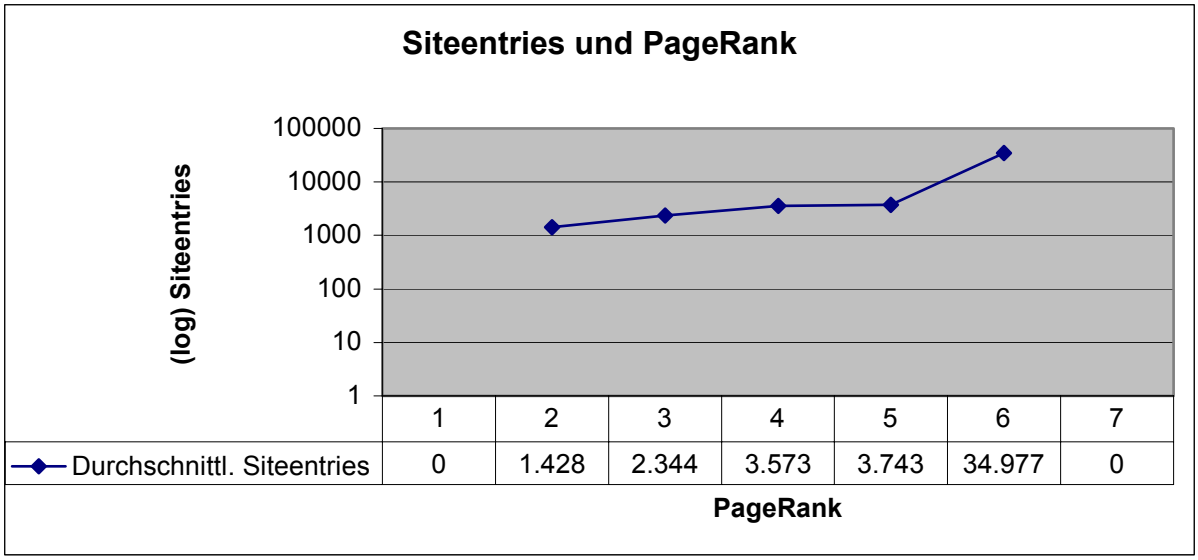


Abbildung 5-10: durchschnittliche Siteentries und PageRank (2002)[50]

Die durchschnittliche Anzahl der Siteentries bezogen auf die drei Navigationsarten (Suchmaschine, Direkt und Referenz) verdeutlichen den Trend, dass über Suchmaschinen-Queries die meisten Siteentries auf die IB-Site erfolgen. Durchschnittlich erhält jede Webseite 2.247 Siteentries über die Navigationsart Suchmaschine. Direkte Navigation spielt mit durchschnittlich 1.205 Siteentries die zweit wichtigste Navigationsart. Siteentries über „Referenzen“ kommen an dritter Stelle. Jede Webseite erhält durchschnittlich lediglich 372 Entries durch diese Navigationsart. Die untere Abbildung 5-11 verdeutlicht diese Ergebnisse. Auf der X-Achse werden die 100 Webseiten (Top 100) aufgetragen. Die Y-Achse (logarithmisierte Skala) stellt die absoluten Siteentry-Werte (jeweils sortiert nach den höchsten Werten) für die drei Navigationsarten dar. Deutlich wird bei dieser Darstellung, dass die Navigationsart Suchmaschine die anderen beiden Navigationsarten bis auf wenige Ausnahmen dominiert. Die detaillierte Betrachtung der nach dem Rang sortierten Verteilung der Siteentry-Werte pro Webseite zeigt, dass unter den 100 Seiten der „Top 100-Liste“ 77 Webseiten über 1000 Siteentries über die Navigationsart „Suchmaschine“ erhalten. Für die beiden anderen Navigationsarten „direkte“ Navigation und „Referenz“ sind das deutlich weniger. Lediglich 15 Webseiten erhalten mehr als 1000 Siteentries über das „direkte“ Navigation. Für die Navigationsart „Referenzen“ sind es sogar nur 6 Webseiten (58 Webseiten erhalten sogar weniger als 100 Siteentries).

[50] Hinweis: logarithmischer Maßstab (siehe log Siteentries)

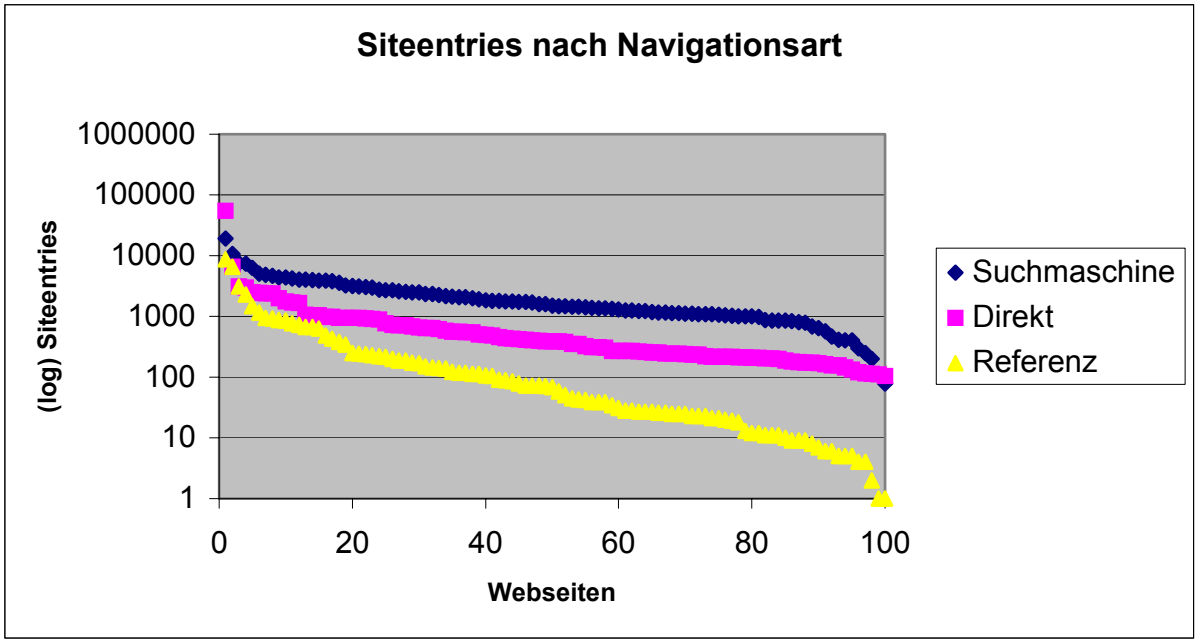
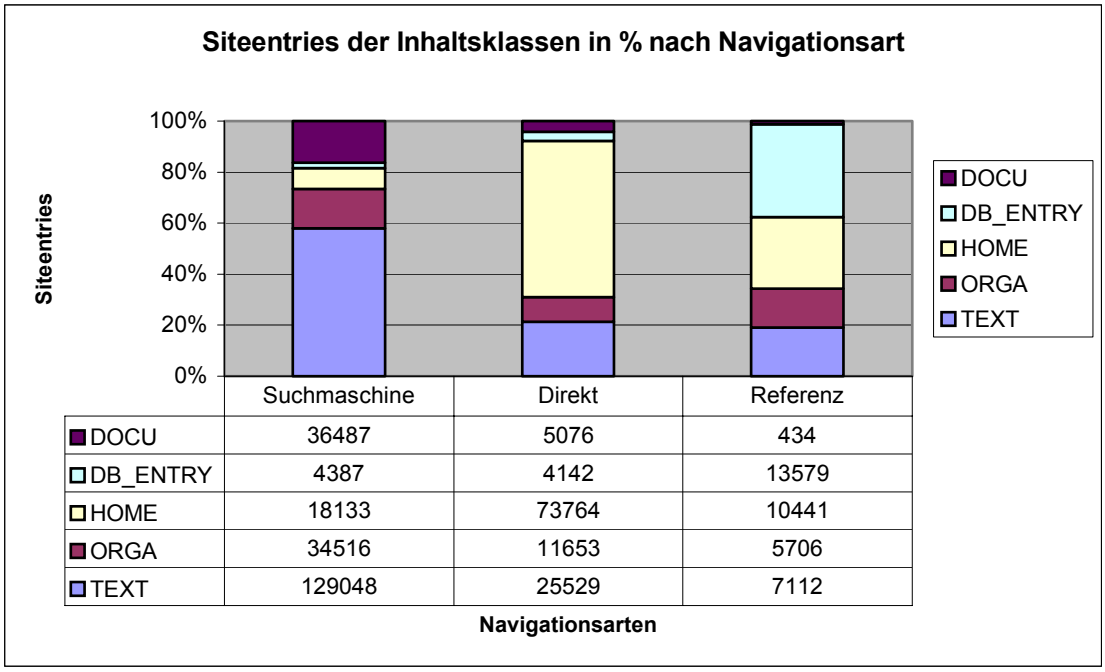


Abbildung 5-11: Top 100-Webseiten mit Entry-Werten nach Navigationsart (2002)[51]

Weitere interessante Accessibility-Tendenzen bezüglich der Seitenkategorien werden aus der unteren Abbildung deutlich. Die Kombination der Siteentries für die Navigationsart und den Inhaltstyp geben Auskunft darüber, wie stark die drei Navigationsarten bei den einzelnen Inhaltstypen der „Top 100"-Liste enthalten sind. Verkürzt gesagt: die untere Abbildung 5-12 macht deutlich, welcher Inhaltstyp bei welcher Navigationsart die wichtigste Rolle spielt.

Bei der Navigationsart Suchmaschine dominiert mit knapp 60% die Inhaltsklasse „Text" (siehe linker Balken, Abb. 5-12). Die weiteren „textorientierten" Inhaltsklassen „Docu" und „Orga" folgen mit jeweils etwas über 15%. Die Inhaltsklassen „Home" und „DB_Entry" erreichen über die Navigationsart Suchmaschine den geringsten Anteil an Siteentries. Die Verteilung der Inhaltsklassen bezogen auf die „direkten" Siteentries ergeben, dass die verschiedenen Homepages inkl. Zugriffe auf die IB-Homepage (siehe Inhaltsklasse „Home") den größten Anteil der Siteentries (ca. 61%) über die Navigationsart „direkte" Navigation erhalten. Bis auf die „Text-Seiten", die ca. 21% der Siteentries generieren, erreichen die übrigen Inhaltsklassen keine nennenswerten Ergebnisse. Die letzte untersuchte Navigationsart Referenz zeigt eine Verschiebung der Anteile der fünf analysierten Inhaltsklassen. Den Hauptanteil (36%) der „referenzierten" Siteentries („Backlinks") erzielt die relativ „kleine" (insgesamt sechs Webseiten) Inhaltsklasse „DB_Entry". Den zweithöchsten Anteil (28%) erreicht hier die „Homepage-Klasse". Die folgenden Positionen nehmen „Text", „Orga" und „Docu-Seiten" ein.

[51] Hinweis: logarithmischer Maßstab (siehe log Siteentries)



	Suchmaschine	Direkt	Referenz
DOCU	36487	5076	434
DB_ENTRY	4387	4142	13579
HOME	18133	73764	10441
ORGA	34516	11653	5706
TEXT	129048	25529	7112

Abbildung 5-12: Siteentries der Inhaltsklassen in % nach Navigationsart (2002)

Folgende Navigationstrends werden mit Hilfe der Inhaltskategorisierung der Webseiten der Top 100 deutlich.

- „Textseiten“ (n = 58) werden vorwiegend über Suchmaschinen gefunden; wenige „Textseiten“ schaffen es, durch ihre „Originalität“ „gebookmarkt“ oder „verlinkt“ zu werden. Da sich unter den hier analysierten „Textseiten“ auch eine ganze Reihe Seiten befinden, die Printveröffentlichungen (Artikeln) entsprechen (z.B. „Rehm-Texte“, „Handreichungen“, „Zahn- und Umstätter-Artikel“), ließe sich die Hypothese aufstellen, dass es auch im Web nur sehr wenige „Texte“ (Hypertexte) gibt, die es schaffen, von anderen Autoren „referenziert“ und dann auch wahrgenommen (genutzt) zu werden. Dafür werden diese ansonsten unsichtbaren Texte auf Webservern durch Suchmaschinen unter Umständen sehr sichtbar (vgl. *Lawrence*, 2001 [57]). Ein Grund für die hohe Sichtbarkeit der „Textseiten“ durch Suchmaschinen sind die größeren Textanteile, die die „Textseiten“ auszeichnen. Dies zeigt sich auch daran, dass die analysierten „Textseiten“ fast alle zu den „großen“ Seiten, bezogen auf die Bytezahl, gehören.
- Die „Orga-Seiten“ (n = 21) sind in der Regel Seiten mit vergleichsweise vielen internen und auch externen Links („Outlinks“). Sie sind mit weniger Text (vgl. Seitengrößen) ausgestattet als die „Textseiten“ und dienen vielfach der internen Navigation. Die meisten Entries erreichen die „Orga-Seiten“ auch durch Suchmaschinen. Auffälligerweise sind sie besser referenziert als die „Textseiten“. Grund dafür könnte sein, dass die „Orga-Seiten“ Information in komprimierter Form (Links und weniger Text) enthalten und daher für Link-Autoren anscheinend bessere Linkziele sind.
- „Homepages“ (n = 10) werden „direkt“ aufgerufen; dafür sind sie auch konstruiert. Hier sollen Websitebesuche idealerweise starten. In der Regel haben sie die kürzeste Adresse (URL), die sich der Besucher merkt und beim Folgebesuch „direkt“ eingibt. Das scheint auch bei der IB-Seite der Fall zu sein. Die zehn Homepages der Top 100 erhalten den Großteil ihrer Entries über

„direkte“ Eingaben. Der andere wichtige Zugangsweg sind „Referenzen“. Die Homepage des IB führt die Liste der Homepages (und ebenfalls der Top 100) aufgrund ihrer Ausnahmeposition und Aufgabe mit großem Abstand an. Suchmaschinen-Entries spielen für die IB-Homepage eine etwa genauso große Bedeutung wie die Referenzen. Ansonsten sind die Homepages durchschnittlich groß.

- Ein interessantes Ergebnis liefern die Entry-Zahlen für die Klasse „DB_Entry“ (n = 6). Die Startseiten zu Datenbankangeboten werden vier mal so häufig über Referenzen gestartet als über die übrigen Navigationsarten. Aufgrund der speziellen Thematiken ihrer Angebote und dem geringen Textanteil (im Gegensatz zu den „Textseiten“ gehören die „DB_Entry-Seiten“ zu den kleinsten Seiten der Top 100) finden nur wenige Suchmaschinenbesucher diese „Datenbank-Seiten“. Der geringe Anteil an „direkten“ Entries (über Bookmarks) auf „Datenbank-Seiten“ verwundert, da es sich bei den Angeboten um originelle, z.T. einmalige Angebote handelt.
- Überraschend sind auch die äußerst geringen Entry-Werte für die Klasse „Docu“ (n = 4) auf Seiten der „Referenzen“ und „direkten“ Entries. Obwohl die vier „Docu-Seiten“ so häufig über Suchmaschinen gefunden werden, erhalten sie praktisch keine Entries über „Referenzen“. Die Seiten sind also trotz vieler Suchmaschinenbesuche nicht verlinkt worden. Dieses Ergebnis verwundert, weil gerade angenommen wird, dass Dokumentationen bzw. Tutorials verlinkt oder „gebookmarkt“ werden.

5.7 Besondere Webseiten unter den Top 100

Unter den Webseiten der „Top 100-Liste“ lassen sich folgende interessante Seitengruppen identifizieren (vgl. Anhang [10.1]).

- Texte von Margarete Rehm (n = 13)
- Berliner Handreichungen zur Bibliothekswissenschaft (n = 14)
- div. Link-, Adress-, und Literaturlisten (n = 17)
- Ausreißer und Anomalien

Die dreizehn Texte von „Margarete Rehm“ sowie die vierzehn Texte der „Berliner Handreichungen zur Bibliothekswissenschaft“ fallen zum einen aus dem Rahmen, weil sie bzgl. der Struktur und des Textumfangs eine sehr homogene Gruppe innerhalb der Top 100 bilden. Zum anderen sind die Anteile der drei Navigationsarten für beide Seitengruppen auffällig ähnlich. Die einzelnen Seiten dieser Gruppen zeichnet ein sehr hoher Suchmaschinenanteil (im Falle der „Rehm-Texte“ sogar zwischen 80-90 Prozent) aus. Ein viel heterogeneres Bild zeigt sich bei der Gruppe der „Listen-Seiten“ (siebzehn Webseiten), die alle zur Inhaltskategorie „Orga“ gehören. Für die einzelnen Webseiten dieser Gruppe lässt sich nicht vereinheitlichend sagen, warum (über welche Navigationsart) sie in die „Top 100-Liste“ gekommen sind. Einige der Seiten zeigen sehr hohe Entry-Werte über die

Navigationsart „Suchmaschine“, andere werden bevorzugt „direkt“ gestartet und wiederum andere erhalten sehr hohe „Navigationszahlen“ über „Referenzen“.
Ein Beispiel für eine „Ausreißer-Seite“ wäre die Webseite mit dem Rang 2 in der „Top 100-Liste“. Diese Seite stellt den „ASCII-Codes“ dar und erhält in Untersuchungszeitraum 19.248 Entries über die Navigationsart „Suchmaschine“. Grund für den hohen Platz in der Rangliste ist, dass die Seite zum entsprechenden Suchbegriff „Ascii-Code“ bzw. weiteren Permutationen dieses Begriffs bei Google an vorderster Stelle gelistet wird (Seite 1, Position 1-10).
Beim Vergleich der „Top 100-Listen“ des Jahres 2000 und 2002 fällt auf, dass Dreiviertel der Seiten (n = 75), in beiden Jahren unter den am häufigsten genutzten Einstiegsseiten sind. Ein Viertel der Webseiten (n = 25) ist 2002 neu unter die ersten 100 Einstiegsseiten gekommen. Tendenziell hat sich eher der mittlere und untere Bereich der „Top 100-Liste“ im Jahr 2002 verändert. Der Vergleich der Listen bzgl. der Zugehörigkeit der Webseiten zu einer Inhaltsklasse zeigt, dass sich die Zusammensetzung der Inhaltsklassen nicht nennenswert verändert hat.

5.8 Ergebnisse der Web Entry Faktoren Analyse[52]

Die untere Abbildung 5-13 zeigt die Verteilung der drei Web Entry Faktoren für die Top 100 Einstiegswebseiten der IB-Site (vgl. WEF-Werte in der „Top 100-Liste“, siehe Anhang [10.1]). Die Abbildung verdeutlicht folgende an einigen Stellen bereits erwähnte Tendenzen der Loganalyse.

- Die Anteile der Navigationsart „Suchmaschine“ übertrifft die anderen beiden Navigationsarten für die Top 100 Webseiten deutlich (vgl. Mediane der „WEF- Suchmaschinenwerte“ = 0,81, der „WEF-Direktwerte“ = 0,15, der „WEF-Referenzwerte“ = 0,02)
- Bei allen drei Navigationsarten gibt es Ausreißer. Diese „Ausreißer-Seiten“ zeigen starke Abweichungen bzgl. der WEF-Werte. Diese lassen sich auf der Abbildung 5-13 gut erkennen.
- Die Verteilung der WEF-Werte für die Top 100 zeigt einen positiven Trend bzgl. der Suchmaschinen-WEF’s. Die Trendlinien (siehe Abb. 5-13) für die beiden anderen Navigationsarten haben ein negatives Vorzeichen.
- Die untenstehende Abbildung ermöglicht die Identifikation der „Website-Authorities“[53] der analysierten Website.

[52] Jede untersuchte URL erhält drei Entry-Werte (Suchmaschine, Direkt & Referenz, vgl. „Top 100-Liste“). Die Summe der drei Werte ergibt die Gesamtanzahl an Entries für die URL. Die drei WEF’s (für die drei Navigationsarten) werden für jede URL (bzw. Directory-Urls) der Top 100 errechnet, indem jeder der drei Entry-Werte einer URL durch die Gesamtanzahl an Entries für diese URL dividiert wird. Damit ergeben sich pro URL drei Anteilswerte zwischen 0 und 1. Diese Anteilswerte werden Web Entry Faktoren genannt.
[53] Im Zusammenhang mit den Web Entry Faktoren sind „Website-Authorities“ Webseiten, die über sehr hohe WEF-Werte für die Navigation „Referenz“ verfügen (wef r).

