

Die Logfiles des IB-Webserver - Ergebnisse einer Magisterarbeit aus der Webometrie

Referent: Philipp Mayr

Berliner Bibliothekswissenschaftliches Kolloquium,
Institut für Bibliothekswissenschaft, HU Berlin
am 06.Juli 2004



Agenda

- Logfiles, Logfile Analyse
 - Hintergrund & Grundlagen, Terminologie
 - Einschränkungen und Potenziale
 - Anwendung
 - Webometrie (Einführung)
- Magisterarbeit
 - Konzeption
 - Ergebnisse
 - Diskussion
- Ausblick
- Live-Demo des Web Entry Miner
- Fragen & Diskussion



Hintergrund

- Informationswissenschaftliches Interesse an der Untersuchung des Online Suchverhaltens
 - Interface Design
 - Angebotsoptimierung
- Untersuchungen des Informationsverhalten (z.B. Opacs, Host)
 - bislang hauptsächlich an kleinen homogenen Laborgruppen (meist akademische Nutzer)
 - Kaum interdisziplinäre Untersuchungen



Motivation

- Motivation für eine Magisterarbeit in diesem Bereich
 - Herausforderung Webdatenanalyse
 - Webuser und Inhalte sind sehr heterogen und wenig erforscht
 - Web Log Files bieten eine sehr gute Möglichkeit den Web Use (Online Verhalten) zu untersuchen
- Ziel: Entwicklung eines Instruments zur Visualisierung & Bezifferung von Zugänglichkeit, Sichtbarkeit, Verlinkung im Web



IB-Website 1998

Humboldt-Universität zu Berlin; Institut für Bibliothekswissenschaft - Microsoft Internet Explorer

Adresse <http://web.archive.org/web/19980707230622/http://www.ib.hu-berlin.de/>




Das Institut für Bibliothekswissenschaft begrüßt Sie auf seiner Homepage

Institut für Bibliothekswissenschaft der Humboldt-Universität zu Berlin

Eingang Dorotheenstraße 26

Geschäftsführender Direktor [Prof. Dr. Robert Funk](#)

Informationen zum Institut

- [Zur Geschichte der Bibliothekswissenschaft in Berlin](#)
- [Mitarbeiter](#)
- [Ausbildung \(Direktstudium\)](#)
- [Ausbildung \(Fernstudium\)](#)
- [Datenbankangebote des Instituts](#)
- [Berliner Bibliothekswissenschaftliches Kolloquium \(BBK\)](#)
- [Publikationen des IB](#)
- [Fachschaft des IB](#)
- [Homepages der Studenten](#)
- [NEWS](#)

Informationen für die Bibliothekspraxis

- [Tagungen, Wissenschaftliche Veranstaltungen](#)
- [Weiterbildungsveranstaltungen](#)
- [Suchmöglichkeiten im Internet](#)
- [Bibliotheken, OPACs](#)
- [Zeitschriften](#)
- [Nachschlagewerke](#)
- [Informationscenter, Datenbankanbieter](#)
- [Institutionen](#)

Bibliotheks-, Dokumentations-, Informationswissenschaft weltweit

- [Amerika](#)

Fertig Internet

Screenshot
1998

Vgl. Internetarchive
www.web.archive.org



Grundlagen

Was ist ein Logfile?

- „Ein Logfile ist das automatisch erstellte Protokoll aller oder bestimmter Aktionen von einem oder mehreren Nutzern an einem Rechner, ohne dass diese davon etwas mitbekommen oder ihre Arbeit beeinflusst wird.“ *[Wikipedia, 2004]*
- „Machine-generated records of user activity“ (Nicholas et al., 2002)
- Strukturiert aufgebaute, webserverabhängige Protokolldatei (ASCII-Format) zur nicht-reaktiven Nutzungsmessung
- „Eine Goldmine“



Beispiel Logzeile

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

- Logfile-Felder im Format NCSA Combined Log Format (Apache Webserver)
 - host, ident, authuser, date, request, status, bytes, referer,
useragent



Grundlagen



Allgemeine Fragen die Logfiles beantworten können

- Von welchen Seiten kommen die Besucher? (Interlinking)
- Welche Suchmaschinen bzw. Suchwörter werden genutzt?
- Ist die Site bzw. die Seite gut bei Suchmaschinen gerankt? (Sichtbarkeit)
- Wie lange bleiben die Besucher auf einzelnen Webseiten?
- Wie viele Seiten rufen sie dabei auf? Welche Nutzungspfade werden genutzt?
- Wie lauten IP-Adresse und Hostname (Top Level Domains) des Nutzers?
- Welchen Browser/Betriebssystem hat der Besucher benutzt?
- Kommen die Besucher zurück?