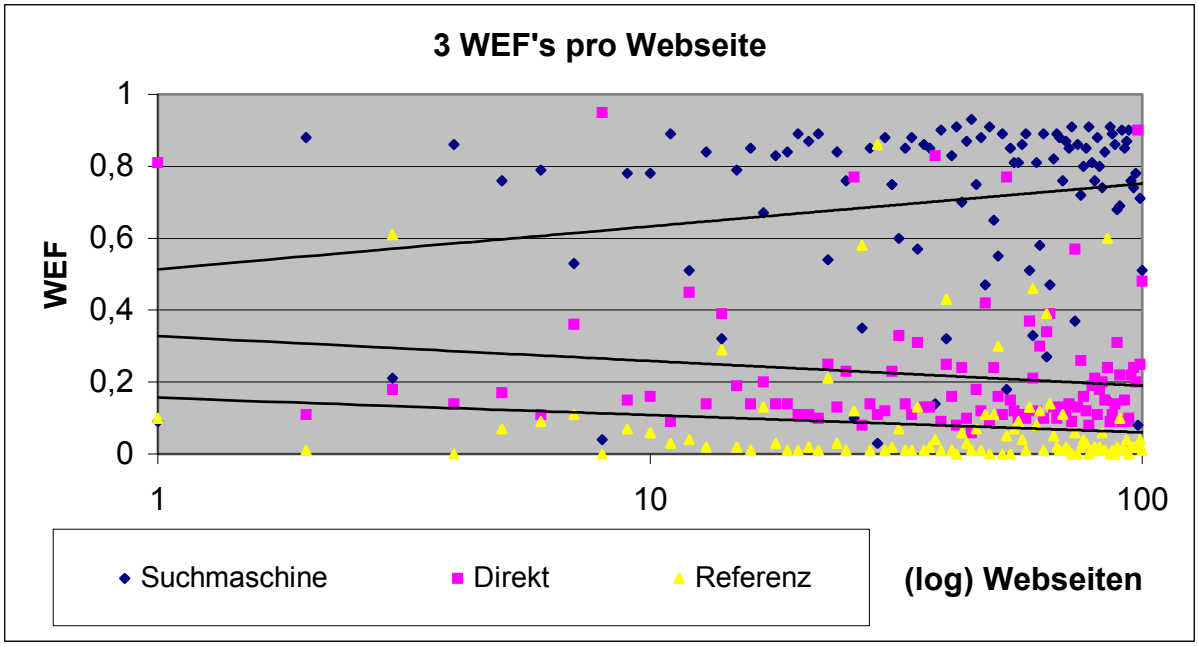


Abbildung 5-13: 3 WEF-Werte für die Top 100 nach Navigationsart (2002)[54]

Die Verteilungen der WEF-Werte für die einzelnen Inhaltskategorien wird auf den unteren Abbildungen deutlich. Auf der X-Achse der folgenden Abbildungen werden in logarithmischem Maßstab die Webseiten der jeweiligen Inhaltskategorien dargestellt. Auf der Y-Achse werden die drei WEF-Werte für jede Seite aufgetragen. Die ersten drei Abbildungen zeigen, dass die „textbasierten" Inhaltskategorien (Text, Orga und Docu) sehr viel häufiger über Suchmaschinen gestartet werden, als die übrigen Inhaltskategorien „DB_Entry" und „Home". Die „Textseiten" und „Docu-Seiten" (Text, Docu) zeigen diesen Trend am deutlichsten. Das gilt in besonderem Maße auch für die „großen" Webseiten (Webseiten mit über 43.000 Byte Dateigröße) zu denen hauptsächlich die Inhaltskategorie „Text" zu zählen ist. Unter den „Orga-Seiten" befinden sich auch einige Webseiten die gut über Backlinks bzw. Bookmarks besucht sind[55]. Die Webseiten der Kategorien „Home" und „DB_Entry" zeigen WEF-Verteilungen mit Tendenzen zu den anderen beiden Navigationsarten. Auf den unteren Abbildungen lassen sich einzelne Ausreißer identifizieren (siehe dazu auch Abschnitt „WEF-Szenarien").

[54] Hinweis: logarithmischer Maßstab (siehe log Webseiten)
[55] Die „Suchmaschinenseite" des IB kann beispielsweise als „Hub-Page" bezeichnet werden, da sie auf Authorities (in den Fall die großen Suchmaschinen) linkt.

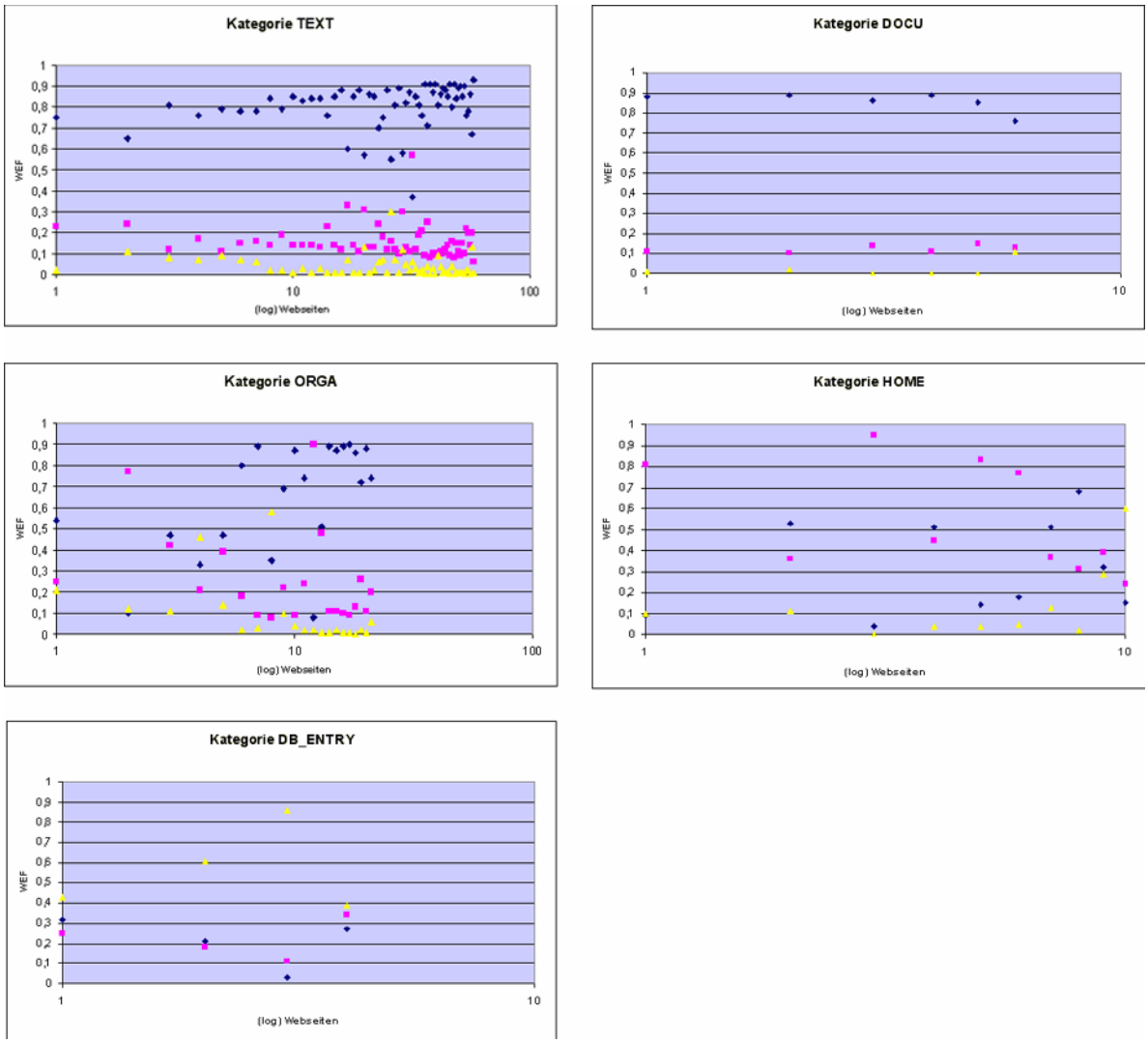


Abbildung 5-14: Verteilung der WEF-Werte der Navigationsarten nach Inhaltskategorien (2002)[56]

5.9 WEF-Faktoren und PageRank

Bei dem Vergleich der gemittelten PageRank-Werte mit den ebenfalls gemittelten WEF-Werten für die verschiedenen Inhaltsklassen fällt auf, dass insbesondere die Zahlenreihen der PageRank-Werte mit den Werten der „Referenz-WEFs“ (WEF_R) positiv korrelieren (vgl. Tab. 5-6). Die Inhaltsklassen, die höhere PageRank-Werte erreichen, erhalten ebenfalls höhere WEF_R-Werte. Besonders auffällig ist hierbei die Inhaltsklasse DB_ENTRY. Das Beispiel DB_ENTRY zeigt, dass stärkere Verlinkung (höherer PageRank-Wert) zufolge hat, dass diese Angebote stärker über „Referenzen“ zugänglich werden. Die „textbasierten“ Inhaltsklassen ORGA, TEXT und DOCU, die hauptsächlich über Suchmaschinen (hohe WEF_S-Werte) zugänglich sind und kaum verlinkt sind, zeigen die niedrigsten PageRank-Werte. Die Korrelation der Zahlenreihe PageRank mit den WEF_R-Werten bzw. WEF_D-Werten ergeben jeweils einen positiven Spearman Korrelationskoeffizienten (1,0 bzw. 0,7). Die PageRank-Werte korrelieren negativ mit den WEF_S-Werten (negativer Spearman Korrelationskoeffizient -1,0).

Inhaltsklasse	PageRank	WEF_R	WEF_D	WEF_S
DB_ENTRY	5,25	0,57	0,22	0,21
HOME	4,90	0,14	0,55 (0,52)	0,32
ORGA	3,90	0,10	0,26	0,65
TEXT	3,71	0,04	0,16	0,80
DOCU	3,17	0,02	0,12	0,86

Tabelle 5-6: Mittelwerte der PageRank und WEF Werte pro Inhaltsklasse (2002)

5.10 WEF-Szenarien

Nachfolgend sollen einzelne WEF-Szenarien vorgestellt und anhand der Ergebnisse der Extraktion beschrieben werden.

Einfache Filter:

- Filter „Suchmaschinen-WEF“: hohe WEF_S-Werte (75%-Quartil = 0,87) - Ergebnis: ausschließlich TEXT-Seiten aus dem mittleren bis unteren Listenbereich. Dieses Ergebnis spricht dafür, dass weniger exponierte Webseiten mit einfachen Textinhalten nahezu nur noch durch Suchmaschinen zugänglich sind.
- Filter „Direkt-WEF“: hohe WEF_D-Werte (75%-Quartil = 0,24) - Ergebnis: HOME und ORGA-Seiten aus allen Bereichen der Liste. Dieses Ergebnis spricht dafür, dass Bookmarks hauptsächlich auf Webseiten der Klassen HOME und ORGA gesetzt werden. Für die Klasse HOME ist dieses Ergebnis auch nicht weiter verwunderlich, überraschender ist, dass auch ORGA-Seiten als Startseiten genutzt werden. Weiterhin verwundert es, dass die DB_ENTRY-Seiten, die ja auch klassische Startseiten sind, nicht in dem Maße über Bookmarks genutzt werden, wie Homepages und Orga-Seiten.
- Filter „Referenz-WEF“: hohe WEF_R-Werte (75%-Quartil = 0,07) - Ergebnis: DB_ENTRY und HOME, ORGA-Seiten. Beim Filtern nach besonders hohen WEF_R-Werten zeigt sich, dass insbesondere die DB_ENTRY-Seiten sehr viele Entries über externe Referenzen erhalten. Einzelne Homepages und ORGA-Seiten kommen auch in diesen Bereich von über 50 % Siteentries über Referenzen (Backlinks).

Die in der unteren Tabelle 5-7 skizzierten WEF-Szenarien (kombinierte Filter) verdeutlichen die bereits oben erwähnten Tendenzen. Die Anzahl der Seiten der jeweiligen Inhaltsklassen befindet sich in Klammern (siehe rechte Spalte).

[56] Hinweis: logarithmischer Maßstab (siehe log Webseiten)

Szenario	WEF_S	WEF_D	WEF_R	Beschreibung	Ergebnisse d. Extraktion
1	hoch	hoch	niedrig	Seiten mit vielen Entries über Suchmaschinen und „direkten“ Eingaben	TEXT (6), ORGA (1)
2	hoch	niedrig	niedrig	Seiten mit hohen Suchmaschinen-Entries	TEXT (6), DOCU (4)
3	hoch	niedrig	hoch	„gut verlinkte Texte“	TEXT (2), DOCU (1)
4	niedrig	-	-	kaum Siteentries über Suchmaschinen	DB_Entry (3), HOME (2)
5	niedrig	hoch	hoch	Authorities[57]	DB_Entry (2), HOME (2)
6	median	median	median	Alle drei Entry-Werte sind durchschnittlich	TEXT (2), ORGA (1)

Tabelle 5-7: Beispiele und Ergebnisse für WEF-Szenarien (2002)

Die folgenden Ergebnispunkte stehen zwar nicht in direktem Zusammenhang zu der vorgestellten Logmetrik WEF, sie werden aber trotzdem in knapper Form angeführt, weil sie weitere interessante Hinweise und Erweiterungen zum Navigationsverhalten im Web bzw. der Analyse des Webuse von akademischen Webangeboten liefern.

5.11 Suchmaschinen Details

Die beiden unteren Abbildungen zeigen die Entwicklung der Siteentries der Navigationsart „Suchmaschine“, unterteilt nach den zehn wichtigsten kommerziellen Suchmaschinen sowie der Sitesuche der Humboldt-Universität (HU Search)[58] und der deutschen Metasuchmaschine Metager. Auf der Y-Achse der Abbildungen sind jeweils die Siteentries für jede Suchmaschine über die 12 Monate dargestellt. Die erste Abbildung zeigt den 12-Monatsverlauf der Suchmaschinen für das Jahr 2000. Hier wird deutlich, dass die „großen“ internationalen Suchmaschinen wie Altavista, Lycos oder Yahoo im Jahresverlauf deutlich durch das starke Wachstum der Siteentries über Google (siehe großes blaues Dreieck und Trendlinie, Abbildung 5-15) zurückgedrängt wurden[59]. Google zieht ab Oktober 2000 an seinen Konkurrenten vorbei. Insgesamt fällt auf, dass die Siteentries für die meisten

[57] Authorities können über die Filterprozeduren für alle Inhaltstypen ausgegeben werden. Der Filter wird so eingestellt, dass nach den Seiten einer Klasse gefiltert wird, die die höchsten WEF_R bzw. WEF_D-Werte haben (siehe 75%-Quartil). So lassen sich beispielsweise auch Authorities unter den „Textseiten“ identifizieren.

[58] Die Sitesuche der Humboldt-Universität besteht aus einer Suchmaschine, die in regelmäßigen Abständen die Webserver der Institute und Fakultäten der Humboldt-Universität indexiert und über eine gemeinsame Oberfläche recherchierbar macht.

[59] Google wurde 1999 gegründet und konnte im Jahresverlauf 2000 seinen Rückstand zu den bestehenden großen Suchmaschinen aufholen. Google ist spätestens seit Ende 2001 die dominierende Suchmaschine. Diese Dominanz zeigt sich zum einen aufgrund der Anzahl der indexierten Seiten (vgl. Search Engine Watch) als auch der Anzahl der Benutzer (vgl. Google’s Angaben, google.com).

Suchmaschinen im Sommer (Juli, August, September) sichtbar ansteigen. Die vorrangig deutschsprachigen Suchmaschinen wie Fireball, Web.de und Metager zeigen im Jahresverlauf 2000 im Vergleich zu den „großen“ internationalen Suchmaschinen relativ hohe und stabile Siteentry-Werte. Die Sitesuche (siehe HU Search) der Humboldt-Universität spielt eine ähnlich geringe Rolle wie die übrigen internationalen Suchmaschinen InfoSeek, Excite, MSN und AlltheWeb die alle weniger als 1.000 Siteentries pro Monat generieren. Eine ganz andere Situation stellt sich im Jahr 2002 dar. Die Abbildung der 12-Monatsentwicklung der Siteentries für die einzelnen Suchmaschinen zeigt sehr deutlich, dass Google seine Konkurrenten klar distanziert hat und den Großteil der Besuche auf die IB-Site vermittelt. Google ist 2002 mit deutlichem Abstand die größte, beliebteste und meist genutzte Suchmaschine mit nach wie vor steigender Tendenz (siehe Trendlinie, Abbildung unten). Der Vergleich der Siteentry-Werte der Suchmaschinen der Jahres 2002 gegenüber 2000 zeigt die Gewinne bzw. Verluste (Siteentries) der großen Suchmaschinen.

- Google (+334.486)
- MSN (+4.559)
- HU Search (+3.203)
- Altavista (-37.878)
- Lycos (-26.729)
- Fireball (-23.031)
- Metager (-15.225)
- Yahoo (-10.353)
- Web.de (-6.727)
- AlltheWeb (-1.718)

Google liefert damit 2002 rund zwölf Mal (+334.486) so viele Siteentries wie im Jahr 2000 (30.928) und rund vier mal soviel wie alle anderen Suchmaschinen zusammen. Die großen Verlierer sind Altavista, Lycos, Fireball etc. Lediglich die Microsoft Internetsuchmaschine MSN und die HU-Sitesuche (HU SEARCH) können die Anzahl der vermittelten Siteentries steigern. Das Wachstum der Siteentries über die Suchmaschinen Google hat sich 2002 deutlich verringert (vgl. Trendlinie 2002, Abb. 5-15).

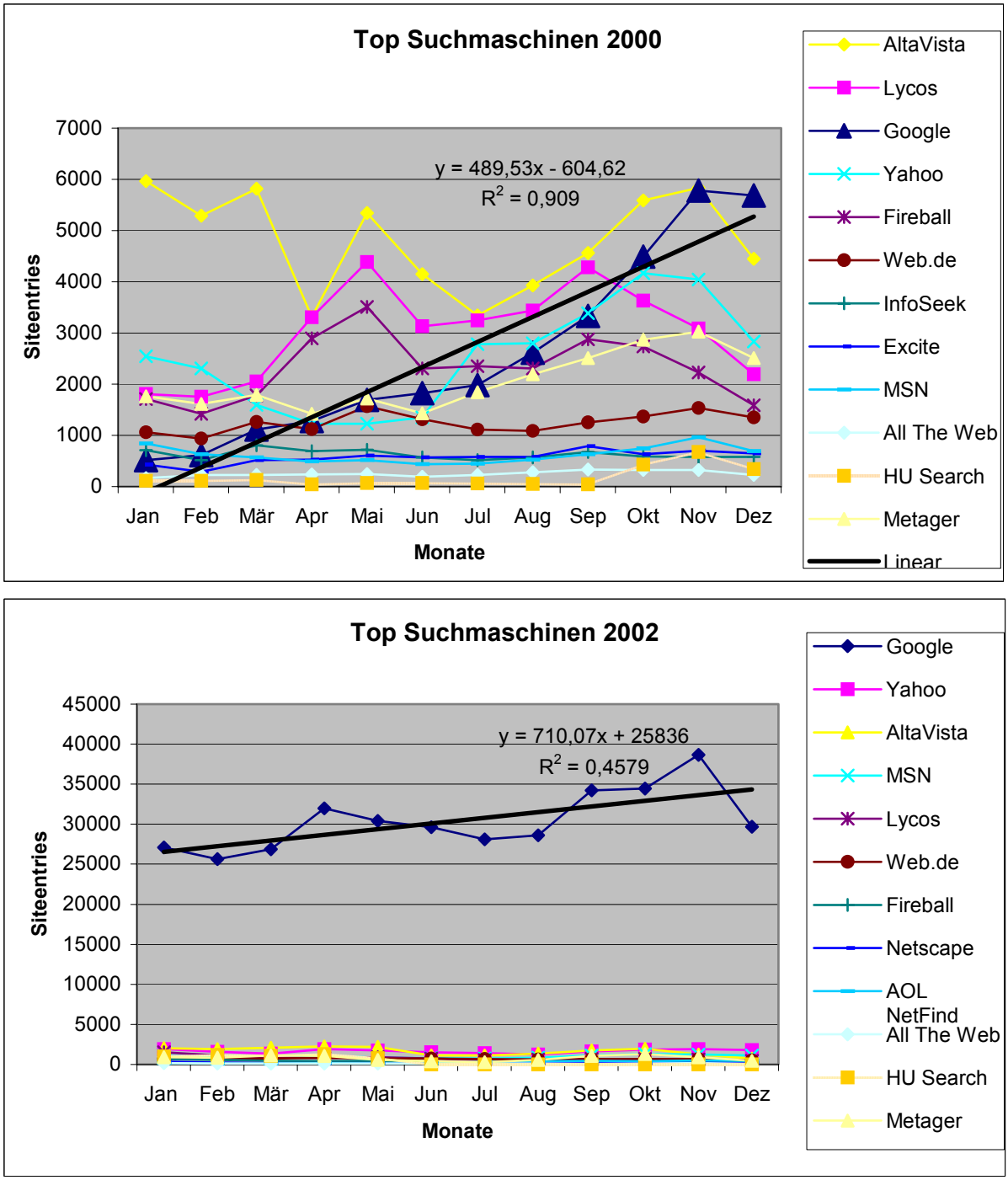


Abbildung 5-15: 12-Monatsverlauf der Top-Suchmaschinen (2000, 2002)

5.11.1 Analyse der Queries

Die Analyse der Suchmaschinen-Queries bietet weitere Potenziale (vgl. *Jansen & Pooch*, 2001 [47], *Cothey*, 2002 [22]). Die Extraktion der einzelnen Queries aus dem Logfile ergab das im Jahr 2000 7.285 verschiedene Queries (von insgesamt 253.044) im Log waren (vgl. *Ross & Wolfram*, 2000 [80]). Im Jahr 2002 waren es rund 3,5 mal so viele verschiedene Suchanfragen (24.888 unterschiedliche Queries von insgesamt 485.963) (vgl. Top 100 Queries im Anhang [10.2]). Die Abbildung 5-16 stellt die Verteilung der unterschiedlichen Suchmaschinen-Queries bzgl. der Häufigkeit

ihres Vorkommens im Log dar. Die Verteilungen der 100 Queries für den Monat April (2002, 2000) nähern sich an ein klassisches Power Law an (Rsq (2002) = 0,99, Rsq (2000) = 0,97). In beiden Untersuchungs-zeiträumen werden nur wenige Queries häufiger als 100 mal benutzt. Im Jahr 2002 werden dreizehn und im Jahr 2000 acht Queries häufiger als 100 mal im Log gefunden (siehe dazu Tabelle 5-8). Die überwiegende Mehrheit der Queries wird bei der Extraktion lediglich ein oder zweimal gefunden.

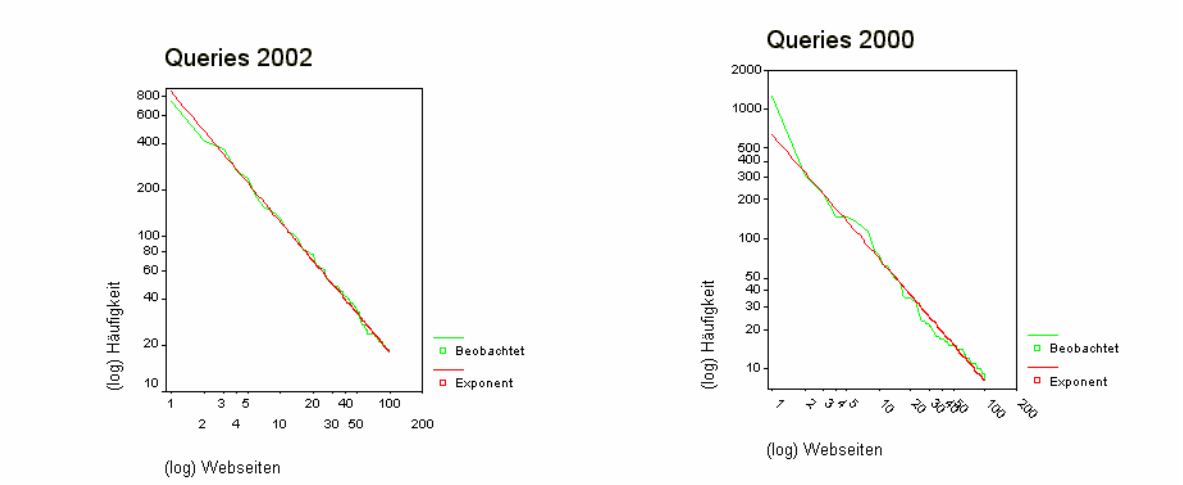


Abbildung 5-16: Verteilung der Queries (Top 100) nach Häufigkeiten (April 2000, 2002)

Die untere Tabelle 5-8 zeigt die zehn meist genutzten Queries für die beiden Untersuchungszeiträume über die Suchmaschinenbesucher auf die IB-Site navigiert sind. Bei der Betrachtung der unteren Tabelle fällt zum einen auf, dass einige Suchbegriffe (z.B. „fernstudium“, „florenz“ und „ascii code“), sowohl 2000 als auch 2002 sehr häufig auf die IB-Seite führen. Zum anderen fällt bei den Top Queries aus dem Jahr 2002 auf, dass sie häufiger als 2000 aus zwei Begriffen bestehen (siehe dazu auch nächsten Abschnitt). Sowohl 2000 als auch 2002 tauchen unter den Top Queries Begriffe auf (z.B. Fireball, Florenz, Amerika) bei denen nicht klar ist, warum sie so häufig auf die IB-Site führen, weil sie nicht unmittelbar zum thematischen Umfeld der Site passen. Hier scheinen Besonderheiten beim Ranking der Suchmaschinen verantwortlich zu sein.

Rang	Top Queries 2000	Häufigkeit	Top Queries 2002	Häufigkeit
1	fireball	1264	ascii code	745
2	fernstudium	305	fernstudium	412
3	florenz	226	römische zahlen	364
4	ascii code	146	ascii-code	263
5	amerika	146	florenz	239
6	bat	136	opac berlin	182
7	bernsteinzimmer	123	dos befehle	157
8	inktomi	116	hildebrandslied	147
9	beutekunst	83	www.humboldt-uni.de[60]	142
10	bibliothek	74	ms-dos befehle	131

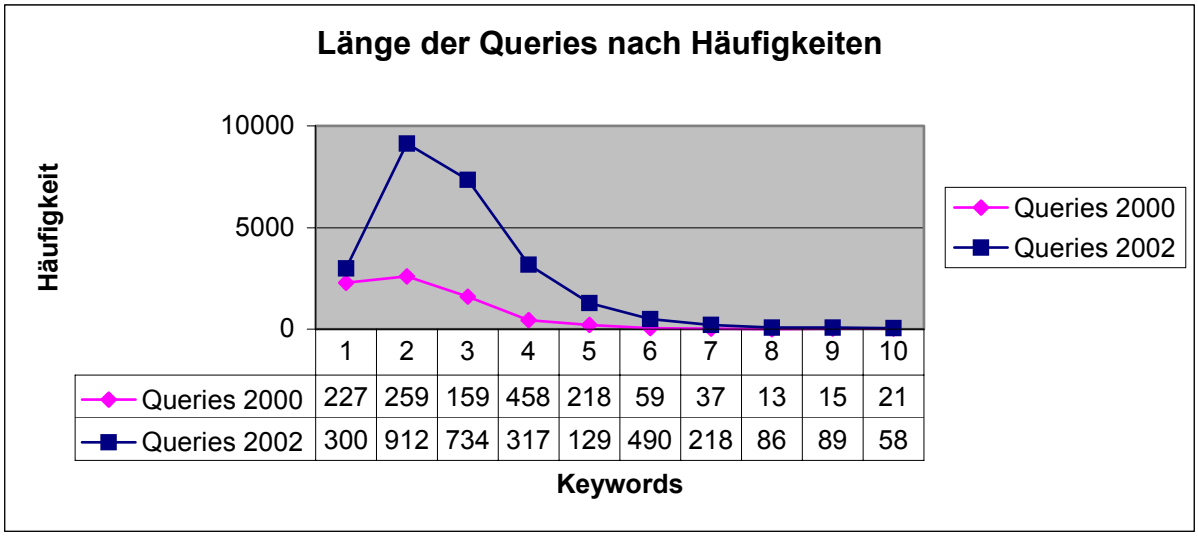
Tabelle 5-8: Top 10 Queries (April 2000, 2002)

5.11.2 Länge der Queries

Die folgende Abbildung zeigt die Verteilung der Länge der Queries, bezogen auf die Anzahl der Keywords (vgl. *Ross & Wolfram*, 2000 [80]). Hier zeigt sich eine Verschiebung des Suchverhaltens bzgl. der Anzahl der verwendeten Suchbegriffe (siehe dazu auch Anhang [10.2]). Die Suchmaschinenbesucher suchen 2002 mit mehr Suchbegriffen als noch im Jahr 2000. Während im Jahr 2000 die meisten Siteentries über Suchmaschinen mit einem oder zwei Suchbegriffen erfolgten, suchen die Suchmaschinenbenutzer 2002 vorwiegende mit zwei bzw. drei Suchbegriffen. Die Länge der Queries hat sich also für den Großteil der Queries durchschnittlich um einen Suchbegriff erweitert. Queries mit mehr als sechs Keywords sind in beiden Jahren eher selten (vgl. *Silverstein et al.*, 1999 [83], *Spink et al.*, 2001[86]).

Ein Blick auf die „erweiterten" Queries, also Queries die z.B. boolesche Operatoren enthalten (AND, OR, NOT) zeigt, dass der Einsatz des bekanntesten Suchoperators AND im April 2000 noch sehr viel häufiger (358 mal) bei der Suchmaschinenrecherche verwendet wurde als im April 2002 (202 mal). „Erweiterte" Queries sind in beiden Samples ausgesprochen selten.

[60] Anmerkung: der Domainname www.humboldt-uni.de ist nicht vergeben. Das Suchdienst MSN vermittelt diese Browsereingaben auf die Homepage des IB.



	1	2	3	4	5	6	7	8	9	10
Queries 2000	227	259	159	458	218	59	37	13	15	21
Queries 2002	300	912	734	317	129	490	218	86	89	58

Abbildung 5-17: Anzahl der Queries bezogen auf die Länge (April 2000, 2002)

5.11.3 Google's Trefferlisteninformation

Die Analyse der ungefähren Position der Seiten innerhalb der Suchmaschinentrefferliste ist über den URL-Bestandteil (z.B. &start=10) der Trefferlisten-URL möglich[61]. Die untere Abbildung 5-18 zeigt für beide Untersuchungszeiträume tendenziell ähnliche Verteilungen. Im Jahr 2002 liegen die Werte gemäß den höheren Siteentry-Zahlen über den Werten aus dem Jahr 2000. Die Verteilung zeigt, dass in der Regel bei Google nur wenige Trefferlistenseiten begutachtet und anschließend „geklickt" werden. Treffer, die über das Suchmaschinenranking einen „schlechteren" Listenplatz (Trefferlistenseite >10) erhalten, werden viel seltener über diese Navigationsart gefunden. Webseiten, die sich auf Listenplätzen auf Trefferlistenseiten > 100 befinden, haben kaum eine Chance über Suchmaschinen gefunden zu werden (vgl. *Silverstein et al.*, 1999 [83], *Spink et al.*, 2001[86]). Die Verteilung der Siteentries für die ersten 48 Trefferseiten (siehe Abb. 5-18) nähert sich ebenfalls für beide Jahre an ein klassisches Power Law an (Rsq (2002) = 0,991, Rsq (2000) = 0,966).

[61] Die meisten Suchmaschinenanbieter liefern innerhalb der Trefferlisten-URL die ungefähre Position der Trefferseite zurück. Bei einem Klick auf einen Suchmaschinentreffer der zweiten Trefferlistenseite liefert Google beispielsweise die Trefferlisten-URL mit dem Kürzel (&start=10, Beispiel Google) zurück.

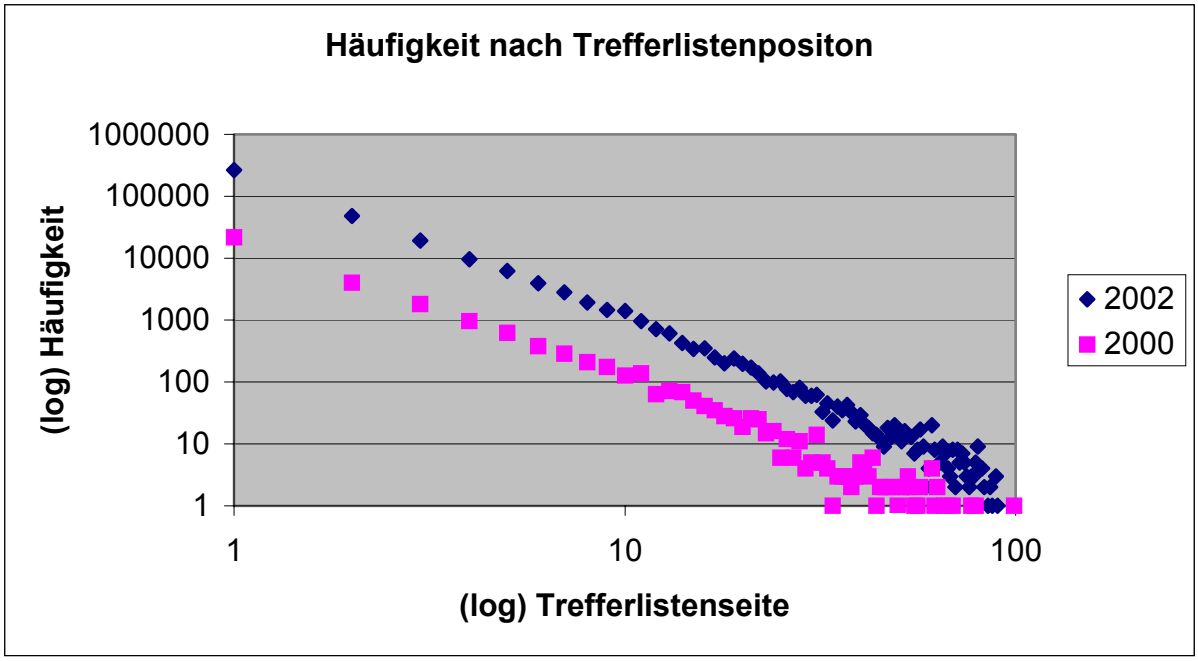


Abbildung 5-18: Häufigkeit nach Trefferlistenpositon (Google) (2002, 2000)[62]

5.12 Details der Backlinks

Insgesamt lassen sich im Jahr 2002 1323 unterschiedliche Websites aus dem Logfile extrahieren, die Links (Referenzen) enthalten, die auf Webseiten der IB-Site verweisen. Nur ein geringer Anteil der Websites mit Backlinks auf die IB-Site wird intensiv genutzt. Der Großteil der Backlinks bringt kaum Traffic auf der IB-Site (vgl. *Thelwall*, 2001 [105]). Die Verteilung der Häufigkeiten nähern sich für beide Untersuchungszeiträume einem klassischen Power Law an (Rsq (2002) = 0,994, Rsq (2000) = 0,992) (vgl. Abb. 5-19 und Anhang, Top Backlinks [10.3]).

Die Abbildung 5-20 zeigt die Verteilung der referenzierenden Sites bezogen auf die TLD's der Website-URL (z.B. www.domain.de). Diese Verteilung zeigt, dass insbesondere Websites aus dem Domainbereich Deutschland (n = 643) am häufigsten Links auf die IB-Site setzen, gefolgt von Backlinks auf kommerziellen Websites (n = 129). Die beiden weiteren deutschsprachigen Domainbereiche Österreich (siehe at, n = 60) und die Schweiz (siehe ch, n = 43) befinden sich auf vorderen Positionen (Platz 3 und 5). Die Websites aus dem Domainbereich edu (n = 49) und org (n = 39) befinden sich an vierter bzw. siebter Position. Die Sprache und die thematische Fokussierung (akademisches Webangebot) einer Website haben somit eine sichtbare Auswirkung auf die Zusammensetzung der Referenzen. Die Verteilung der Websites nach den 29 TLD's nähert sich ebenfalls an ein klassisches Power Law an (Rsq (2002) = 0,960).

[62] Hinweis: doppelt logarithmischer Maßstab (siehe log Häufigkeit und Trefferlistenseite)

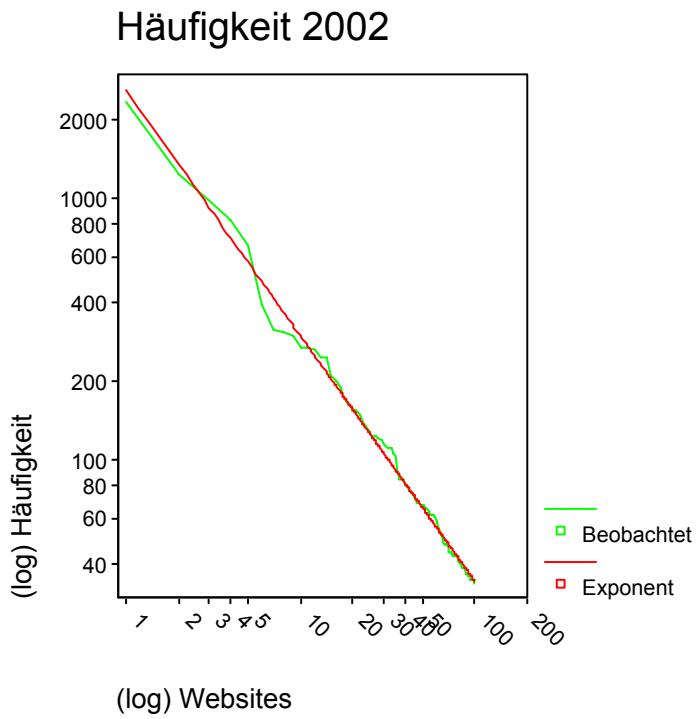


Abbildung 5-19: Häufigkeit der Websitebesuche der Top 100 Websites (Backlinks) (2002)[63]

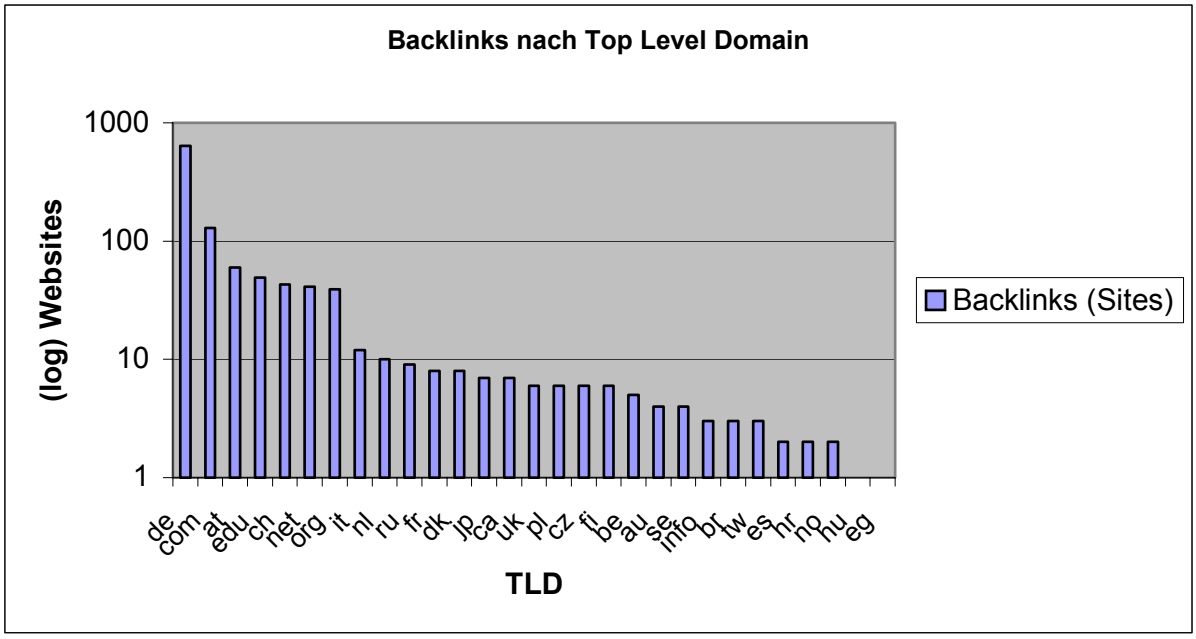


Abbildung: 5-20: Backlink-Websites nach Top Level Domain (2002)[64]

Die untere Tabelle 5-9 zeigt, dass neben der Sprache der referenzierenden Websites (vgl. Häufigkeit der Top Level Domain DE unter den Top 20) auch die räumliche bzw. institutionelle Vernetzung der IB-Site eine sichtbare Rolle spielt. Ein Großteil der Siteentries auf die IB-Site haben ihren Ursprung auf akademischen Universitätswebsites aus dem Großraum Berlin (siehe hu-berlin, fu-berlin, tu-berlin,

[63] Hinweis: doppelt logarithmische Maßstab (siehe log Häufigkeit, Websites)

fh-potsdam, dbi-berlin, udk-berlin). Die meisten Websitebesuche kommen über Links der Mutterorganisation Humboldt-Universität hu-berlin.de auf die Institutsseite (vgl. *Thelwall*, 2002 [98]). Websites aus dem Komplex „Bibliothek & Bibliothekswissenschaft“ bzw. allgemein Universität referenzieren ebenfalls häufig.