Begriffe – Logfile Terminologie

- **Hit:** kleinste Maßeinheit für einen Zugriff (wenig Aussagekraft)
- **Page View (Impression):** Aufruf einer Webseite zu einem best. Zeitpunkt
- **Visit:** Aufrufe eines Besuchers innerhalb eines festgelegten Zeitraums (z.B. 30min), Server Session
- **Visitor:** Besucher, fehleranfällig
- **User Session:** Abfolge von Seitenaufrufen eines Users
- **Episode:** Abfolge von Seitenaufrufen innerhalb einer User Session
- Andere Maße: Time Online, Downloads, ...



allg. Nutzungsmaße - Metriken

- Durchschnittliche Anzahl der angesehenen Seiten/Dokumente pro Besuch
- Durchschnittliche Ansichtszeit einer Seite/Dokument
- Wiederbesuchswerte
- Länge der Sessions/Besuche (Anzahl der Seiten, Zeit)
- Anzahl und Typen der Besucher



Analyse - Schwerpunkte

- Einfache Nutzungsmaße (z.B. Hit, View, Visits)
- Besuchercharakteristiken (Nationalität, Organisation, Vielkäufer, Subscriber)
- Suchcharakteristiken (Themen, Trends, Suchbegriffe vs. Informationsbedürfnis)
- Benutzungspfade (z.B. Muster)
- Kommerzielle Analysen (z.B. Web Mining, Kaufverhalten)
- wissenschaftliche Fragestellungen



Beispiele Hits

Beispiele für Hits im Logfile

- Bilder (keine relevanten Hits)

userb - - [01/Apr/2003:11:37:30 +0200] "GET /pics/**pfeil.gif** HTTP/1.0" 200 51

"http://www.disinfojournal.net/index.html" "Mozilla/3.0 (compatible; WebCapture 1.0; Windows)"

userb - - [01/Apr/2003:11:37:30 +0200] "GET /pics/**question.gif** HTTP/1.0" 200 999

"http://www.disinfojournal.net/index.html" "Mozilla/3.0 (compatible; WebCapture 1.0; Windows)"

- Suchmaschinen-Robots (keine relevanten Hits)

si3001.inktomisearch.com - - [01/Apr/2003:13:35:16 +0200] "GET /robots.txt HTTP/1.0" 200 26 "-" "Mozilla/5.0

(**Slurp**/si; slurp@inktomi.com; http://www.inktomi.com/slurp.html)"

64.68.82.57 - - [01/Apr/2003:10:14:13 +0200] "GET / HTTP/1.0" 200 4614 "-" "**Googlebot**/2.1

(+http://www.googlebot.com/bot.html)"

- Download (2 Hits, aber nur eine Datei)

usera - - [01/Apr/2003:11:49:00 +0200] "GET /downloads/**0103_abstracts.pdf** HTTP/1.0" 206 **28832** "-"

"Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 4.0)"

usera - - [01/Apr/2003:11:49:01 +0200] "GET /downloads/**0103_abstracts.pdf** HTTP/1.0" 206 **1024** "-"

"Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 4.0)"



Weitere Beispiele

9 Hits, 4 Views, 2 Visit, 2 Visitors, 2 User Sessions

```

203.30.5.145 - - [01/Jun/1999:03:09:21 -0600] "GET /Calls/OWOM.html HTTP/1.0" 200
3942 "http://www.lycos.com/cgi-bin/pursuit?query=advertising+psychology-
&maxhits=20&cat=dir" "Mozilla/4.5 [en] (Win98; I)"
203.30.5.145 - - [01/Jun/1999:03:09:23 -0600] "GET /Calls/Images/earthani.gif
HTTP/1.0" 200 10689 "http://www.acr-news.org/Calls/OWOM.html" "Mozilla/4.5 [en]
(Win98; I)"
203.30.5.145 - - [01/Jun/1999:03:09:24 -0600] "GET /Calls/Images/line.gif
HTTP/1.0" 200 190 "http://www.acr-news.org/Calls/OWOM.html" "Mozilla/4.5 [en]
(Win98; I)"
203.252.234.33 - - [01/Jun/1999:03:12:31 -0600] "GET / HTTP/1.0" 200 4980 ""
"Mozilla/4.06 [en] (Win95; I)"
203.252.234.33 - - [01/Jun/1999:03:12:35 -0600] "GET /Images/line.gif HTTP/1.0"
200 190 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"
203.252.234.33 - - [01/Jun/1999:03:12:35 -0600] "GET /Images/red.gif HTTP/1.0" 200
104 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"
203.252.234.33 - - [01/Jun/1999:03:12:35 -0600] "GET /Images/earthani.gif
HTTP/1.0" 200 10689 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"
203.252.234.33 - - [01/Jun/1999:03:13:11 -0600] "GET /CP.html HTTP/1.0" 200 3218
"http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"
203.30.5.145 - - [01/Jun/1999:03:13:25 -0600] "GET /Calls/AWAC.html HTTP/1.0" 200
104 "http://www.acr-news.org/Calls/OWOM.html" "Mozilla/4.5 [en] (Win98; I)"

```



Probleme

- Die Analyse von Web Logdaten öffentlicher Webserver steckt voller Schwierigkeiten und Probleme
- Eindeutige Besucheridentifikation sehr schwierig
 - Virtual users ungleich end user (keine Individuen, sondern lediglich Computer)
- Firewall und Caching-Problematik
 - Welche Transaktionen wurden nicht aufgezeichnet?
 - under/over reporting Problematik
- HTTP-Protokoll ist zustandslos
 - Registrierung von atomaren Transaktionen (kein Logoff)
 - Konzeption als technisches Protokoll, nicht zur Untersuchung von Onlinebehavior und -retrieval



Probleme cont.

- URL's häufig nichtssagend und müssen daher semantisch angereichert werden (z.B. Parameteranalyse)
 - Größe der Logfiles (AOL.com ca. 400 Views/s)
 - Zugang zu Logfiles anderer Webserver
 - Anomalien innerhalb der Daten
 - Filtern und Säuberung der Daten vor der Analyse dringend erforderlich (Hit ist nicht gleich Hit)
- „logs as dirty information source“ [Nicholas & Huntington, 2003]



Einschränkungen

- In der Regel nur Analyse einer Website bzw. eines Webservers
- Kaum richtige Standards, eher Richtwerte
- Standardauswertungen (Webtrends) liefern i.d.R. nur sehr allgemeine Aussagen und weisen nicht auf die Einschränkungen hin
- Beachtung von Datenschutz-Richtlinien
- Besucherklassifikation noch sehr schwierig
 - IP reverse lookup recht ungenau [Nicholas et al., 2003]
 - Ideal sind authentifizierte Besucher



Potenziale

- Konkrete Nutzung(smaße) von internen und externen Linkstrukturen gegenüber der Untersuchung der Existenz von Verlinkungen (Webometrie)
- große und heterogene Besucherschaft
 - Makro und Mikro-Untersuchungen
- Datenerhebung 24 h und 7 Tage
- Sehr zeitnahe Analysen und Aussagen
- Untersuchung des Informationsverhalten, -aufnahmeverhalten im Web aus einer Website-Perspektive
- Hinweise zur Optimierung, Evaluation (?) und Adaption von Websites



Typische allg. Auswertungen

1. Wie viel Traffic erhält die Site, Directory, Page, ...?
2. Die beliebtesten Bereiche der Site
3. Die wichtigsten Einstiegsseiten der Site
4. Woher kommen die Besucher?
5. Welche Suchmaschinen bringen Traffic?
6. Wird die Navigation richtig eingesetzt? Navigationspfade?
7. Steigen die Besucher zu früh aus (Exit Pages)? Wo sind Löcher in meiner Site? Single Access Pages
8. Kommen die Besucher zurück? Neue vs. alte Besucher
9. Gibt es technische Probleme mit der Website?