RANG	REF. WEBSITE	SITEENTRIES	BESCHREIBUNG (Klasse)
1	ub.hu-berlin.de	7217	Universität
2	www.hu-berlin.de	3654	Universität
3	physik.fu-berlin.de	2063	Universität
4	fh-potsdam.de	1708	Universität
5	de.dir.yahoo.com	1442	Directory
6	www2.hu-berlin.de	1140	Universität
7	zfuw.uni-koblenz.de	1114	Universität
8	dbi-berlin.de	1012	Akademisch
9	ub.tu-berlin.de	1002	Universität
10	udk-berlin.de	846	Universität
11	amor.rz.hu-berlin.de	744	Universität
12	www.uni-koblenz.de	693	Universität
13	hbz-nrw.de	654	Akademisch
14	sewanee.edu	639	Universität
15	niester.de	587	Privat
16	home.t-online.de	525	Privat
17	ub.fu-berlin.de	456	Universität
18	bsz-bw.de	436	Akademisch
19	phil.uni-erlangen.de	323	Universität
20	pitt.edu	262	Universität

Tabelle 5-9: Anzahl der Siteentries über Backlinks (Top 20 Websites) (2002)

Der folgende Abschnitt diskutiert die wichtigsten Ergebnisse der Loganalyse.

[64] Hinweis: logarithmischer Maßstab (siehe log Websites)

6 Diskussion

Die explorative Logfile-Untersuchung der drei methodisch voneinander abgrenzbaren Navigationsarten im Web (Suchmaschine, Direkt und Referenz) am Beispiel einer umfangreichen Website aus dem akademischen Bereich hat eine Reihe neuer Ergebnisse hervorgebracht, die im abschließenden Diskussionsteil im Mittelpunkt stehen sollen. Neben der Diskussion der wichtigsten Ergebnisse der Analyse sollen vor allem der Wert und die Potenziale erweiterter detaillierter Logfileanalysen (Mikro-Loganalysen) diskutiert werden.

6.1 Suchmaschinen als wichtigste Trafficlieferanten

Die vorliegenden Ergebnisse machen deutlich, dass die großen kommerziellen Such-maschinen eine dominierende Rolle bzgl. des Websitetraffics für die analysierte Website www.ib.hu-berlin.de spielen. 59% Prozent der Websitebesucher der IB-Site gelangen im Jahr 2002 über Suchmaschinen auf die Site, wobei der überwiegende Teil der Websitebesuche im Jahr 2002 von Google vermittelt wurde (vgl. auch *Sullivan*, 2003 [90]). Die Suchmaschinen liefern in der untersuchten Zeitreihe zunehmend mehr Besucher zu einem sehr breiten Spektrum von Themen und Begriffen (siehe Anzahl der unterschiedlichen Queries). Die beiden anderen untersuchten Navigationsarten „direkte Navigation“ und „Navigation durch Referenzen“ verändern sich an Anteilen kaum (siehe „Direkt“) bzw. verlieren im Untersuchungszeitraum an Bedeutung (siehe „Referenzen“). Diese und andere Untersuchungen im Bereich akademischer Websites (*Oldenburg*, 2002 [71], *Thelwall*, 2001 [105]) und anderer Websites (*Sullivan*, 2001 [89]) zeigen somit, dass Suchmaschinen heute die wichtigsten und beliebtesten Instrumente für Informationssuchende sind, um im Internet neue Webseiten zu finden. Insbesondere die akademischen Websites mit ihren z.T. umfangreichen Textsammlungen und internen und externen Verlinkungsstrukturen (*Thelwall*, 2002 [98]), die inzwischen von der Suchmaschinentechnologie beim Ranking der Treffer berücksichtigt werden, ziehen viele Suchmaschinen-Nutzer an. Die Gründe für die Dominanz der Suchmaschinen als Trafficlieferanten für spezifische Websites sind sicherlich komplex und vielfältig. Erste einfache Erklärungsansätze können aus den einzelnen Ergebnissen dieser Untersuchung aber skizziert werden.

1. Suchmaschinen indexieren und finden heute nach wie vor Text. Je mehr unterschiedlichen Text (Types und Tokens[65]) eine Webseite bzw. Site enthält, desto mehr Möglichkeiten bestehen theoretisch für einen Informationssuchenden, dass er die Webseite über eine Suchmaschinen-Query findet. Die Untersuchung kann zeigen, dass der Großteil der Einstiegsseiten der IB-Site (Top 100) einfache Textseiten (Artikel, Aufsätze, Skripte, …) sind, die überdurchschnittlich viel

[65] Es wird zwischen Type und Token in Dokumenten unterschieden. Längere Dokumente beinhalten viele Tokens, wobei jedes Wort als Token gezählt wird, auch wiederholte Worte. Die Zahl der Types (verschiedene Worte) wächst somit langsamer als die Zahl der Tokens, weil wiederholte Worte nicht als neuer Type gezählt werden (vgl. *Mayr*, 2002 [62]).

Text enthalten. Gemessen an der durchschnittlichen Dateigröße aller 100 untersuchten Webseiten (Ø der Top 100 ist 53 kbyte), enthält der Durchschnitt der 52 Textseiten 16 Kilobyte mehr Text (Ø der Textseiten ist 69 kbyte)[66]. Die Textseiten mit durchschnittlich mehr Text werden nachweislich häufiger über Suchmaschinen aufgerufen als die übrigen Seitentypen.

2. Der zweite Erklärungsansatz basiert auf den internen und externen Links einer Site bzw. Webseite, die von den Suchmaschinen analysiert und bewertet werden. Seitdem die referenzierenden Linkstrukturen einer Website mit in das Ranking der großen Suchmaschinen einfließen (vgl. link popularity), insbesondere ist hier der PageRank (PR) der Google Suchmaschine zu nennen [72], spielen vor allem die Backlinks einer Webseite eine große Rolle. Die gewachsenen Linkstrukturen einer Site stehen somit in direktem Zusammenhang zu dem zu erwartenden Traffic durch Suchmaschinen. Am Beispiel der PageRank-Werte des Rankingalgorithmus PageRank, den Google für jede indexierte Webseite ausgibt, konnte dies deutlich nachgewiesen werden. Die untersuchten Seiten mit einem höheren PageRank, sprich einer höheren Anzahl von Links und Backlinks, erhalten messbar mehr Traffic (Entries) als Seiten mit einem niedrigeren PR-Wert (vgl. Ergebnisse [5.9]). Allgemein lässt sich für die IB-Site aussagen, dass der mittlere PR aller untersuchten Seiten mit Ø 3,9 recht hoch ist. Unter den auf dem Webserver befindlichen untersuchten 100 Webseiten, existierte keine einzige Webseite, für die Google keinen PageRank (PR 0) berechnet hatte.

Weitere externe Faktoren, wie das Alter, die Sprache, der inhaltliche Kontext und besonders die Struktur und Zugänglichkeit (einschließlich der Kodierung der Webseiten) einer Site beeinflussen die Eigenschaften und Intensitäten seiner Nutzung. Inwieweit die angeführten externen Faktoren für die hohen Entry-Werte auf Seiten der Suchmaschinen mitverant-wortlich gemacht werden können, kann nur gemutmaßt werden, da sich die exakte Funktionsweise der Suchmaschinenalgorithmen (auch Google's PageRank) selbst dem wissenschaftlichen Betrachter als Blackbox darstellt. Gerade das Logfile kann an dieser Stelle unter Umständen sehr konkrete Hinweise geben. Hypothetisierend ließe sich in diesem Zusammenhang fragen, ob Websites, die umfangreiche, intern und extern verlinkte und für Suchmaschinenroboter indexierbare Textsammlungen anbieten, nach einem bestimmten Zeitraum mit hoher Wahrscheinlichkeit einen Großteil der Besucher über Suchmaschinen erhalten.

6.2 Einstiegsseiten im Fokus der Untersuchung

Ein Ziel dieser Untersuchung ist es den Blick ganz bewusst auf die Einstiegsseiten von Websitebesuchen zu lenken und diese über ihre Nutzungshäufigkeiten zu charakterisieren. Die Einstiegsseiten (Entry Pages oder Entry Points) einer Website haben grundlegende Bedeutung für websitebezogene Untersuchungen, da sie die ersten und unter Umständen letzten Webseiten sind, mit denen die Websitebesucher bei ihren Websitebesuchen in Berührung kommen. Diese Webseiten entscheiden mitunter über den Erfolg und die Dauer der einzelnen Websitebesuche. Der hohe Anteil

[66] Vgl. dazu die empirisch erhobenen durchschnittlichen Webseitengrößen folgender Untersuchungen (*Bergman*, [9], *Koehler*, 1999 [51], *Mayr*, 2002, [62]).

der Websitebesucher über Suchmaschinen hat in dieser Untersuchung sichtbare Auswirkungen auf die Zusammensetzung der wichtigsten Einstiegsseiten der Site. Die Ergebnisse zeigen, dass Suchmaschinen weniger zu „klassischen" Startseiten (z.B. Homepages) von Websites führen, sondern vielmehr Traffic in andere, für Suchmaschinen indexierbare Websitebereiche transferieren. Der Grund wurde oben schon angedeutet; die großen Suchmaschinen finden vor allem Text und der befindet sich nicht zwangsläufig auf den Homepages und anderen Startseiten (z.B. DB_Entry), sondern auf textbasierten Seitentypen. Die untersuchten Textkategorien (Text, Docu und Orga) machen im Jahr 2002 über 90% der Entries der wichtigsten Navigationsart Suchmaschine für die IB-Site aus und werden damit sehr häufig, vom Autor der Webseite unintendiert, zu Startseiten von Websitebesuchen. Als Folge dessen, erhalten Textseiten, die meist wenig auf intensive Nutzung vorbereitet sind, unter Umständen (Blackbox Suchmaschine) sehr viel Suchmaschinen-Traffic (Entries über Suchmaschinen, z.B. „Ascii-Seite", Top 2 [10.1]), der in vielen Fällen schon auf der ersten Seite endet. *Sullivan* (2001) bezeichnet dieses auch in anderen Untersuchungen festgestellte Phänomen als „search gap"[67] [89]. Diese Suchlücke „search gap" konnte *Oldenburg* (2003) für die Jahre 1999 und 2001 auf dem IB-Server ansatzweise bestätigen [71]. Suchmaschinenbesucher rufen demnach weniger Webseiten bei ihren Websitebesuchen auf als Websitebesucher, die über Backlinks auf die Site navigieren. Als mögliche Erklärung der „search gap" am IB könnte angeführt werden, dass ev. die Vielzahl der textbasierten, eher „funktionsarmen" Webseiten für das Navigationsverhalten der Webuser verantwortlich sind. Die Bedeutung der Einstiegsseiten wird außerdem sehr deutlich, wenn man sich zudem vor Augen führt, dass die 100 am häufigsten genutzten Einstiegsseiten der IB-Site, dass entspricht im Jahr 2002 etwa acht Promille aller zugegriffener Webseiten des IB-Webservers, ca. 50% aller Entries auf sich vereinbaren. Optimierungs- bzw. Anpassungsbemühung bzgl. des Webangebots einer Site, könnten sich dabei insbesondere auf die Einstiegsseiten konzentrieren und damit Suchmaschinenbesuchern die Navigation innerhalb der Website erleichtern. Die Ergebnisse zeigen weiterhin, dass sich bei „klassischen" Startseiten (siehe Seitenkategorien Home und DB_Entry) eine gänzlich andere Zugänglichkeit feststellen lässt als bei den textbasierten Seitenklassen. In beiden Untersuchungszeiträumen spielt die Navigationsart Suchmaschine für diese Klasse eine untergeordnete Rolle. Die „klassischen" Startseiten werden mehrheitlich direkt oder über Backlinks aufgerufen. Am deutlichsten zeigt sich dieses Verhalten an der wichtigsten Startseite (Top 1), der Homepage der Gesamtsite, die mit großem Abstand die meisten Entries aller Webseiten erhält (67.272 Entries). Die Homepage wird im Jahr 2002 zu 81% direkt aufgerufen. An dieser Seite, aber auch anderen klassischen Startseiten lässt sich ablesen, dass Linkautoren[68] und Websitenutzer, ihre Links, Bookmarks und Starteinstellungen sehr viel häufiger auf Homepages und andere Startseiten setzen, als auf spezifische Textseiten. Offensichtliche Gründe für dieses Verhalten, liegen neben der größeren Bekanntheit und häufigeren Aktualisierung von Homepages, sicherlich auch an den elaborierten Navigationsmöglichkeiten, die diese Startseiten meist bieten. Die Verteilung der

[67] Der Suchende besucht über eine Suchmaschinen-Query eine Seite, nutzt das Webangebot nicht weiter (gap), weil er sein Informationsbedürfnis bereits befriedigen konnte, oder sich auf der Seite nicht zurecht fand.

[68] Anmerkung: Mit Linkautoren sind Webnutzer gemeint, die die Möglichkeit besitzen, Hyperlinks im Web anzulegen. Über die Motivation der Linkautoren, einen Backlink auf das Angebot einer anderen Institution zu setzen, wurde an verschiedenen Stellen nachgedacht [48, 23, 95].

Navigationsarten auf die IB-Site für einzelne untersuchte Besuchertypen (siehe IBdHUB, HUB, EDU, COM) gibt Anlass zu folgender Aussage: je besser die Webbesucher der IB-Site die Angebote des Instituts für Bibliothekswissenschaft kennen, desto direkter navigieren sie auf die IB-Site bzw. nutzen die interinstitutionellen Verlinkungen. Das trifft insbesondere auf Besuchertypen zu, die institutionell mit dem IB verbunden sind (z.B. Angehörige der Humboldt-Universität). Neue Besucher navigieren hingegen überwiegend über Such-maschinen auf die IB-Site.
Im Zusammenhang eines regelmäßigen Webcontrollings einer Website sollte den Einstiegsseiten nach Ansicht des Autors künftig mehr Aufmerksamkeit geschenkt werden. Dazu gehört zumindest die wichtigsten Entry Pages seiner Website zu kennen und folglich zu wissen, warum die einzelnen Seiten so häufig Einstiegsseiten von Websitebesuchen werden. Die Ergebnisse der 100 am stärksten frequentierten Seiten einer Site zeigen, dass sich die Zugänglichkeit bzw. die Verteilung der drei Navigationsarten für die einzelnen Webseiten, Seitentypen und Besuchertypen z.T. sehr stark unterscheiden. Eine genauere Analyse und Bezifferung der Zugänglichkeit der Websiteentitäten (Site, Directory, Page), unterschieden nach Navigationsarten, wird notwendig, um konkrete Aussagen zu einzelnen Webseiten bzw. Webseitentypen (z.B. Ausreißer-Seiten, Webseitencluster) treffen zu können. An diesem Punkt setzt das entwickelte Konzept „Web Entry Faktoren“ (WEF) an.

6.3 WEF als neuer Nutzungs- und Navigationsindikator

Die Web Entry Faktoren (WEF) sind ein neues nutzungsbezogenes Loganalyse-Verfahren für freizugängliche Einstiegsseiten im Web (vgl. Web Usage Mining). Mit Hilfe der WEF’s werden eine Reihe informationswissenschaftlicher aber auch anderweitig interessanter Accessibility-Untersuchungen möglich. Im Zentrum des Verfahrens stehen die Nutzungsdaten freizugänglicher Webseiten sowie die Unterscheidung und Bezifferung definierter Einstiegszugriffe innerhalb des Logfiles. Die Anwendung der Logmetrik WEF auf unterschiedliche Websiteentitäten (Site, Directory, Page) zeichnet sich dadurch aus, dass die nötigen Daten auf allen gängigen Webservern ohne zusätzlichen Aufwand erhoben werden können und die Berechnung der WEF’s auf unterschiedlich granulierten Niveaus (Entities) erfolgen kann. Mit dieser neuen Form der erweiterten Logfileanalyse steht Informations-wissenschaftlern und Websiteverantwortlichen ein einfach zu implementierendes Konzept zur Verfügung, mit dem detaillierte Messungen der Nutzung und Sichtbarkeit des Webangebots auf einer Mikroebene vorgenommen werden können.
Die Ergebnisse der WEF-Berechnung der 100 wichtigsten Einstiegswebseiten des Instituts für Bibliothekswissenschaft demonstrieren, dass die Unterscheidung der Einstiegsnavigation einer Website (Suchmaschine, Direkt, Referenz) auf einem seitengenauen Niveau (Mikroebene) neue Formen der Bewertung und Analyse von Webinhalten im Kontext quantitativer Nutzungsmessung eröffnen. Beispielsweise lässt sich auf Grundlage der WEF-Werte für eine spezifische Webressource aussagen, ob und wie sie über Suchmaschinen gefunden wird, oder ob die Webseite sich anderweitig, etwa durch Backlinks oder Bookmarks als Startseite etabliert hat. Mit den WEF-Werten einer Ressource ist es somit möglich, ohne genaue Kenntnis der einzelnen Ressource, allgemeine Schlussfolgerungen bzgl. Sichtbarkeit und Akzeptanz ihres Inhalts und sogar der Bedeutung der Seite

für die Site anzustellen. WEF's identifizieren und klassifizieren auf einem sehr allgemeinen Level. Die WEF's einer größeren Anzahl von Webseiten sind beispielsweise in der Lage Cluster von Seiten zu identifizieren, die beinahe ausschließlich über „direkte Navigation" oder eine andere Navigationsart zugänglich werden. Die Identifikation einzelner Typen von Webseiten anhand ihrer Zugänglichkeit kann zu ganz neuen und an einigen Stellen skizzierten Anwendungen dieser Arbeit (z.B. WEF-Szenarien) führen. Praktikabel erscheint beispielsweise die Identifikation von Seiten, die besonders viele verschiedene genutzte Backlinks auf sich vereinigen, also Webseiten, die in der Internetforschung allgemein als „authorities" bezeichnet werden (vgl. *Kleinberg*, 1998, [49]). Ob die Seiten, die laut ihrer WEF-Werte (z.B. hohe Referenz-WEF's) auch wirklich „Autoritäten" (authoritative sources) innerhalb einer Gemeinschaft darstellen, sei dahingestellt, zumindest würde die WEF-Berechung auf Seiten hindeuten, die anhand ihrer Nutzungszahlen aus dem Rahmen fallen. Die WEF-Berechnung ist in der Lage auf die „Autoritäten" innerhalb eines Webangebots hinzuweisen. Die Kenntnis der Webseiten, die sehr viele Backlinks auf sich vereinigen, die nachweislich auch genutzt werden (vgl. *Thelwall*, 2001 [105]), kann für die Planung und Umsetzung neuer Webanbote sehr hilfreich sein. Hohe WEF's einer Ressource bei einer Navigationsart haben aufgrund der Zusammensetzung der Faktoren immer niedrige Werte bei den übrigen beiden Navigationsarten zur Folge. Niedrige WEF-Werte geben folglich Defizite bzgl. der Zugänglichkeit preis (vgl. z.B. Suchmaschinen-Sichtbarkeit der Webseitenkategorie DB_Entry). An dieser Stelle könnten Optimierungsbemühungen bestehender und einseitig zugänglicher Webinhalte ansetzen. Auf Basis der drei Zugriffswerte bzw. der abgeleiteten WEF's der genutzten Webangebote stehen Detailinformationen zur Verfügung, die neben den etablierten Standardauswertungen üblicher Loganalyzer neue Hinweise über Nutzungsmuster, Sichtbarkeit und Akzeptanz einer Website liefern. Die Ergebnisse der IB-Site bestätigen beispielsweise die allgemein festgestellte Tendenz, dass Websitebesucher sehr häufig über Suchmaschinen navigieren, zeigen aber auch wie sehr die Einstiegszahlen über Suchmaschinen variieren können.