March April

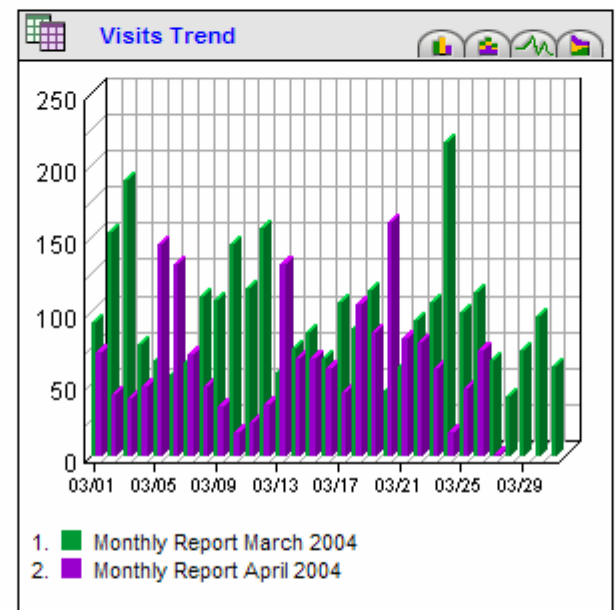
S	M	T	W	T	F	S	Q1	S	M	T	W	T	F	S	Q2
29	1	2	3	4	5	6	W10	28	29	30	31	1	2	3	W14
7	8	9	10	11	12	13	W11	4	5	6	7	8	9	10	W15
14	15	16	17	18	19	20	W12	11	12	13	14	15	16	17	W16
21	22	23	24	25	26	27	W13	18	19	20	21	22	23	24	W17
28	29	30	31	1	2	3	W14	25	26	27	28	29	30	1	W18

2004 2004

Table of Contents

- Overview
- Commerce
- Marketing
- Visitors
- Pages and Files
- Parameter Analysis
- Navigation
- Technical
- Activity
- Browsers
- Glossary

the corresponding page.

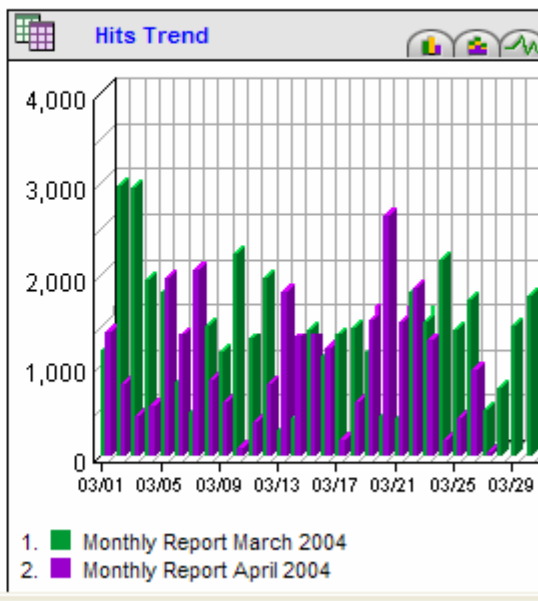


Visitor Summary

	Monthly Report March 2004	Monthly Report April 2004	
Unique Visitors	1,659	1,255	-24.35%
Visitors Who Visited Once	1,380	1,060	-23.19%
Visitors Who Visited More Than Once	279	195	-30.11%
Average Visits per Visitor	1.81	1.43	-20.99%

Visit Summary

	Monthly Report March 2004	Monthly Report April 2004	
Visits	3,011	1,795	-40.39% ▼
Average per Day	97	59	-39.18% ▼
Average Visit Length	00:05:33	00:05:10	-6.91% ▼
Median Visit Length	00:01:55	00:01:56	+0.87% ▲
International Visits	0.00%	0.00%	0.00%
Visits of Unknown Origin	100.00%	100.00%	0.00%





Loganalyse

Schritte zur eigene Anwendungen

1. Sammlung der Logdaten
2. Fehlerbereinigung (z.B. Codierung)
3. Ev. Stichprobe ziehen, wenn Daten zu umfangreich
4. Preprocessing – Data Cleaning (Bilder, Robots, ... entfernen)
5. Parsing und Import der Zeilen in eine separate Anwendung (z.B. Datenbank)
6. Analyse und weiteres Parsen und Extrahieren v. spez. Daten
7. Wiederholung der Schritte 1-6



Loganalyse cont.