Folgendes Szenario verdeutlicht die erweiterten Möglichkeiten der WEF-Metrik gegenüber Standardauswertungen: Eine neu erstellte Webseite, wird zu einem bestimmten Zeitpunkt auf eine von Suchmaschinen regelmäßig indexierten Website gestellt und mit dieser verlinkt. Externe Verlinkungen (Backlinks) zu dieser Seite bestehen zu diesem Zeitpunkt noch nicht. Zu einer sehr allgemeinen Suchmaschinen-Suchanfrage (z.B. „internet research") wird die Webseite nach einer Woche gelistet (z.B. vierte Seite bei Google). Zu diesem Zeitpunkt erhält die Webseite hauptsächlich Entries über die Navigationsart Suchmaschine, was sich an einem sehr hohen WEF-Suchmaschinen-Wert (z.B. 0,9) ausdrückt. Wenige Wochen später, die absoluten Zugriffszahlen der Webseite haben sich stark erhöht, fällt bei der Analyse der WEF-Werte auf, dass sich der Zugangstraffic überwiegend aus Backlinks und direkten Zugriffen zusammensetzt. Der Suchmaschinen-Traffic spielt ein ungeordnete Rolle (z.B. 0,1). Was ist passiert? Einige der Besucher, die die Website über eine Suchmaschinen-Query entdeckt haben, haben z.B. aufgrund der hohen Qualität bzw. Originalität der Ressource, Backlinks von deren starkbesuchten Websites auf die Webseite gesetzt. Diese Backlinks und Bookmarks werden aktuell sehr stark genutzt. Nach einem Jahr erhält die Webseite wieder überwiegend Traffic über die Suchmaschine Google (vgl. „Suchmaschinen-WEF"). Google bewertet die Backlinks der anderen Websites (vgl. Hubs, *Kleinberg*, 1999 [49]) und listet inzwischen die

Webseite auf Seite eins. Die Logfileauswertungen üblicher Standard-Loganalyzer hätte im Zusammenhang mit diesem Szenario lediglich ein Wachstum der Visit-Zahlen feststellen können. Die WEF-Metrik kann demgegenüber zu jedem Zeitpunkt anhand der drei WEF-Werte sehr genaue Aussagen bzgl. der Sichtbarkeit der einzelnen Webseite treffen.
Was macht die WEF's so interessant? Die WEF's liefern neue Website-Informationen (Nutzungs- und Sichtbarkeitsindikatoren), die sowohl für die wissenschaftliche Betrachtung (z.B. Website-Interlinking, Web Searching) als auch für die praktische Nutzungsauswertung und Evaluation von Webangeboten (siehe Szenario oben) künftig eine Rolle spielen können.
Die WEF-Werte geben aussagekräftige Hinweise über die Nutzung der externen Verlinkungsstrukturen und bilden damit eine nützliche Erweiterung der in der Webometrie begonnenen wissenschaftlichen Untersuchung von Hyperlinks. Während die Webometrie bislang nur die Existenz von Linkstrukturen nachweist und misst (durch Auszählen der Links zu einem festen Zeitpunkt), ist die hier entwickelte WEF-Metrik in der Lage, die Nutzung der Linkstrukturen einer Website aus dem Blickwinkel eines Knotenspunkts (Website) exakt wieder zugeben. Der besondere Wert der WEF-Metrik besteht darin, dass sie nutzungsbezogene Sichtbarkeitsindikatoren liefert, die unabhängig von den absoluten Nutzungszahlen sind. Folge der WEF-Berechnung sind skalierbare Kennzahlen der Sichtbarkeit und des Gebrauchs einer Website, die Hinweise auf die Bedeutung der definierten Navigationsarten für die Website aber auch Unterstrukturen, sowie inhaltliche Kategorien (siehe WEF-Werte der Seitenklassen) geben. Außerdem konzentriert sich das Konzept auf den spezifischen Webseitentyp „Einstiegsseite" (Entry Page), der für die Analyse von Webinhalten eine bislang unterschätzte, aber nach Ansicht des Autors entscheidende Bedeutung für die Akzeptanz und den Erfolg einer Website trägt.
Das folgende Zitat verdeutlicht die allgemeine Einsicht, Web Logfileanalysen künftig in erweiterten Untersuchungsanordnungen stattfinden zu lassen.

"Essentially, current measures are crude and interpretations are too simplistic for words. Visits, page impressions, time online, the aforementioned 'hits' and internet protocol (IP) addresses are variously used to measure use and users. The logs are also thin on content and quite raw and beg to be clothed with the meaning that only interviews, questionnaires and subscriber databases and cookies can furnish." Nicholas et al., 1999, S.265 [69]

6.4 Queries und Backlinks – die Schlüssel zur Nutzung?

Die Analyse der Queries und Backlinks, die zu einer Website bzw. deren Entitäten führen und aus dem Logfile extrahiert werden können (vgl. *Thelwall*, 2001 [105]), stellen ein spannendes und in dieser Arbeit lediglich per Stichprobe skizziertes Untersuchungsfeld dar. Das Interessante an diesen Daten sind zum einen die Intensitäten ihres Gebrauchs zum anderen die konkreten Ausprägungen der Suchanfragen (Queries) und externen Verlinkungen (Backlinks). Nach Auffassung des Autors liegt in den Query- und Backlink-Daten des Logfiles der Schlüssel nur Nutzung einer beliebigen Website. Die

nachfolgenden Ergebnisse und Ideen können fragmentarisch zur Entschlüsselung der Nutzungsdaten öffentlicher Websites dienen.
Eine kombinierte Analyse der konkreten Query- und Backlink-Informationen einer Website mit den zuvor berechneten WEF-Werten - als Konkretisierung und Untermauerung der drei aggregierten Entry-Werte - wäre die logische Fortführung für erweiterte Logfileanalysen, die den Untersuchungsfokus auf die Bestimmung der Sichtbarkeit und Zugänglichkeit von Websites legen.

6.4.1 Queries in den Logdaten

Die Stichproben der extrahierten Queries, die über die Suchmaschinen zur IB-Site geführt haben, verdeutlichen, dass die Analyse und Interpretation der anfallenden Daten ein ausgesprochen anspruchsvolles und aufwendiges Unterfangen ist. Zum einen fallen bei der Extraktion der Queries aus den Suchmaschinen-Entries enorme Datenmengen an (vgl. 24.888 unterschiedliche Queries allein im April 2002), die nicht mehr rein intellektuell gesichtet und ausgewertet werden können. Auf der anderen Seite existiert ein allgemeines methodo-logisches Problem im Zusammenhang mit der Interpretation der einzelnen Queries (vgl. *Thelwall*, 2001 [105]). Die Beschreibung bzw. Entschlüsselung des Informationsbedürfnis eines Informationssuchenden anhand der Suchmaschinen-Query, die zum Besuch einer Website geführt hat, stellt in vielen Fällen eine schwierige bis unmögliche Aufgabe dar (siehe Anhang, Top 100 Queries [10.2]). Noch viel schwieriger scheint in diesem Zusammenhang die Beurteilung zu sein, ob das ursprüngliche Informationsbedürfnis durch die aufgerufene Webseite befriedigt werden konnte oder ob der Websitebesucher keine hilfreichen Informationen auf seine Anfrage erhalten hat.

“A fundamental methodological problem here is that it is impossible to infer the precise need which lead to the query formulation. … the queries chosen by visitors were suggestive of at least a flexibility in information needs and perhaps also the ability to be easily distracted from the chosen goal, … “ Thelwall, 2001, S. 222 [105]

Die Extraktion der Queries anhand einer Stichprobe (Monat April 2002 und 2000) zeigt recht deutlich, dass die Verteilung der Häufigkeiten der Queries sich klassischen Power Law Verteilungen (Potenzgesetz) annähert. Die Ergebnisse der Stichprobe zeigen, dass wenige Queries im Logfile sehr häufig auftauchen, während sehr viele Queries nur ein oder zwei mal vorkommen. Die Power Law Verteilungen der Queries im Log deuten auf eine Ungleichverteilung der Sichtbarkeit der einzelnen Webseiten und Site-Themen bei den Suchmaschinen hin. Demnach existieren innerhalb der untersuchten Website Inhalte, die von Suchmaschinennutzern besonders häufig über die gleiche Suchanfrage aufgerufen werden (vgl. Anhang, z.B. „ascii code“ oder „fernstudium“). Auf der anderen Seite befinden sich auf der Website eine große Menge von Webseiten, die über sehr viele verschiedene Suchkombinationen (vgl. große Text-Seiten) auffindbar sind, aber über diese nur selten besucht werden. Ein Grund für die veränderlichen Verteilungen bei den protokollierten Besucherqueries und vor allem der Veränderlichkeit der Sichtbarkeit einzelner Webseiten zu bestimmten Themen (Keywords), sind die Ranking-Algorithmen der Suchmaschinen, die die

Treffergewichtung stetig aktualisieren und somit die Zusammensetzung der Trefferlisten beeinflussen (z.B. Google Dance). Bei der genaueren Betrachtung der Query-Information (siehe URL der Query) wird zudem deutlich, dass die Benutzer von Suchmaschinen sehr wenige Seiten der Ergebnistreffer durchsuchen. Diese Tatsache hat deutliche Auswirkungen auf die Verteilung und Häufigkeit der Queries und die Sichtbarkeit einzelner Inhalte bei Suchmaschinen. Ein weiteres Ergebnis (vgl. [5.11.3]) zeigt, dass die Wahrscheinlichkeit für Webseiten aufgerufen zu werden, rapide sinkt, je tiefer sie innerhalb der Ergebnisliste gelistet werden (siehe Power Law Verteilung). Eine Webseite, die auf Seite eins bei einer großen Suchmaschine gelistet wird (z.B. „Ascii-Seite" der IB-Site), erhält deutlich mehr Traffic, als eine Webseite die sich auf Seite zehn befindet. Dieses nachvollziehbare Verhalten der Suchmaschinenbenutzer, kann durch die Verteilung der Positionsangaben (Ergebnisseite 1, 2, 3, … 100 innerhalb der URL) für beide Untersuchungszeiträume bestätigt werden. Die Tendenz der Suchmaschinennutzer lediglich wenige Ergebnisseiten durchzusehen wird sich aller Voraussicht nicht ändern, da viele Informationssuchende nach Autopsie der „ersten" Treffer entweder zufrieden mit den Ergebnissen sind, oder ihre Suchanfrage verändern. Kein „Mensch" durchsucht 100 Ergebnisseiten.
Ein weiteres Ergebnis der Query-Analyse zeigt ein interessantes Online-Verhalten im Zusammenhang mit der Nutzung von Suchmaschinen generierten Ergebnislisten. Die Suchmaschinennutzer aus dem Jahr 2002 benutzen laut IB-Logfile deutlich mehr Suchbegriffe in ihren Queries als noch zwei Jahre zuvor, um auf die IB-Site zu gelangen (vgl. Ergebnis [5.11.2]). Dies kann als Anpassung an die stark gestiegenen Treffermengen gewertet werden, die Suchmaschinen heute zu einfachen „one term"-Queries ausgeben. Während *Lawrence & Giles* [58] im Jahr 1999 aussagen, dass das öffentliche indexierbare Web aller großen kommerziellen Suchmaschinen zusammen etwa 800 Millionen Webseiten umfasst, gibt Google heute alleine an, das 3,3 Milliarden Webseiten über seinen Index recherchiert werden können (www.google.com, November 2003). Die Suchmaschinenbenutzer reagieren auf die gestiegenen Informationsmengen mit längeren Suchanfragen.

6.4.2 Nutzung der Backlinks

Die genaue Betrachtung der Backlinks (Referenzen), die auf die Website ib.hu-berlin.de zeigen, entwirft ein weiteres, neues Bild dieses Instituts sowie seiner Angehörigen im Web und gibt wichtige Hinweise für die Erklärung des Sitetraffics. Backlink-Zählungen sind heute trotz ihrer Ungenauigkeit[69] [46, 85, 83, 29, 23, 7] die wichtigste Grundlage für die Webometrie. Außerdem spielen die Backlinks neben der Forschung auch bei Suchmaschinen eine nicht mehr wegzudenkende Rolle. Das Rankingkriterium Backlink (vgl. PageRank) hat sich gegenüber anderen Rankingverfahren durchgesetzt und ist von den Suchmaschinen-benutzern aufgrund der verbesserten Qualität der Suchergebnisse und der Transparenz und Nachvollziehbarkeit des Verfahrens breit akzeptiert. Die Analyse der „genutzten" Backlinks der untersuchten Website anhand des Logfiles, ev. in Kombination mit anderen Backlinkzählungen, liefern weitere Informationen, die nachfolgend andiskutiert werden sollen.

[69] "It has been pointed out that web data is inherently more unreliable and technically problematical than citations." *Thelwall*, 2001 [101]

Im Gegensatz zu Backlink-Analysen der Webometrie liefert das Logfile konkrete Informationen bzgl. der Intensität der Nutzung der bestehenden Backlinks einer Website. Bei der Analyse der Backlinks interessieren vor allem die Backlinks, die die meisten Besucher auf die Website bringen, aber auch die Backlinks, die keinen oder nur sehr wenig Traffic generieren. Neue Erkenntnisse im Zusammenhang mit Backlinkanalysen verspricht ein Abgleich bzw. Vergleich verschiedener Backlinkstatistiken. Zum einen die Backlinks die das Logfile aufzeichnet, also die real genutzten Backlinks einer Site, zum anderen die Backlinks, die die großen Suchmaschinen als Referenzen einer Site ausgeben (vgl. *Thelwall*, 2001 [105]).[70]

„A combination of search engine backlink discovery facilities and web log analysis is likely to yield a fuller picture of the link structure surrounding a site than either taken on their own." Thelwall, 2001, S. 222 [105]

Der Abgleich wäre beispielsweise in der Lage genauere Informationen darüber zu liefern, welche Backlinks sowohl im Logfile auftauchen als auch von der Suchmaschine ausgewiesen werden. Weiterhin zeigt der Abgleich aber auch wo die Suchmaschine Defizite bei der Abdeckung (weist Backlinks nicht nach) aufweist. Interessant wäre beispielsweise der Prozentsatz der protokollierten, aber von der Suchmaschine nicht nachgewiesenen Backlinks.[71]
Welche Backlinks sind nun aber die wichtigsten für eine Website? Das sie wichtig für den zu erwartenden Traffic einer Website sind, zeigen die einzelnen Ergebnisse dieser Untersuchung (PR und Entries-Korrelationen) sowie die Ausführungen zur allgemeinen Funktion und Wirkung der Google PR-Werte. Sind beispielsweise die Backlinks die wichtigsten, die „substanziellen Traffic"[72] auf die Website bringen, oder ist es vor allem wichtig, viele Backlinks hochgeschätzter Website (hoher PR) zu erhalten um selber hoch gerankt zu werden? Welche Backlinks heben den allgemeinen PR der Website am meisten bzw. welche konkreten Webseiten erhalten warum Backlinks wichtiger Hubs? Auf alle diese Fragen können nur kombinierte mehrstufige Analysen Antworten liefern, die sowohl webometrische, logfilebasierte und weitere Kennzahlen vereinbaren.
Hypothetisierend lässt sich formulieren: die IB-Site erhält insgesamt gesehen deshalb so viele Entries über Suchmaschinen, weil einzelne ihrer Backlinks sehr große Bedeutung (gemessen an ihrem PR-Wert) innerhalb eines Teiles des Webs (community) haben. Diese Backlink-Sites (Referenzen) geben über ihre Verlinkung Bedeutung an die IB-Site ab und beeinflussen durch das Suchmaschinenranking den Traffic der Site indirekt. Die hohen Suchmaschinen-anteile einer gesamten Website sind somit die Folge einzelner Verlinkungen bzw. vererbter Link-Bewertungen sein. Die konkrete Nutzung von Backlinks (Navigationsart Referenz) tritt in den Hintergrund, während allein die Existenz und der Nachweis der Backlinks immer wichtiger wird. Neben dem Traffic den Backlinks direkt oder indirekt

[70] Google und die anderen großen Suchmaschinen bieten die Möglichkeit, die Backlinks einer Webseite über eine spezielle Abfrage auszugeben (z.B. Goggle-Syntax: link:http://www.domain.de/).
[71] Thelwall identifiziert und beziffert in seiner Untersuchung über eine kombinierte Analyse der beiden Backlinkzählungen drei Bereiche: 1. Sowohl im Log als auch Suchmaschine, 2. nur im Log, 3. nur in der Suchmaschine. Der Großteil der Backlinks (48%) tauchen in seiner Untersuchung aus dem Jahr 2001 nur im Logfile auf (vgl. Thelwall, 2001, S. 221 [105]).
[72] Mit substanziellem Traffic sind häufige Websitebesuche gemeint, die das Angebot der Website intensiv (gemessen an der Anzahl der aufgerufenen Webseiten) nutzen.

generieren, ist es für die Zukunft denkbar, dass Backlinks auch als Evaluationsgrundlage immer wichtiger werden (vgl. *Thomas & Willet*, 2000 [113]). In diesem Zusammenhang lohnt ein Blick auf *Kleinbergs* Konzept der „Hubs“ und „Authorities“[73] [49]. *Kleinbergs* Konzept folgend, ließen sich mit Hilfe des Logfiles, der Suchmaschinenangaben und weiteren Analysen (z.B. Inhaltsanalyse) für jede Website abgeleitete Hub- bzw. Authority-Werte für die einzelnen Webseiten errechnen. Daraus ließen sich gänzlich neue Seitentypen konzipieren. Für alle folgenden Seitentypen ließen sich die Nutzungswerte (Logfile) als zusätzlichen Faktor integrieren:

1. Hubs innerhalb der Site: grundsätzlich jede Webseite, die Links (outlinks) zu Webseiten (authorities) anderer Websites enthält.
2. Hubs außerhalb der Site: über die Backlinkzählung (Hubs außerhalb der Site) ließen sich die internen Authorities (siehe unten) einer Site identifizieren. Über das Logfile ließe sich zusätzlich eine Gewichtung der einzelnen externen Hubs vornehmen (z.B. über den PR-Wert der Hubs oder die Anzahl der vermittelten Entries pro Hub).
3. Authorities innerhalb der Site: durch Zählung und Gewichtung der externen Hubs (Backlinks) außerhalb der Site
4. Authorities außerhalb der Site: durch Einbeziehen des PR-Wertes der identifizierten Ressourcen

Wie sehen die Backlinks der IB-Site aber nun aus? Ein Großteil der Backlinks (n = 1323), die auf Seiten des Instituts für Bibliothekswissenschaft verweisen, haben ihren Ursprung auf Websites aus dem TLD-Bereich „de“ (Deutschland, n = 643) bzw. aus dem deutschsprachigen Raum Österreich (at) (n = 60) und Schweiz (ch) (n = 43). Das verwundert nicht weiter, da die Hauptsprache der Webseite deutsch ist. Es verwundert auch nicht, dass die meisten internationalen Backlinks von Websites der TLD com (n = 129) und die weiteren Domains mit großem Abstand folgen edu (n = 49), gefolgt von net (n = 41) und org (n = 39). Auffällig ist jedenfalls die Nutzung der Backlinks nach Website, die nähert sich an eine Power Law Verteilung an (Rsq (2002) = 0,96) (vgl. *Rousseau*, 1997 [82], *Cui*, 1999 [25]). Die Betrachtung der meistgenutzten Backlinks der IB-Site (Top 20, siehe Ergebnisse [5.12]) zeigt, dass die wichtigsten Backlinks dieser universitären Website ebenfalls auf Webservern akademischer und universitärer Institutionen liegen (siehe Backlinkklassen Universität, Akademisch, vgl. dazu auch *Thelwall*, 2001, S. 220 [105]). Eine genauere Analyse (Klassifikation) der einzelnen genutzten Backlinks könnte hier detaillierte Aussagen bzgl. der intra- bzw. interuniversitären Verlinkung sowie die Nutzung der wissenschaftlichen Netzwerke (scholarly communication networks) liefern.
Die Autopsie der Webseiten auf denen sich die Backlinks befinden inkl. der Kontexte und Eigenschaften der Links, bietet die Möglichkeit für weitere anschließende Untersuchungen. Beispielsweise ließe sich in diesem Zusammenhang aussagen, welche Backlinks den klassischen Zeitschriftenzitationen am ehesten entsprechen würden.

[73] Das Konzept von *Kleinberg* sieht vor auf Grundlage einer definierten Anzahl von Webseiten (root set) anhand extensiver Linkzählungen (iterative algorithm) zwei Typen von Webseiten zu errechnen. Er führt dazu die beiden Begriffe Authority und Hub ein. “Hyperlinks encode a considerable amount of latent human judgment, and we claim that this type of judgment is precisely what is needed to formulate a notion of authority.” (*Kleinberg*, 1997 [49]).