Beispiel eines geparsten Files [Nicholas et al., 1999]

Table 1. The parsed and processed file: a sample.

<i>IP address</i>	<i>date of search</i>	<i>hrs.</i>	<i>mins.</i>	<i>secs.</i>	<i>GMT offset</i>	<i>destination page</i>	<i>HTTP version</i>	<i>Service status code</i>	<i>Bytes sent</i>
134.117.1.22	11/Mar/1998	00	20	49	0000	regis	1.0"	200	926
209.63.114.28	11/Mar/1998	00	24	31	0000	home	1.1"	200	2801
195.44.0.224	11/Mar/1998	00	26	11	0000	front	1.0"	200	1424
160.96.179.5	11/Mar/1998	02	20	07	0000	fpcon	1.0"	200	8880
153.35.110.150	11/Mar/1998	02	25	45	0000	timbi	1.1"	200	9531
207.102.33.157	11/Mar/1998	02	26	16	0000	timnw	1.0"	200	9930
166.72.168.140	11/Mar/1998	02	41	34	0000	timfg	1.0"	200	9151
203.10.130.1	11/Mar/1998	03	21	33	0000	timnw	1.0"	200	14442
198.7.150.54	11/Mar/1998	03	24	49	0000	home	1.0"	200	2801
198.53.4.182	11/Mar/1998	03	30	04	0000	timin	1.0"	200	13581
202.242.209.55	11/Mar/1998	03	30	59	0000	timin	1.0"	200	10896



Webometrics

- Forschungsgebiet seit Mitte der 90er
- Starke methodische Verwandtschaft zur Informetrie und Bibliometrie
- Ziel: neue Regeln, Charakterisierungen und wissenschaftliche Ergebnisse über das Netzwerk-Phänomen Internet als Zitationsnetzwerk
- “We define webometrics ... to be the quantitative study of web-related phenomena“ [Thelwall, Vaughan, Björneborn, 2004]
- Disziplin: Information Science, Communication Studies, Statistical Physics, Computer Science



Webometrics cont.

- Teilbereiche der Webometrie/Cybermetrie:
 - Suchmaschinenanalyse (Abdeckung, Performance, Qualität, ...)
 - Inhaltsanalyse
 - Linkanalysen
 - z.B. Web Impact Factor [Ingwersen, 1998]
 - Lotka-Law in der Verteilung der Top Level Domains der Treffer [Rousseau, 1997]
 - Onlineverhalten
 - Knowledge discovery in databases (KDD)
 - Web Mining (Web pages, link structures, user's information behaviour)
 - Graphen, „small-world“-Phänomene, Path analysis
 - Logfile Analysis (Suchmaschinen, Websites)



Magisterarbeit

- **Titel:** Entwicklung und Test einer logfilebasierten Metrik zur Analyse von Website Entries am Beispiel einer akademischen Universitäts-Website
- **Betreuung:** Prof. Umstätter, PD Dr. Wagner-Döbler, Dipl. Math. Heinz
- **Abgabe** Dezember 2003
- **Vorgängeruntersuchung** am IB von Heike Oldenburg, 2003
 - Unterscheidung der 3 Navigationsarten: Suchmaschine, Direkt, Backlink
 - Untersuchung des Navigationsverhaltens (Click-Stream, Pfadlängen)
 - Backlink-Besucher rufen mehr Seiten auf
- **Thematische Einordnung:** Web Usage Mining, Webometrie, Logfile Analyse
- **AOIR 2004, Workshop:** The web as a mirror of scientific and technical achievements: issues in access and measurement (Mike Thelwall et al.)