Angemerkt sein, dass die Ergebnisse der Untersuchung zum einen nur eine Momentaufnahme der Zugänglichkeit der IB-Site der Jahre 2002 und 2000 darstellen. Zum anderen beschränken sich die Ergebnisse dieser Untersuchung auf die Logdaten eines singulären Universitäts-Webservers, folglich sind Aussagen zu allgemeinen Navigationstrends, wie sie an einigen Stellen in dieser Arbeit vorkommen, nur hypothetisch möglich. Voraussichtlich lassen sich ähnliche Zugangswerte bzw. -verteilungen aber auch für andere textbasierte Website aus dem akademischen Bereich feststellen.

7 Zusammenfassung

Die Zusammenfassung stellt die wichtigsten Ergebnisse und Diskussionsstränge dieser Arbeit in sehr verkürzter Form dar.

1. Die Ergebnisse der Untersuchung zeigen, dass die großen Suchmaschinen die wichtigsten Trafficlieferanten für die untersuchte (akademische) Website sind. Die Gründe hierfür liegen zum einen in der inhaltlichen Zusammensetzung (viele textbasierte Webseiten) und Struktur der Website sowie an den umfangreichen externen Verlinkungen (Backlinks) der Website. Hierfür konnten deutliche Hinweise in den untersuchten Daten gefunden werden. Der große Anteil der Navigationsart Suchmaschine hat deutlich sichtbare Auswirkungen auf die Ergebnisse der Untersuchung. Die Backlinks spielen indirekt für die Verteilung der Navigationsarten eine sehr wichtige Rolle.

2. Die 100 extrahierten und nach *Haas & Grams* [41] kategorisierten Einstiegswebseiten der Website setzen sich mehrheitlich aus textbasierten Seitenklassen aus allen Bereichen und Hierarchieebenen der Website zusammen, die sicherlich nicht zu den „klassischen“ Startseiten einer Website gehören. Neben den Textklassen (Text, Orga und Docu) finden sich weitere Seitentypen (Home, DB_Entry) unter den 100 am häufigsten genutzten Startseiten. Die Verteilung der drei alternativen Navigationsarten für die Einstiegsnavigation zu der IB-Site folgt vereinfacht folgender Faustregel. Suchmaschinen finden vor allem Textseiten, während die „klassischen“ Startseiten (z.B. Homepages) überwiegend durch direktes Aufrufen oder über Backlinks zugänglich werden. Ausreißer bzgl. der Zugänglichkeit finden sich unter den Einstiegseiten in allen untersuchten Seitenkategorien. Auffällig ist weiterhin, dass zum einen einzelne Cluster von Webseiten in ihrer Zugänglichkeit sehr ähnlich sind, zum anderen, dass die protokollierten Nutzungsvorgänge der 100 Einstiegsseiten sich einer Power Law Verteilung annähert.

3. Die Logfilemetrik Web Entry Faktoren (WEF) verfolgt das Ziel, über die Differenzierung der Nutzungszahlen neue quantitative Kennzahlen für die Loganalyse bereitzustellen, die zur Beschreibung der Sichtbarkeit, Zugänglichkeit und Nutzung unterschiedlicher Web-Entitäten (Site, Directory, Page) einer Website dienen. Anhand der WEF-Kennzahlen konnte beispielsweise detailliert gezeigt werden, in welchem Maße die untersuchten Seitenkategorien durch die drei Navigationsarten zugänglich werden. Es wurde weiterhin gezeigt, dass die WEF-Metrik aufgrund ihrer Granulierung als ein wirkungsvolles Instrument angesehen werden kann, dass zur Optimierung bzw. zur Evaluierung von freizugänglichen Webangeboten dienen kann. Zudem konnte demonstriert werden, dass mit Hilfe der WEF’s einer Site, Zugänglichkeits-Phänomene definierter Webseiten transparenter gemacht werden können.

4. Die stichprobenartige Analyse der Queries und Backlinks der untersuchten Website, die im Anschluss an die eigentliche Untersuchung vorgenommen wurde, konnte zeigen, dass Logfiles eine Vielzahl von Daten liefern, die bislang nicht ausreichend untersucht wurden (mit Ausnahme von *Thelwall*, 2001 [105]) und zum Teil methodologische Probleme aufwerfen. Die Untersuchung der Queries konnte eine Reihe von Ergebnissen anderer Untersuchungen bestätigen, außerdem konnte die Existenz von Power Law Verteilungen in den extrahierten Daten nachgewiesen werden. Die Analyse der Queries und Backlinks einer Website kann als die unterste Ebene (Mikroebene) der Logfileanalyse angesehen werden, hier liegt nach Ansicht des Autors der Schlüssel zum Verständnis der Nutzung einer Website. Logfiles sind mehr den je reiche Fundstellen zur Untersuchung von unterschiedlichsten Online-Phänomenen.

8 Ausblick

Die Untersuchung konnte zeigen, dass Web Logfileanalysen eine Reihe neuer informationswissenschaftlich interessanter Anwendungsgebiete bieten (vgl. *Thelwall, Vaughan, Björneborn*, 2004 [111]) und heute in keiner Weise als wissenschaftlich ausgereizt angesehen werden können. Im Gegenteil, quantitative Nutzungsmessung wird nach Ansicht des Autors in seiner Bedeutung und Anwendungsvielfalt (z.B. Websiteevaluation, -adaption) sowohl in der Wissenschaft als auch im kommerziellen Kontext zunehmen. Diese und andere neuere Untersuchungen deuten diese Tendenz bereits an (*Nicholas et al.*, 2003 [70]). Die Unterscheidung und separate Bezifferung der drei alternativen Navigationsmöglichkeiten (Suchmaschine, Direkt und Backlink) zu Einstiegsseiten einer Website über das Logfile ist nach heutigem Kenntnisstand in der hier angewendeten Form neu. Dieser Typ Logfileanalyse zeigt, trotz seiner eingeschränkten Aussagekraft für das gesamte Web, dass kombinierte mehrdimensionale Analysen die Aussagekraft der Ergebnisse deutlich steigern können (vgl. *Nicholas et al.*, 1999 [69], *Thelwall*, 2001 [105]) und weitere Potenziale für die Webometrie bzw. das Web Mining bieten. Das vorgestellte Logfileanalyse-Verfahren WEF, dass lediglich am Beispiel einer singulären Website erprobt wurde, bietet sich an, auch auf andere Websitestypen (z.B. kommerzielle Websites, akademische Universitätswebsites) bzw. Websitegrößen und -sprachen angewendet zu werden. Dabei wäre es interessant über einen regionalen Vergleich Unterschiede, aber auch Gesetzmäßigkeiten im Webuse zwischen den einzelnen Websitetypen zu identifizieren. Die parallele Analyse mehrerer zusammen-hängender Websites, sowie die Verfolgung und Messung der Nutzung zwischen den Websites, ist eine weitere denkbare aber organisatorisch schwer zu realisierende Untersuchungserweiterung. Eine noch detaillierte Analyse einzelner exponierter Webentitäten, wie der IB-Homepage, aber auch anderer Webseiten (z.B. Online-Artikel) oder Verzeichnisse, über einen längeren Zeitraum (z.B. Vierjahres-Zeitreihe), wäre mit Sicherheit ebenfalls lohnend. Hierbei könnte die Integration und genauere Analyse der Queries und Backlinks im Mittelpunkt stehen, die in dieser Untersuchung nur am Rande dargestellt werden konnte.

Finally, we are totally convinced, not of the quality of the cyber use data, but of the fact that this data (in enhanced form, no doubt) will hold the key to the customisation and individualisation of information. The blueprint is being laid down now. It would be a fool who thought otherwise. The tail will wag the dog - and how! Nicholas et al., 1999, S. 268 [69]

9 Literatur

1. Abdulla, Ghaleb: Analysis and Modeling of World Wide Web Traffic. 1998, available: http://citeseer.nj.nec.com/abdulla98analysis.html [19. November 2003].
2. Almind, T. C.; Ingwersen, Peter: Informetric analyses on the world wide web: methodological approaches to 'webometrics'. In: Journal of Documentation, Vol. 53, 1997, S. 404-426.
3. Barabasi, Albert-László: Linked – The New Science of Networks. Cambridge, Mass.: 2002.
4. Bar-Ilan, Judit: Data collection methods on the web for informetric purposes – A review and analysis. In: Scientometrics, Vol. 50, 2001, S. 7-32.
5. Bar-Ilan, Judit: Results of an extensive search for "S&T indicators" on the Web: a content analysis. Scientometrics, Vol. 49, 2000, S. 257-277.
6. Bar-Ilan, Judit: Search Engine Results over Time – A Case Study on Search Engine Stability. In: Cybermetrics, Vol. 2/3, 1998/99, available: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html [19. November 2003].
7. Bar-Ilan, Judit: The Web as an information source on Informetrics? A content analysis, Journal of the American Society for Information Science, Vol. 51, 2000, S. 432-443.
8. Batista, Paolo; Silva, Mário J.: Mining Web Access Logs of an On-line Newspaper. 2002, available: http://citeseer.nj.nec.com/535021.html [19. November 2003].
9. Bergman, Michael K.: The Deep Web: Surfacing Hidden Value. (Whitepaper), 2000, available: http://citeseer.nj.nec.com/llc00deep.html [19. November 2003].
10. Bjöneborn, Lennart; Ingwersen, Peter: Perspectices of webometrics. In: Scientometrics, Vol. 50, 2001, S. 65-82.
11. Björneborn, Lennart: Small-world link structures on the Web. Extended version of presentation at NetLab, Lund University Libraries, Sweden April 2002, available: http://www.db.dk/lb/2002smallworld.pps [19. November 2003].
12. Boudourides, Moses; Sigrist, Beatrice; Alevizos, Philippos: Webometrics and the Self-Organization of the European Information Sociaty. Draft Report, 1999, available: http://hyperion.math.upatras.gr/webometrics/ [19. November 2003].
13. Brin, S., Page, L.: The anatomy of a large scale hypertextual web search engine. In: Computer Networks and ISDN Systems, Vol. 30, 1998, S. 107-117, available: http://citeseer.nj.nec.com/brin98anatomy.html [19. November 2003].
14. Broder, Andrei: A taxonomy of web search. In: ACM SIGIR Forum, Vol. 36, 2002, S. 3-10, available: http://portal.acm.org/citation.cfm?id=792552&jmp=cit&dl=ACM&dl=ACM&CFID=11111111&CFTOKEN=2222222 [19. November 2003].
15. Catledge, Lara D.; Pitkow, James E.: Characterizing Browsing Strategies in the World-Wide Web. In: Computer Networks and ISDN Systems, Vol. 27, 1995, S. 1065-1073, available: http://citeseer.nj.nec.com/catledge95characterizing.html [19. November 2003].
16. Chakrabarti, Soumen et al.: Mining the Link Structure of the World Wide Web. 1999, available: http://citeseer.nj.nec.com/chakrabarti99mining.html [19. November 2003].

17. Chakrabarti, Soumen: Data mining for hypertext: A tutorial survey. In: SIGKDD Explorations, Volume 1, 2000, S. 1-11, available: http://citeseer.nj.nec.com/chakrabarti00data.html [19. November 2003].
18. Clever Project: Hypersearching the Web. In: Scientific American, 1999, S. 44-52, available: http://www.sciam.com/article.cfm?articleID=000BC474-9440-1CD6-B4A8809EC588EEDF&ref=sciam [19. November 2003].
19. Cockburn, Andy; McKenzie, Bruce: What do Web users do? An empirical analysis of Web use. In: International Journal of Human-Computer Studies, Vol. 54, 2001, S. 903-922.
20. Cooley, R. Mobasher; B., Srivastava, J.: Web Mining: Information and Pattern Discovery on the World Wide Web. In: Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, 1997, S. 558 –567.
21. Cooley, R.; Mobasher, B.; Srivastava, J.: Data Preparation for Mining World Wide Web Browsing Patterns. In: Knowledge and Information Systems, Vol. 1, 1999, S. 5-32.
22. Cothey, Vivian: A Longitudinal Study of World Wide Web Users' Information-Searching Behavior. In: Journal of the American Society for Information Science and Technology, Vol. 53, 2002, S. 67–78.
23. Cronin, Blaise et al.: Invoked on the Web. In: Journal of the American Society for Information Science and Technology, Vol. 49, 1998, S.1319 –1328.
24. Cronin, Blaise: Bibliometrics and Beyond: Some thoughts on webometrics and influmetrics. Freedom of Information Conference, 2000, available: http://www.biomedcentral.com/meetings/foi/cronin-tr.asp [19. November 2003].
25. Cui, Lei: Rating Health Web sites using the principles of Citation Analysis: A Bibliometric Approach. In: Journal of Medical Internet Research, Vol.1, 1999, available: http://www.jmir.org/1999/1/e4/index.htm [19. November 2003].
26. Dillon, Andrew; Gushrowski, Barbara, A.: Genres and the Web: Is the personal home page the first uniquely digital genre? In: Journal of the American Society for Information Science, Vol. 51, 2000, S. 202-205.
27. Dublin Core Metadata Initiative Homepage. Available: http://dublincore.org/ [19. November 2003].
28. Egghe, Leo: Discussions on Informetrics of the Internet and other social networks. 2002, available: http://lepont.univ-tln.fr/isdm/PDF/isdm6/isdm6a45_egghe.pdf [19. November 2003].
29. Egghe, Leo: New informetric aspects of the Internet: some reflections – many problems. In: Journal of Information Science, Vol. 26, 2000, S. 329-335.
30. Eschenfelder, Kristin; Wyman, Steven, et al.: Using Log Files to Assess Web-Enabled Information System Usage. 1997, available: http://hsb.baylor.edu/ramsower/ais.ac.97/papers/eschen.htm [19. November 2003].
31. Etzioni, Oren: The World Wide Web: quagmire or gold mine? In: Communications of the ACM, Vol. 39, 1996, S. 65-68.
32. Fenstermacher, Kurt D.; Ginsberg, Mark: Client-Side Monitoring for Web Mining. In: Journal of the American Society for Information Science and Technology, Vol. 54, 2003, S. 625-637.
33. Ford, Nigel; Miller, David; Moss, Nicola: Web search strategies and retrieval effectiveness: an empirical study. In: Journal of Documentation, Vol. 58, 2002, S. 30-48.

34. Fu, Yongjian; Shih, Ming-Yi; Creado, Mario: Reorganizing Web Sites Based on User Access Patterns. Available: http://web.umr.edu/~mingyi/paper/ijis02.pdf [19. November 2003]
35. Fühles-Ubach, Simone: Web-Statistik – Potenziale und Grenzen. In: B.I.T. online, H. 4, 2001, available: http://www.b-i-t-online.de/archiv/2001-04/fach1.htm [19. November 2003].
36. Garfield, Eugene: The most-cited papers of all time, SCI 1945-1988, part 1A: The SCI top100: Will the Lowry method ever be obliterated? In: Current Contents, Vol. 7, 1990, S. 3-14.
37. Gibson, David; Kleinberg, Jon; Raghavan, Prabhakar: Structural Analysis of the World Wide Web. 1998, available: http://www.w3.org/1998/11/05/WC-workshop/Papers/kleinber1.html [19. November 2003].
38. Google.com. Available: http://www.google.com/ [Novemver 2003].
39. Google Dance - The Index Update of the Google Search Engine. Available: http://dance.efactory.de [19. November 2003].
40. Gottlieb, Lisa; Dilevko, Juris: User Preferences in the Classification of Electronic Bookmarks: Implications for a Shared System. In: Journal of the American Society for Information Science and Technology, Vol. 52, 2001, S. 517–535.
41. Haas, Stephanie; Grams, Erika: Readers, Authors, and Page Structure: A Discussion of Four Questions Arising from a Content Analysis of Web Pages. In: Journal of the American Society for Information Science and Technology, Vol. 51, 2000, S. 181–192.
42. Hargittai, Eszter: Beyond Logs and Surveys: In-Depth Measures of People 's Web Use Skills. In: Journal of the American Society for Information Science and Technology, Vol. 53, 2002, S. 1239 – 1244.
43. Hernández-Borges, A.A. et al.: Can examination of WWW usage statistics and other indirect quality indicators distinguish the relative quality of medical Web sites?. In: Journal of Medical Internet Research, Vol. 1, 1999, available: http://www.jmir.org/1999/1/e1/index.htm [19. November 2003].
44. Huberman, Bernardo, et al.: Strong Regularities in World Wide Web Surfing. In: Science, Vol. 280, 1998, S. 95-97.
45. Iivonen, Mirja; White, Marilyn D.: The choice of initial Web Search Strategies: A Comparision between finnish and american searchers. In: Journal of Documentation, Vol. 57, 2001, S. 465–491.
46. Ingwersen, Peter: The calculation of Web Impact Factors. In: Journal of Documentation, Vol.54, 1998, S. 236-243.
47. Jansen, B.J.; Pooch, U.: A review of Web searching studies and a framework for future research. In: Journal of the American Society for Information Science and Technology, Vol. 52, 2000, S. 235-246.
48. Kim, Hak: Motivations for Hyperlinking in Scholarly Electronic Articles: A Qualitative Study. In: Journal of the American Society for Information Science and Technology, Vol. 51, 2000, S. 887– 899.
49. Kleinberg, Jon M.: Authoritative Sources in a Hyperlinked Environment. In: Journal of the ACM, Vol. 46, 1999, S. 604-32.
50. Kleinberg, Jon M.; Lawrence, Steve: The structure of the web. In: Science, Vol. 294, 2001, S. 1849-1850.

51. Koehler, Wallace: An Analysis of Web Page and Web Site Constancy and Permanence. In: Journal of the American Society for Information Science and Technology, Vol. 50, 1999, S. 162–180.
52. Koehler, Wallace: Web Page Change and Persistence – A Four-Year Longitudinal Study. In: Journal of the American Society for Information Science and Technology, Vol. 53, 2002, S. 162–171.
53. Kosala, Raymond; Bockeel, Hendrik: Web mining research: A survey. In: SIGKDD Explorations, Vol. 2, 2000, S. 1–15.
54. Koutsoupias, Nikos: Exploring Web Access Logs with Correspondence Analysis. In: Proceedings of the 2nd Hellenic Conference on AI, 2002, S. 229-236.
55. Larson, Ray R.: Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace. Paper presented at ASIS 96, 1996, available: http://sherlock.berkeley.edu/asis96/asis96.html [19. November 2003].
56. Lawrence, Steve: Context in Web Search. In: IEEE Data Engineering Bulletin, Vol. 23, 2000, S. 25–32.
57. Lawrence, Steve: Online or invisible? In: Nature, Vol. 411, 2001, S. 512.
58. Lawrence, Steve; Giles, Lee C.: Accessibility of information on the web. In: Nature, Vol. 400, 1999, S. 107-109.
59. Lawrence, Steve; Giles, Lee C.: Context and Page Analysis for Improved Web Search. In: IEEE Internet Computing, Vol. 2, 1998, S. 38-46.
60. Lawrence, Steve; Giles, Lee C.: Searching the World Wide Web. In: Science, Vol. 280, 1998, S. 98-100.
61. Lucas, Wendy; Topi, Heikki: Form and Function: The Impact of Query Term and Operator Usage on Web Search Results. In: Journal of the American Society for Information Science and Technology, Vol. 53, 2002, S. 95–108.
62. Mayr, Philipp: Das Dateiformat PDF im Web - eine statistische Erhebung. In: NFD – Nachrichten für Dokumentation, Jg. 53, 2002, S. 475-481, available: http://www.informatik.hu-berlin.de/~mayr/arbeit/pdf_im_web.pdf [19. November 2003].
63. McClure, Charles; Wyman, Steven; Beachboard, John: User and System-Based Quality Criteria for Evaluating Information Resources and Services Available from Federal Web Sites. 1997, available: http://slis-two.lis.fsu.edu/~cmcclure/exec_sum.html [19. November 2003].
64. Mell, Wolf-Dieter: Methodische Anmerkungen zur Auswertung der WWW-Log-Dateien des Servers www.gesis.org. IZ-Arbeitsbericht Nr. 26, 2002, available: http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab_26.pdf [19. November 2003].
65. Mettrop, Wouter; Nieuwenhuysen, Paul: Internet Search Engines – Fluctuations in Document Accessibility. In: Journal of Documentation, Vol. 57, 2001, S. 623–651.
66. Middleton, Iain; McConnell, Mike; Davidson, Grant: Presenting a model for the structure and content of a university World Wide Web site. In: Journal of Information Science, Vol. 25,1999, S. 219-227.
67. Nicholas, David et al.: Developing and testing methods to determine the use of web sites: case study newspapers. In: Aslib Proceedings, Vol. 51, 1999, S. 144-154.