Magisterarbeit - Fragestellung

Links bzw. Suchmaschinen-Trefferlisten als Untersuchungsgebiet etabliert

- Welche informatrischen Potenziale verbergen sich in den Logdaten freizugänglicher Webserver?
- Lassen sich Gesetzmäßigkeiten/Regelhaftigkeiten für das Informationsverhalten im Web in den Web Logdaten eines Webservers finden und beschreiben?
- Kann über die Feststellung der Nutzungshäufigkeiten beliebiger Websiteinhalte auf die Bedeutung dieser Entitäten geschlossen werden?



Magisterarbeit - Idee

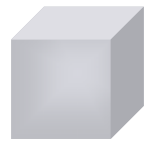
- Darstellung neuer Aspekte und Analyse-Möglichkeiten für Standard Web Log Files
- Beschreibung d. Hyperlink-Nutzung und Online-Navigation via Logfile
- Untersuchungsfokus der Studie: Entwicklung eines Konzepts zur Messung und Visualisierung der Sichtbarkeit, Zugänglichkeit und Verlinkung von Websites bzw. Website Entitäten
- Spezielle Anwendung: Web Entries am Beispiel einer umfangreichen deutschsprachigen akademischen Website



Magisterarbeit - Konzept

Einstiegsseiten

- Entries = Zugriff auf die erste Seite (Einstiegsseite, Entry Point, Entry Page) eines Besuchs
 - Alle folgenden Zugriffe sind Navigationszugriffe
- Hintergrund: Entry pages sind entscheidend für die Länge des Besuchs und ev. Wiederbesuch
- Kenntnis der Entry pages ist insbesondere für umfangreiche Websites interessant



Magisterarbeit - Konzept

- Untersuchung der Einstiegsseiten

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



Magisterarbeit - Konzept

Website Entitäten

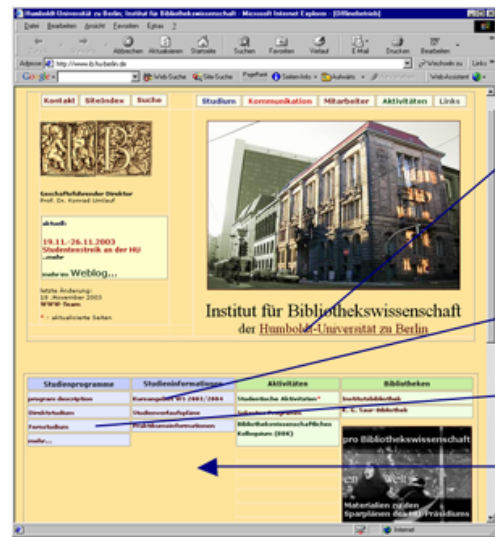
- Entitäten-Konzept ermöglicht die Untersuchung granularer Website-Bestandteile
- vgl. „Advanced Web Document Models“, Thelwall & Harries, 2003) als Methodologie zur Reliabilitätsprüfung von Linkzählungen
 - Site = alle Verzeichnisse samt Webseiten
(/*)
 - Directory = einzelne Verzeichnisse eines Webserver
(/dir/*)
 - Page = einzelne URL's
(/index.html)



Magisterarbeit - Konzept

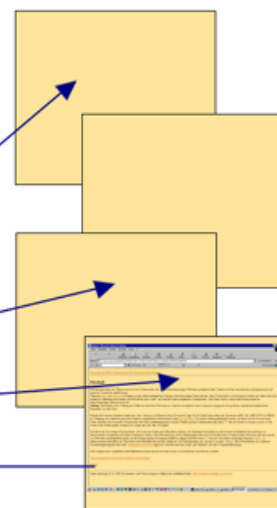
- Darstellung von Website Entitäten (Site, Directory, Page)

Site

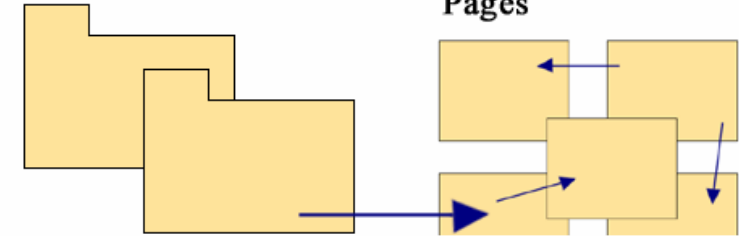


Homepage

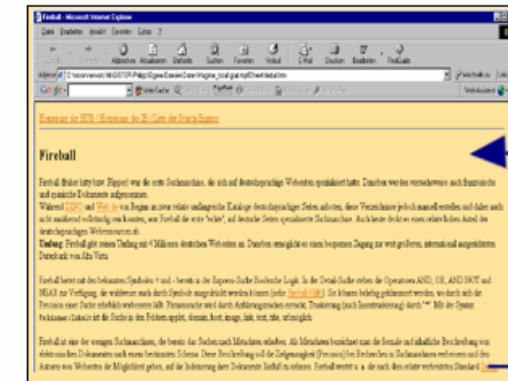
Pages



Directory



Page

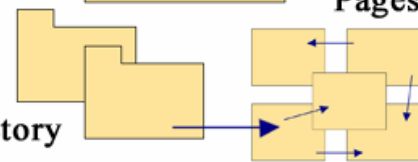


backlink

link

Directory

Pages





Magisterarbeit - Konzept

Navigationarten im Web

1. Suchmaschinen = Verfolgen eines Links, der durch eine Suchanfrage (Query) mittels Suchmaschine generiert wurde (z.B. Suchanfrage „Bibliothekswissenschaft“)
2. Direkt = Direktes Aufrufen einer Seite (z.B. Bookmark, Verlauf, Browserfeatures)
3. Backlinks, Inlinks, Referenzen = Verfolgen eines externen Links (Backlinks)

Untersuchungen nur möglich, wenn das Logfile im
Combined Logformat (referer)!



Magisterarbeit - Konzept

Identifikation der **Navigationsarten** im Logfile

1. Navigationsart Suchmaschine (Logfile-Eintrag):

```
"http://www.google.de/search?q=fernstudium  
&start=20&sa=N,"
```

2. Navigationsart Direkt (Logfile-Eintrag):

```
" _ "
```

3. Navigationsart Backlink (Logfile-Eintrag):

```
"http://www.inf-wiss.uni-  
konstanz.de/FG/IV/mitarbeiter.html"
```



Magisterarbeit - Konzept

Web Entry Factors (WEF) = detaillierte Zahlen der Sichtbarkeit, Zugänglichkeit und Verlinkung von Web Entitäten

- WEF als Aggregat allgemeiner Nutzungshäufigkeiten
- Sind eigentlich keine Faktoren sondern Anteilsraten

$$WEF_s(URL) =$$

$$Entries_s(URL) / Entries_{total}(URL)$$

$$WEF_s + WEF_r + WEF_d = 1$$

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: Entry-Werte und WEF-Faktoren für unterschiedliche Webentitäten



Magisterarbeit - Konzept

Klassifikation der Top 100 nach Haas & Grams, 2000

- **Organizational** (Orga) = Seiten, die Hilfestellungen in bezug auf Webseitestructur und Seitennavigation bieten (z.B. Indexseiten, Sitemap, Inhaltsverzeichnisse)
- **Documentation** (Docu) = Seiten, die Beschreibungen und Erklärungen zu einem spezifischen Thema enthalten und als Referenz fungieren (z.B. FAQ, Tutorial, How-to)
- **Text** (Text) = Seiten, die beliebige Texte von Personen, Gruppen etc. enthalten (z.B. Artikel, Paper, Forschungsbericht, Vertrag, Bibliographie, Lebenslauf)
- **Home Page** (Home) = Seiten, die eine Organisation oder Person einführend darstellen und Links zu weiterführenden Seiten enthalten (z.B. Berufliche bzw. private Home Pages)
- **Database Entry** (DB_Entry) = Seiten, die den Einstieg zu einer Datenbank darstellen (z.B. Bibliografische Daten, Katalogdaten)
- **Multimedia, Tools** (auf IB-Site nicht in den Top 100)