68. Nicholas, David et al.: Digital journals, Big Deals and online searching behaviour: a pilot study. In: Aslib Proceedings, Vol. 55, 2003, S. 84-109.
69. Nicholas, David, et al.: Cracking the code: web log analysis. In: Online & CD-ROM Review, Vol. 23, 1999, S. 263-269.
70. Nicholas, David; Huntington, P.; Williams, P.: Micro-mining log files: a method for enriching the data yield from Internet log files. In Journal of Information Science, to appear 2003.
71. Oldenburg, Heike: Analysen von Webserver-Logfiles zur Kategorisierung des Navigationsverhaltens von Nutzern. Magisterarbeit am Institut für Bibliothekswissenschaft, 2003.
72. Page, L. et al.: The PageRank Citation Ranking: Bringing Order to the Web. 1998, available: http://citeseer.nj.nec.com/page98pagerank.html [19. November 2003].
73. Park, Ham W.; Barnett, George A.; Nam, In-Yong: Hyperlink-Affiliation Network Structure of Top Web Sites: Examining Affiliates with Hyperlink in Korea. In: Journal of the American Society for Information Science and Technology, Vol. 53, 2002, S. 592–601.
74. Pennock, David; Flake, Gary; Lawrence, Steve, et al.: Winners don't take all: Characterizing the competition for links on the web. In: Proceedings of the National Academy of Sciences, Vol. 99, 2002, S. 5207-5211.
75. Perkowitz, Mike; Etzioni, Oren: Towards adaptive Web sites: Conceptual framework and case study. In: Artificial Intelligence, Vol. 118, 2000, S. 245–275.
76. Pirolli, Peter; Pitkow, James; Rao, Ramana: Silk from a Sow's Ear: Extracting Usable Structures from the Web. In: Proceedings of 1996 Conference on Human Factors in Computing Systems, 1996, S. 118-125.
77. Pitkow, James: Summary of WWW Characterizations. In: Computer Networks and ISDN Systems, Vol. 30, 1998, S. 551–558.
78. Prime, Camille; Bassecoulard, Elise; Zitt, Michel: Co-citations and co-sitations: A cautionary view on an analogy. In: Scientometrics, Vol. 54, 2002, S. 291–308.
79. Rieh, Soo Young: Judgment of Information Quality and Cognitive Authority in the Web. In: Journal of the American Society for Information Science and Technology, Vol. 53, 2002, S. 145–161.
80. Ross, Nancy C. A.; Wolfram, Dietmar: End User Searching on the Internet: An Analysis of Term Pair Topics Submitted to the Excite Search Engine. In: Journal of the American Society for Information Science and Technology, Vol. 51, 2000, S. 949-958.
81. Rousseau, Roland: Daily time series of common single word searches in AltaVista and NorthernLight. In: Cybermetrics, Vol. 2/3., 1999, available: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html [19. November 2003].
82. Rousseau, Roland: Sitations: an exploratory study. In: Cybermetrics, Vol. 1, 1997, available: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html [19. November 2003].
83. Silverstein, C.; Henzinger, M.; Marais, H.; Moricz, M.: Analysis of a very large Web search engine query log. In: SIGIR Forum, Vol. 33, 1999, S. 6–12.
84. Smith, A.; Thelwall, Mike: Web Impact Factors for Australasian universities. In: Scientometrics, Vol. 54, 2002, S. 363-380.

85. Snyder, Herbert; Rosenbaum, Howard: Can Search Engines be used as tools for Web-Link Analysis? A critical review. In: Journal of Documentation, Vol. 55, 1999, S. 375–384.
86. Spink, Amanda: Introduction to the Special Issue on Web Research. In: Journal of the American Society for Information Science and Technology, Vol. 53, 2002, S. 65–66.
87. Spink, Amanda; Wolfram, Dietmar; Jansen, Major: Searching the Web: The Public and Their Queries. In: Journal of the American Society for Information Science and Technology, Vol. 52, 2001, S. 226-234.
88. Srivastava, Jaideep; Cooley, Robert, et al.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. In: SIGKDD Explorations, Vol. 1, 2000, available: http://citeseer.nj.nec.com/srivastava00web.html [19. November 2003].
89. Sullivan, Danny: Avoiding the Search Gap. SearchEngineWatch.com, 2001, available: http://www.searchenginewatch.com/sereport/article.php/2163711 [19. November 2003].
90. Sullivan, Danny: comScore Media Metrix Search Engine Ratings. SearchEngineWatch.com, 2003, available: http://searchenginewatch.com/reports/article.php/2156431 [19. November 2003].
91. Sullivan, Danny: Direct Navigation To Sites Rules, But Search Engines Remain Important. SearchEngineWatch.com, 2002, available: http://searchenginewatch.com/sereport/article.php/2164571 [19. November 2003].
92. Tauscher, L.; Greenberg, S.: Revisitation patterns in World Wide Web navigation. In: Proceedings of the Conference on Human Factors in Computing Systems, 1997.
93. Thelwall, et al.: European Union associated university websites. In: Scientometrics, Vol. 53, 2002, S. 95–111.
94. Thelwall, Mike, Wilkinson, David: Graph Structure in Three National Academic Webs: Power Laws with Anomalies. In: Journal of the American Society for Information Science and Technology, Vol. 54, 2003, S. 706-712.
95. Thelwall, Mike: A comparison of sources of Links for academic Web Impact Factor Calculations. In: Journal of Documentation, Vol. 58, 2002, S. 60-72.
96. Thelwall, Mike: A research and institutional size-based model for national university Web site interlinking. In: Journal of Documentation, Vol. 58, 2002, S. 683-694.
97. Thelwall, Mike: A Web Crawler Design for Data Mining. In: Journal of Information Science, Vol. 27, 2001, S. 319-325.
98. Thelwall, Mike: An initial exploration of the link relationship between UK university Web sites. In: Aslib Proceedings, Vol. 54, 2002, S.118-126.
99. Thelwall, Mike: Can Google's PageRank be used to find the most important academic Web pages?. In: Journal of Documentation, Vol. 59, 2003, S. 205-217.
100. Thelwall, Mike: Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university web sites. In: Journal of the American Society for Information Science and Technology, Vol. 53, 2002, S. 995-1005.
101. Thelwall, Mike: Extracting Macroscopic Information from Web Links. In: Journal of the American Society for Information Science and Technology, Vol. 52, 2001, S. 1157-1168.

102. Thelwall, Mike: Methods for reporting on the targets of links from national systems of university Web sites. In: Information Processing and Management, to appear 2003.
103. Thelwall, Mike: Research Note: in praise of Google: finding law journal Web Sites. In: Online Information Review, Vol. 26, 2002, S. 271-272.
104. Thelwall, Mike: The top 100 linked pages on UK university Web sites: high inlink counts are not usually directly associated with quality scholarly content. In: Journal of Information Science, 2002, Vol. 28, S. 485-493.
105. Thelwall, Mike: Web log file analysis: Backlinks and Queries. In: Aslib Proceedings, Vol. 53, 2001, S. 217-223.
106. Thelwall, Mike: Web use and peer interconnectivity metrics for academic Web sites. In: Journal of Information Science, 29, 2003, S. 11-20.
107. Thelwall, Mike: What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. In: Information Research, Vol. 8, 2003, available: http://informationr.net/ir/8-3/paper151.html [19. November 2003].
108. Thelwall, Mike; Harris, Gareth: The Connection between the Research of a University and Counts of Links to its Web Pages: An Investigation Based Upon a Classification of the Relationships of Pages to the Research of the Host University. In: Journal of the American Society for Information Science and Technology, Vol. 54, 2003, S. 594-602.
109. Thelwall, Mike; Smith, Alastair: Interlinking between Asia-Pacific University Web sites. In: Scientometrics, Vol. 55, 2002, S.363-376.
110. Thelwall, Mike; Tang, Rong; Price, Liz: Linguistic patterns of academic Web use in Western Europe. In: Scientometrics, Vol. 56, 2003, S. 417-432.
111. Thelwall, Mike; Vaughan, Liwen ; Björneborn, Lennart: Webometrics. In: ARIST, Vol. 39, 2004, preprint.
112. Thelwall, Mike; Wilkinson, David: Three Target Document Range Metrics for University Web Sites. In: Journal of the American Society for Information Science and Technology, Vol. 54, 2003, S. 489-496.
113. Thomas, O.; Willet, P.: Webometric analysis of departments of librarianship and information science. In: Journal of Information Science, Vol. 26, 2000, S. 421-428.
114. Vaughan, Liwen; Hysen, Kathy: Relationship between links to journal Web sites and impact factors. In: Aslib Proceedings, Vol. 54, 2002, S. 356-361.
115. Vaughan, Liwen; Thelwall, Mike: Scholarly Use of the Web: What are the Key Inducers of Links to Journal Web Sites? In: Journal of the American Society for Information Science and Technology, Vol. 54, 2003, S. 29-38.
116. Wang, Peiling; Berry, Michael W.; Yang, Yiheng: Mining Longitudinal Web Queries: Trends and Patterns. In: Journal of the American Society for Information Science and Technology, Vol. 54, 2003, S. 743 –758.
117. Web Characterization Terminology & Definitions Sheet. W3C Working Draft. 1999, available: http://www.w3.org/1999/05/WCA-terms/ [19. November 2003].

118. Weibel, Stuart: The Dublin Core Metadata Initiative. In: Zeitschrift für Bibliothekswesen und Bibliographie, Jg. 47, 2000, S.3 ff.

119. Wilkinson, David; Harries, Gareth, et al.: Motivations for academic web site interlinking: evidence for the Web as a novel source of information on informal scholarly communication. In: Journal of Information Science, Vol. 29, 2003, S. 49-56.

120. Wormell, Irene: Informetrics for informed decision making. Seminar paper, 2001, available: http://www.hb.se/bhs/seminar/semdoc/wormell.htm [19. November 2003].

121. Zawitz, Marianne W.: Web statistics - Measuring user activity – An analysis of the Bureau of Justice Statistics (BJS) website usage statistics. Working paper No. 19, Conference of European Statisticians, 1998, available: http://www.ojp.usdoj.gov/bjs/abstract/wsmua.htm [19. November 2003].

10 Anhang

10.1 Die “Top 100-Liste”

Rang	URL	Beschreibung	Kategorie	Size	PR	Entry S	Entry D	Entry R	WEF S	WEF D	WEF R	Entries total
1	/	Homepage des Instituts für Bibliothekswissenschaft	Home	av	6	6345	54369	6558	0,09	0,81	0,10	67272
2	/~mh/gedv/ascii.htm	Referenz der ASCII-Code Kodierung	Docu	av	4	19248	2399	187	0,88	0,11	0,01	21834
3	/~mh/projekte/metaopac/	Startseite des „Meta-Opac Berlin-Brandenburg“	DB Entry	av	5	2952	2490	8677	0,21	0,18	0,61	14119
4	/~is/computerkurs/ms-dos.html	Computertutorial zum Thema “MS-DOS”	Docu	av	3	10710	1745	43	0,86	0,14	0,00	12498
5	/~rfunk/lv/scripts/bwl/bwl.html	Vorlesungsscript zum Thema „Betriebswirtschaftslehre“	Text	lg	4	7530	1656	719	0,76	0,17	0,07	9905
6	/~mh/gedv/romzs.htm	Text zum Thema „Römisches Zahlensystem“	Text	av	4	7393	1051	858	0,79	0,11	0,09	9302
7	/~fern/	Ehemalige Homepage des Bereichs Fernstudium	Home	av	5	4371	2947	939	0,53	0,36	0,11	8257
8	/~wumsta/infopub/	Neue Homepage von Prof. Umstätter	Home	av	5	295	6566	19	0,04	0,95	0,00	6880
9	/~wumsta/rehm1.html	Textkapitel von Margarete Rehm	Text	lg	4	5025	974	427	0,78	0,15	0,07	6426
10	/~kumlau/handreichungen/h64/	Heft der „Berliner Handreichungen zur Bibliothekswiss.“	Text	lg	4	4647	978	344	0,78	0,16	0,06	5969
11	/~pz/florenz/florenbn.htm	Linkliste zum Thema „Florenz“	Orga	av	4	4879	479	150	0,89	0,09	0,03	5508
12	/~fern/fernstudium/postgradual/postgrad.html	Startseite des postgradualen Fernstudiums	Home	av	5	2689	2368	199	0,51	0,45	0,04	5256
13	/~wumsta/rehm4.html	Textkapitel von Margarete Rehm	Text	lg	4	4373	705	118	0,84	0,14	0,02	5196
14	/~hab/arnd/Start.html	Homepage des “Hildebrandslieds”	Home	av	4	1606	1996	1471	0,32	0,39	0,29	5073
15	/~is/rel-db2.htm	Studientext zum Thema “Relationale Datenbanken”	Text	av	4	3996	933	107	0,79	0,19	0,02	5036
16	/~kumlau/handreichungen/h54/kapitel1bis3.html	Kapitel der „Berliner Handreichungen zur Bibliothekswiss.“	Text	lg	4	4205	717	26	0,85	0,14	0,01	4948
17	/~pz/zahnpage/librdisc.htm	Vorlesungstext zum Thema „Universität des Mittelalters“	Text	lg	4	3180	946	633	0,67	0,20	0,13	4759
18	/~wumsta/rehm9.html	Textkapitel von Margarete Rehm	Text	lg	4	3786	652	117	0,83	0,14	0,03	4555
19	/~wumsta/rehm10.html	Textkapitel von Margarete Rehm	Text	lg	4	3836	646	68	0,84	0,14	0,01	4550
20	/~mh/projekte/oeb/pb/druckerei.html	Adressliste „Druckereien in Berlin“	Orga	av	3	4004	495	23	0,89	0,11	0,01	4522
21	/~mh/projekte/oeb/pb/	Adressliste	Orga	av	3	3875	500	85	0,87	0,11	0,02	4460

	second.html	„Second Hand-Läden in Berlin“										
22	/~mh/projekte/oeb/pb/theater.html	Adressliste „Theater in Berlin“	Orga	av	3	3930	455	28	0,89	0,10	0,01	4413
23	/inf/studium.htm	Informationen zum Studium am IB	Orga	av	5	2316	1052	919	0,54	0,25	0,21	4287
24	/~wumsta/rehm8.html	Textkapitel von Margarete Rehm	Text	lg	4	3537	557	112	0,84	0,13	0,03	4065
25	/~kumlau/handreichungen/h54/kapitel5bis8.html	Kapitel der „Berliner Handreichungen zur Bibliothekswiss.“	Text	lg	4	3070	950	45	0,76	0,23	0,01	4065
26	/inf/i_suche.htm	Linkliste zu Suchmaschinen	Orga	av	5	414	3077	485	0,10	0,77	0,12	3976
27	/jaw/Html/studwohn.html	Adressliste von Berliner Studentenwohnheimen	Orga	av	4	1373	308	2275	0,35	0,08	0,58	3956
28	/~wumsta/rehm11.html	Textkapitel von Margarete Rehm	Text	lg	4	3252	555	39	0,85	0,14	0,01	3846
29	/inf/bbbform.html	Start zur Recherche des „Berlin-Brandenburgischen Bibliotheksverzeichnisses“	DB Entry	sm	5	108	397	3095	0,03	0,11	0,86	3600
30	/~wumsta/rehm2.html	Textkapitel von Margarete Rehm	Text	lg	4	3040	402	21	0,88	0,12	0,01	3463
31	/~kumlau/handreichungen/h56/	Heft der „Berliner Handreichungen zur Bibliothekswiss.“	Text	lg	5	2514	768	73	0,75	0,23	0,02	3355
32	/~kumlau/handreichungen/h69/	Heft der „Berliner Handreichungen zur Bibliothekswiss.“	Text	lg	4	1898	1044	215	0,60	0,33	0,07	3157
33	/~wumsta/rehm3.html	Textkapitel von Margarete Rehm	Text	lg	4	2667	438	28	0,85	0,14	0,01	3133
34	/~wumsta/rehm71.html	Textkapitel von Margarete Rehm	Text	lg	4	2724	349	31	0,88	0,11	0,01	3104
35	/~kumlau/handreichungen/h58/	Heft der „Berliner Handreichungen zur Bibliothekswiss.“	Text	lg	4	1694	916	378	0,57	0,31	0,13	2988
36	/~wumsta/rehm6.html	Textkapitel von Margarete Rehm	Text	lg	4	2565	387	27	0,86	0,13	0,01	2979
37	/~wumsta/rehm7.html	Textkapitel von Margarete Rehm	Text	lg	4	2501	380	71	0,85	0,13	0,02	2952
38	/~fern/fernstudium/	Homepage des Bereichs Fernstudium	Home	av	5	400	2419	112	0,14	0,83	0,04	2931
39	/~mh/projekte/oeb/pb/buchhand.html	Adressliste von Berliner Buchhandlungen	Orga	av	3	2484	245	18	0,90	0,09	0,01	2747
40	/~pbruhn/russgus.htm	Startseite der Datenbank „RussGUS“	DB Entry	sm	6	862	676	1144	0,32	0,25	0,43	2682
41	/~sbuett/pm/strat_pm2.html	k.A.	k.A.	k.A.	k.A.	2149	422	25	0,83	0,16	0,01	2596
42	/~rfunk/lv/scripts/recht.html	Vorlesungsscript zum Thema „Rechtswissenschaft“	Text	av	3	2329	217	10	0,91	0,08	0,00	2556
43	/~pz/florenz/euroflor.htm	Text „Bedeutende Bibliotheken Europas“	Text	lg	4	1760	613	139	0,70	0,24	0,06	2512
44	/~pz/florenz/michendt.htm	Text zum Thema „Michelangelo“	Text	av	3	2118	255	73	0,87	0,10	0,03	2446
45	/~rfunk/lv/scripts/iud.ht	Script	Text	lg	4	2260	157	23	0,93	0,06	0,01	2440

	ml	„Information und Dokumentation“										
46	/~kumlau/handreichungen/h54/kapitel4bis43.html	Kapitel der „Berliner Handreichungen zur Bibliothekswiss.“	Text	lg	4	1808	432	173	0,75	0,18	0,07	2413
47	/~wumsta/rehm5.html	Textkapitel von Margarete Rehm	Text	lg	4	2065	270	12	0,88	0,12	0,01	2347
48	/inf/biblio.htm	Linkliste „Bibliotheken und Bibliothekskataloge“	Orga	av	5	1082	950	248	0,47	0,42	0,11	2280
49	/~kumlau/handreichungen/h38/kapitel5.htm	Kapitel der „Berliner Handreichungen zur Bibliothekswiss.“	Text	lg	3	2053	202	5	0,91	0,09	0,00	2260
50	/~kumlau/handreichungen/h66/	Heft der „Berliner Handreichungen zur Bibliothekswiss.“	Text	lg	5	1468	543	238	0,65	0,24	0,11	2249
51	/~hab/arnd/text.html	„Das Hildebrandlied“	Text	av	4	1226	348	670	0,55	0,16	0,30	2244
52	/~mh/css/css2/fonts.html	Engl. Tutorial zum Thema “CSS und Fonts”	Docu	lg	3	1999	235	2	0,89	0,11	0,00	2236
53	/~wumsta/	Alte Homepage von Prof. Umstätter	Home	av	5	405	1696	104	0,18	0,77	0,05	2205
54	/werner/lehrgang/html2/javascript/javascript.htm	Javascript-Tutorial	Docu	av	3	1842	324	5	0,85	0,15	0,00	2171
55	/~rfunk/fernstudium/funkmana.html	Glossar zum Lehrgebiet „Management“	Text	av	4	1752	270	144	0,81	0,12	0,07	2166
56	/~pz/zahnpage/vened.htm	Text zur Vorlesung “Bedeutende Bibliotheken - „Venedig“	Text	lg	3	1716	214	188	0,81	0,10	0,09	2118
57	/~pz/zahnpage/wienonb.htm	Text „Wien. Österreichische Nationalbibliothek“	Text	lg	3	1719	212	73	0,86	0,11	0,04	2004
58	/~mh/gedv/binzs.htm	Text „Binäres Zahlensystem“	Text	sm	4	1781	209	13	0,89	0,10	0,01	2003
59	/~kumlau/	Homepage von Prof. Umlauf	Home	lg	5	980	709	243	0,51	0,37	0,13	1932
60	/~wumsta/rehm.html	Start der Textsammlung „Margarete Rehm“	Orga	sm	5	639	408	875	0,33	0,21	0,46	1922
61	/~pz/big/cusanus/kues.htm	Text zu “Nikolaus von Kues“	Text	av	5	1505	216	141	0,81	0,12	0,08	1862
62	/~kumlau/handreichungen/h67/	Heft der „Berliner Handreichungen zur Bibliothekswiss.“	Text	lg	4	1045	547	221	0,58	0,30	0,12	1813
63	/~pbruhn/gruppe05.htm	Bibliographie „Frauen in Russland“	Text	lg	3	1599	180	11	0,89	0,10	0,01	1790
64	/~pbruhn/b-kunst.htm	Startseite zur Datenbank bzw. Bibliographie „Beutekunst“	DB Entry	av	5	465	579	663	0,27	0,34	0,39	1707
65	/inf/handrei.htm	Startseite der „Berliner Handreichungen zur Bibliothekswiss.“	Orga	lg	5	788	643	228	0,47	0,39	0,14	1659
66	/~wumsta/pub54.html	Publizierter Artikel “Die Wissenschaftlich	Text	av	4	1350	215	91	0,82	0,13	0,05	1656

		keit im Darwinismus“										
67	/~mbank/dienste/altavist.htm	Beschreibung des Funktionsumfangs der Altavista-Suchmaschine	Docu	av	4	1461	161	27	0,89	0,10	0,02	1649
68	/~pz/zahnpage/flordia.htm	Beschreibung von Dispositiven „Bibliothek – Florenz“	Text	lg	3	1427	188	12	0,88	0,12	0,01	1627
69	/~is/worduebung.html	Übungsaufgaben „Arbeit mit MS-Word“	Docu	av	2	1227	212	170	0,76	0,13	0,11	1609
70	/~pz/zahnpage/bi1_2g.htm	Text „Geschichte des Bibliothekswesens“	Text	av	4	1397	169	39	0,87	0,11	0,02	1605
71	/~wumsta/pub71.html	Publizierter Artikel „Photokina 1992“	Text	av	3	1344	220	9	0,85	0,14	0,01	1573
72	/~pz/zahnpage/vendia.htm	Text zum Thema „Venedig, Bibliothek“	Text	av	3	1426	145	1	0,91	0,09	0,00	1572
73	/~kumlau/handreichungen/h54/	Heft der „Berliner Handreichungen zur Bibliothekswiss.“	Text	sm	4	573	884	89	0,37	0,57	0,06	1546
74	/~mh/projekte/oeb/pb/copy.html	Adressliste mit Berliner Copy-Shops	Orga	av	3	1318	200	6	0,86	0,13	0,00	1524
75	/~wumsta/wistru/defalph.html	Alphabetische Liste mit Links eines digitalen Handbuchs	Orga	lg	3	1091	389	25	0,72	0,26	0,02	1505
76	/~hab/christine/gaudi1.html	Text zu „Antoni Gaudi“	Text	sm	3	1169	241	58	0,80	0,16	0,04	1468
77	/~kumlau/handreichungen/h65/	Heft der „Berliner Handreichungen zur Bibliothekswiss.“	Text	lg	4	1246	175	39	0,85	0,12	0,03	1460
78	/~pz/zahnpage/theafor.htm	Text zum Thema „Theaterwissenschaft“	Text	av	3	1299	121	4	0,91	0,08	0,00	1424
79	/~kumlau/handreichungen/h54/kapitel44.html	Kapitel der „Berliner Handreichungen zur Bibliothekswiss.“	Text	lg	4	1147	268	8	0,81	0,19	0,01	1423
80	/~wumsta/infopub/textbook/definitions/thesauindex.html	Index-Seite eines digitalen Thesausrus	Text	lg	4	1075	303	34	0,76	0,21	0,02	1412
81	/~wumsta/Milkau/karte.html	Linkliste einer Diasammlung	Orga	av	3	1245	156	11	0,88	0,11	0,01	1412
82	/amerika.html	Linkliste zu Ausbildungs-einrichtungen	Orga	av	5	1118	255	27	0,80	0,18	0,02	1400
83	/~wumsta/wistru/d0.htm	Alphabetische Liste mit Links eines digitalen Handbuchs (ältere Version)	Orga	av	3	991	267	79	0,74	0,20	0,06	1337
84	/~wumsta/pub74.html	Publizierter Artikel zum Thema „Informations-kompression“	Text	lg	3	1100	202	7	0,84	0,15	0,01	1309
85	/~hab/arnd/	Homepage des “Hildebrandslied” (inhaltlich identisch mit Top 14)	Home	sm	4	199	310	776	0,15	0,24	0,60	1285
86	/~kumlau/handreichun	Kapitel der	Text	lg	4	1161	115	1	0,91	0,09	0,00	1277

	gen/h34/d915.html	„Berliner Handreichungen zur Bibliothekswiss.“										
87	/~pz/zahnpage/geicdrom.htm	Literaturliste Prof. Zahn	Text	av	3	1132	134	9	0,89	0,11	0,01	1275
88	/~gtw/lehrgang/multimedia.html	Text zum Thema „Multimedia – Internet“	Text	av	2	1070	171	6	0,86	0,14	0,00	1247
89	/~fern/fernstudium/magister/magister.html	Startseite des Fernstudiums Magister	Home	av	5	843	384	20	0,68	0,31	0,02	1247
90	/~pbruhn/bszimmer.htm	Informationsseite der Bibliographie „Bernstein-zimmer“	Orga	sm	4	848	266	122	0,69	0,22	0,10	1236
91	/~pz/zahnpage/unilit.htm	Literaturliste Prof. Zahn	Text	lg	3	1105	111	9	0,90	0,09	0,01	1225
92	/~hab/christine/sagrada1.html	Text zum Thema „Sagrada Familia“	Text	sm	3	1000	172	11	0,85	0,15	0,01	1183
93	/~pz/zahnpage/hohand1.htm	Literaturliste mit Verweisen	Orga	av	4	990	104	50	0,87	0,09	0,04	1144
94	/~wumsta/dc/dc.html	Volltext-Dissertation	Text	lg	3	1015	113	5	0,90	0,10	0,00	1133
95	/~pz/florenz/medicdt.htm	Text zum Thema „Medici“	Text	sm	3	854	245	26	0,76	0,22	0,02	1125
96	/~wumsta/sgml/links.html	Linkliste zu den Themen „HTML, SGML, XML“	Orga	av	4	803	257	25	0,74	0,24	0,02	1085
97	/~wumsta/umbank.html	Hausarbeit zum Thema „Internet“	Text	lg	3	838	209	21	0,78	0,20	0,02	1068
98	/~fern/fernstudium/postgradual/ablauf.html	Studienablaufplan des Fernstudiums Magister	Orga	av	4	78	913	23	0,08	0,90	0,02	1014
99	/~hab/arnd/manuskript.html	Seite mit gescannter Handschrift	Text	sm	4	683	239	42	0,71	0,25	0,04	964
100	/~fern/fernstudium/magister/ablauf/	Studienablaufplan des Fernstudiums Postgradual	Orga	av	4	250	234	4	0,51	0,48	0,01	488

Tabelle 10-1: „Top 100-Liste“ (2002)

10.2 Top Queries

Rang	Queries April 2002	Häufigkeit 2002		Queries April 2000	Häufigkeit 2000
1	ascii code	745		fireball	1264
2	fernstudium	412		fernstudium	305
3	römische zahlen	364		florenz	226
4	ascii-code	263		ascii code	146
5	florenz	239		amerika	146
6	opac berlin	182		bat	136
7	dos befehle	157		bernsteinzimmer	123
8	hildebrandslied	147		inktomi	116
9	www.humboldt-uni.de	142		beutekunst	83
10	ms-dos befehle	131		bibliothek	74
11	ms dos befehle	115		pompeji	64
12	relationale datenbanken	108		bibliotheken	63
13	druckerei berlin	103		zelle	54
14	ascii	99		hildebrandslied	49
15	darwinismus	92		lycos	48
16	second hand berlin	82		deutsche bibliothek frankfurt	44
17	ms-dos	81		ascii code tabelle	37
18	sagrada familia	80		bundesangestelltentarifvertrag	35
19	amerika	77		bibliothekswissenschaft	35
20	bernsteinzimmer	77		java applet download	35
21	theaterkassen berlin	65		nationalbibliothek wien	34
22	binäres zahlensystem	65		fireball and alta vista	34
23	tabellarischer lebenslauf	64		alexandria ägypten and karte alexandria	28
24	beutekunst	62		hu berlin	26
25	theaterkasse berlin	61		amber room	24
26	mediengeschichte	56		sagrada familia	23
27	dezimalsystem	53		alta vista	23
28	ludwig xiv	53		Ägypten	22
29	antoni gaudi	50		relationale datenbanksysteme	22
30	bat eingruppierung	49		biblioteca apostolica vaticana	21
31	ms dos	49		antoni gaudi	21
32	römisches zahlensystem	48		news	20
33	mittelalter	48		nationalbibliothek	19
34	geschichte der kommunikation	47		java calendar	18
35	michelangelo	44		interpretation	18
36	druckereien berlin	44		römische zahlen	18
37	zahlensystem	44		ebenol	18
38	eingruppierung	42		opac berlin	17
39	berlin opac	42		praktikumsstellen	17
40	hans jonas	41		udssr	17
41	halbwertszeit	41		thomas müntzer	17
42	sgml tutorial	40		müntzer	16

43	rÃ¶mische zahlen	40		die zelle	16
44	dos-befehle	38		ascii-code	16
45	udssr	38		zellen	15
46	nikolaus von kues	38		humboldt universität berlin	15
47	sgml	35		oesterreichische nationalbibliothek	15
48	thomas müntzer	35		sgml	15
49	relationale datenbank	35		ascii code list	15
50	duales zahlensystem	33		zahlensystem	15
51	capitalis quadrata	33		amaryllis	15
52	gaudi	33		schriftgeschichte	15
53	rainer kuhlen	30		mediengeschichte	15
54	unformat	29		computerkurs	14
55	w@rez	27		nikolaus von kues	14
56	amaryllis	27		darwinismus	14
57	betriebswirtschaftslehre	27		staatsbibliothek berlin	14
58	handschrift	27		geschichte der kommunikation	14
59	ben kaden	27		donatello	14
60	sponsoringkonzept	26		british museum	14
61	word übungen	25		stabliniensystem	14
62	pienza	25		russland	14
63	bibliothekswissenschaft	24		eingruppierung	13
64	bundes-angestelltentarifvertrag	24		calendar applet	13
65	fernstudium master	24		selfhtml	13
66	distributionspolitik	24		katharina von medici	12
67	second hand hifi	24		römisches zahlensystem	12
68	17. Jun 53	24		sponsoringvertrag	12
69	wortstamm	24		gus	12
70	karte von england	24		meta opac	12
71	sponsoringvertrag	24		fernstudium architektur	12
72	praktikumsstellen	23		dezimalsystem	11
73	bwl	23		berlin opac	11
74	eingruppierung bat	23		deutsche bibliothek	11
75	xml beispiel	23		united states of amerika	11
76	gewohnheitsrecht	22		boole	11
77	theaterkassen	22		alta vista europe	11
78	allgemeine betriebswirtschaftslehre	22		relationale datenbanken	11
79	venedig karte	21		katodenstrahlröhre	11
80	arbeitsplatzbeschreibung	21		humboldt-universität berlin	11
81	konkurrenzanalyse	21		sgml tutorial	11
82	geburtstagsgedichte	21		zellaufbau	11
83	fernstudium webdesign	21		ägypten	11
84	kommunikation geschichte	21		ambrose bierce	10
85	xml beispiele	21		walther von der vogelweide	10
86	humboldt universität berlin	21		österreichische nationalbibliothek	10
87	din 69901	20		helvetica download free font	10
88	kaizen	20		bundes-	10

				angestelltentarifvertrag	
89	berliner theater	20		bachelor	10
90	grundlagen datenbanken	20		kirchenbau	10
91	kapitalwertmethode	20		vatikan	10
92	bühnenjahrbuch	20		suchmashine	10
93	literaturepochen	19		17. Jun 53	9
94	bat tätigkeitsmerkmale	19		infoseek	9
95	cosimo medici	19		sss	9
96	60. geburtstag	19		sponsoring	9
97	fernstudium berlin	19		pflanzen kölle	9
98	bat vergütungsgruppen	19		michelangelo buonarroti	9
99	theaterkassen in berlin	19		humboldt universität	9
100	bewährungsaufstieg	19		monte amiata	8

Tabelle 10-2: Top 100 Queries (April 2002, 2000)

10.3 Top Backlinks

Rang	Entries über Backlinks (Websites 2002)	Entries über Backlinks (Websites 2000)
1	http://www.ub.hu-berlin.de/	http://www.hu-berlin.de/
2	http://www.hu-berlin.de/	http://www.ub.hu-berlin.de/
3	http:// www.physik.fu-berlin.de/	http://de.dir.yahoo.com/
4	http://de.dir.yahoo.com/	http://amor.rz.hu-berlin.de/
5	http://www2.hu-berlin.de/	http://www.ub.tu-berlin.de/
6	http://www.sewanee.edu/	http://www.hbz-nrw.de/
7	http://www.dbi-berlin.de/	http://www.sewanee.edu/
8	http://www.fh-potsdam.de/	http://www.dbi-berlin.de/
9	http://www.hbz-nrw.de/	http://www.tu-dresden.de/
10	http://www.niester.de/	http://www.uni-koblenz.de/
11	http://www.udk-berlin.de/	http://www.zlb.de/
12	http://home.t-online.de/	http://www.ub.fu-berlin.de/
13	http://www.ub.tu-berlin.de/	http://www2.hu-berlin.de/
14	http://amor.rz.hu-berlin.de/	http://www.beutekunst.de/
15	http://www.pitt.edu/	http://www.rz.hu-berlin.de/
16	http://www.ub.fu-berlin.de/	http://www.fh-potsdam.de/
17	http://www.uni-koblenz.de/	http://www.bsz-bw.de/
18	http:// www.zfuw.uni-koblenz.de/	http:// informant.dartmouth.edu/
19	http://www.bsz-bw.de/	http:// www.physik.fu-berlin.de/
20	http:// www.phil.uni-erlangen.de/	http://rubriken.fireball.de/
21	http://www.tu-harburg.de/	http://www2.rz.hu-berlin.de/
22	http:// www.geschichte.hu-berlin.de/	http:// www.informatik.hu-berlin.de/
23	http:// altedrucke.sbb.spk-berlin.de/	http://www.ub.uni-siegen.de/
24	http:// did.mat.uni-bayreuth.de/	http:// www.geschichte.hu-berlin.de/
25	http://www.geocities.com/	http://www.pitt.edu/
26	http://www.wr-unterricht.de/	http:// server1.schule.uni-halle.de/
27	http://www.zlb.de/	http://www.biologie-lk.de/
28	http://www.geschenkzeitung.de/	http:// www.sigel.spk-berlin.de/
29	http://www.freenet.de/	http://www.infobroker.de/
30	http://www.gragert.de/	http://www.tu-berlin.de/
31	http:// bak-information.ub.tu-berlin.de/	http:// www.diff.uni-tuebingen.de/
32	http://bibliothek.bbaw.de/	http://home.t-online.de/
33	http:// www.inf-wiss.uni-konstanz.de/	http:// www.ub.uni-bielefeld.de/
34	http:// www.bibliothek.uni-augsburg.de/	http://info.ub.uni-potsdam.de/
35	http:// www.wissen-unserer-zeit.de/	http://www.hbi-stuttgart.de/
36	http:// www.nationalsozialismus.de/	http:// studweb.euv-frankfurt-o.de/
37	http://www.jadu.de/	http:// www.phil.uni-erlangen.de/
38	http://www.uni-leipzig.de/	http://www.library.uiuc.edu/
39	http://www2.rz.hu-berlin.de/	http://www.suub.uni-bremen.de/
40	http://www.iaea.org	http://wwwl.sbb.spk-berlin.de/
41	http://forge.fh-potsdam.de/	http://www.gragert.de/
42	http://www.itcs.com/	http://www.oei.fu-berlin.de/
43	http://subscriber.chello.at/	http://www.fu-berlin.de/

44	http://www.fu-berlin.de/	http://www.tu-harburg.de/
45	http://directory.google.com/	http:// www.rz.uni-karlsruhe.de/
46	http://www.uni-duesseldorf.de/	http://www.uni-duesseldorf.de/
47	http://www.library.uiuc.edu/	http://www.hausarbeiten.de/
48	https://www.voebb.de/	http://www-opac.bib-bvb.de/
49	http://www.ik.fh-hannover.de/	http://subscriber.chello.at/
50	http://www.rechenzentrum.org/	http://www.voebb.de/
51	http://nibis.ni.schule.de/	http://dbv-berlin.zlb.de/
52	http://www.htwk-leipzig.de/	http:// www.papyrus-germany.com/
53	http://www.vdb-online.org/	http://www.itaw.hu-berlin.de/
54	http://orb.rhodes.edu/	http://www.dgd.de/
55	http://www.latine.de/	http://www.kloster-metten.de/
56	http://www.clickfish.com/	http://www.goethe.de/
57	http://www.beutekunst.de/	http://info.dai.bund.de/
58	http://info.ub.uni-potsdam.de/	http://www.itcs.com/
59	http://www.biologie-lk.de/	http://www.konbib.nl/
60	http://www.datenquell.de/	http:// www.phil-fak.uni-duesseldorf.de/
61	http:// www.ub.uni-bielefeld.de/	http://www.webtop.com/
62	http://members.chello.at/	http://www.german.sbc.edu/
63	http://www.hbi-stuttgart.de/	http:// www.mathematik.uni-halle.de/
64	http://www.blind-guardian.com/	http://www.fernuni-hagen.de/
65	http://www.dgd.de/	http://www.fbi.fh-koeln.de/
66	http://vivisimo.com/	http://arama.superonline.com/
67	http://InformationR.net/	http://snake.cs.tu-berlin.de/
68	http://www.emporia.edu/	http:// www.physik.hu-berlin.de/
69	http://www.fh-augsburg.de/	http:// www.nationalsozialismus.de/
70	http://www.pinselpark.de/	http://www.yale.edu/
71	http://www.lapkereso.hu/	http://www.uni-leipzig.de/
72	http:// www.physik.hu-berlin.de/	http:// bak-information.ub.tu-berlin.de/
73	http://www.parthey.de/	http://www.shef.ac.uk/
74	http://www.ita.rwth-aachen.de/	http://www.zfuw.uni-kl.de/
75	http://ripley.wo.sbc.edu/	http://www.nettz.de/
76	http:// homepages.uni-tuebingen.de/	http://www.latine.de/
77	http://www.uni-frankfurt.de/	http://www.fortunecity.de/
78	http://www.oeaw.ac.at/	http://bibliothek.bbaw.de/
79	http://langlab.uta.edu/	http://www.biost.de/
80	http:// www.Informatik.hu-berlin.de/	http:// web.ub.uni-greifswald.de/
81	http://www.stub.unibe.ch/	http://www.tfh-berlin.de/
82	http://www.bbpp.de/	http://www.lepak.com/
83	http://www.german.sbc.edu/	http://www.ewetel.net/
84	http://www.kloster-metten.de/	http:// www.bibliothek.uni-regensburg.de/
85	http://www.uni-koeln.de/	http://www.iaea.org
86	http:// www.wissenschaftsforschung.de/	http://www.b.shuttle.de/
87	http:// www.sigel.spk-berlin.de/	http:// homepages.uni-tuebingen.de/
88	http:// www.kreimeier-online.de/	http://www.uni-kiel.de/
89	http://www.infobroker.de/	http://www.htwk-leipzig.de/
90	http://www.uni-kiel.de/	http:// wwwtcs.inf.tu-dresden.de/
91	http://www.his.uni.torun.pl/	http://www.s-line.de/

92	http://www.bib-bvb.de/	http://www.kobv.de/
93	http:// www.diff.uni-tuebingen.de/	http://www.unibw-hamburg.de/
94	http:// www.hildebrandtfoundation.org/	http://www.ik.fh-hannover.de/
95	http://magazine.orf.at/	http://www.t-online.de/
96	http://www.mypage.bluewin.ch/	http:// www.swbv.uni-konstanz.de/
97	http:// www.gym-muentzer.bildung-lsa.de/	http://www.uni-koeln.de/
98	http://www.yale.edu/	http://www.toeppner.de/
99	http://www.sims.berkeley.edu/	http://www.geschi.de/
100	http:// www.papyrus-germany.com/	http:// www.fhsbib.uni-frankfurt.de/

Tabelle 10-3: Top 100 Backlinks (Websites) (2002, 2000)