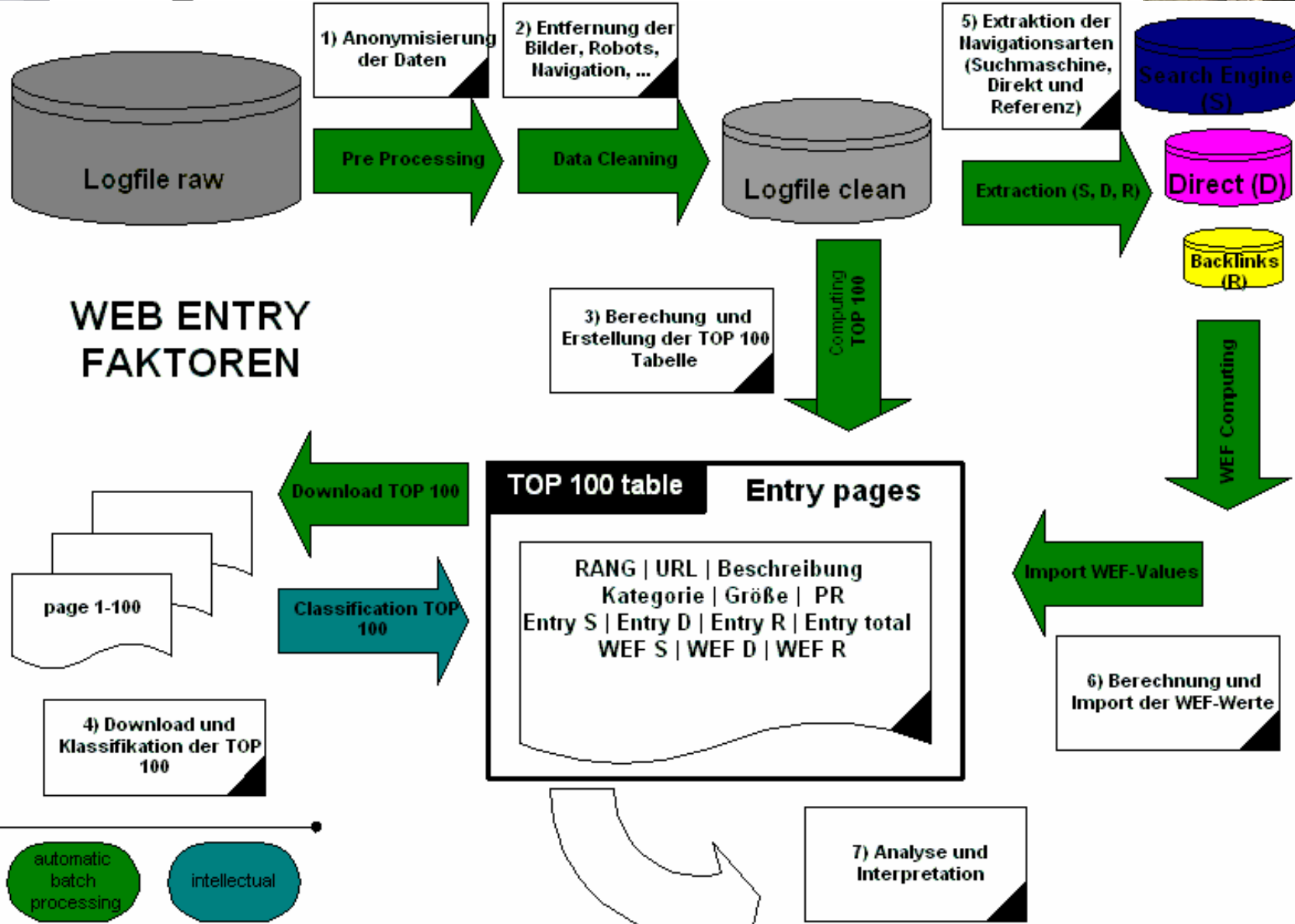
Magisterarbeit - Konzept

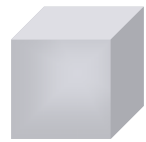
Weitere Merkmale der Top 100

- Größenswert einer Seite
 - groß ($lg = > 43329$ byte)
 - durchschnittlich ($av = > 5182 < 43329$ byte)
 - klein ($sm = > 1 < 5182$ byte)
- PageRang-Wert einer Seite
 - Über die Google Toolbar (toolbar.google.com)
 - Werte von 0-10



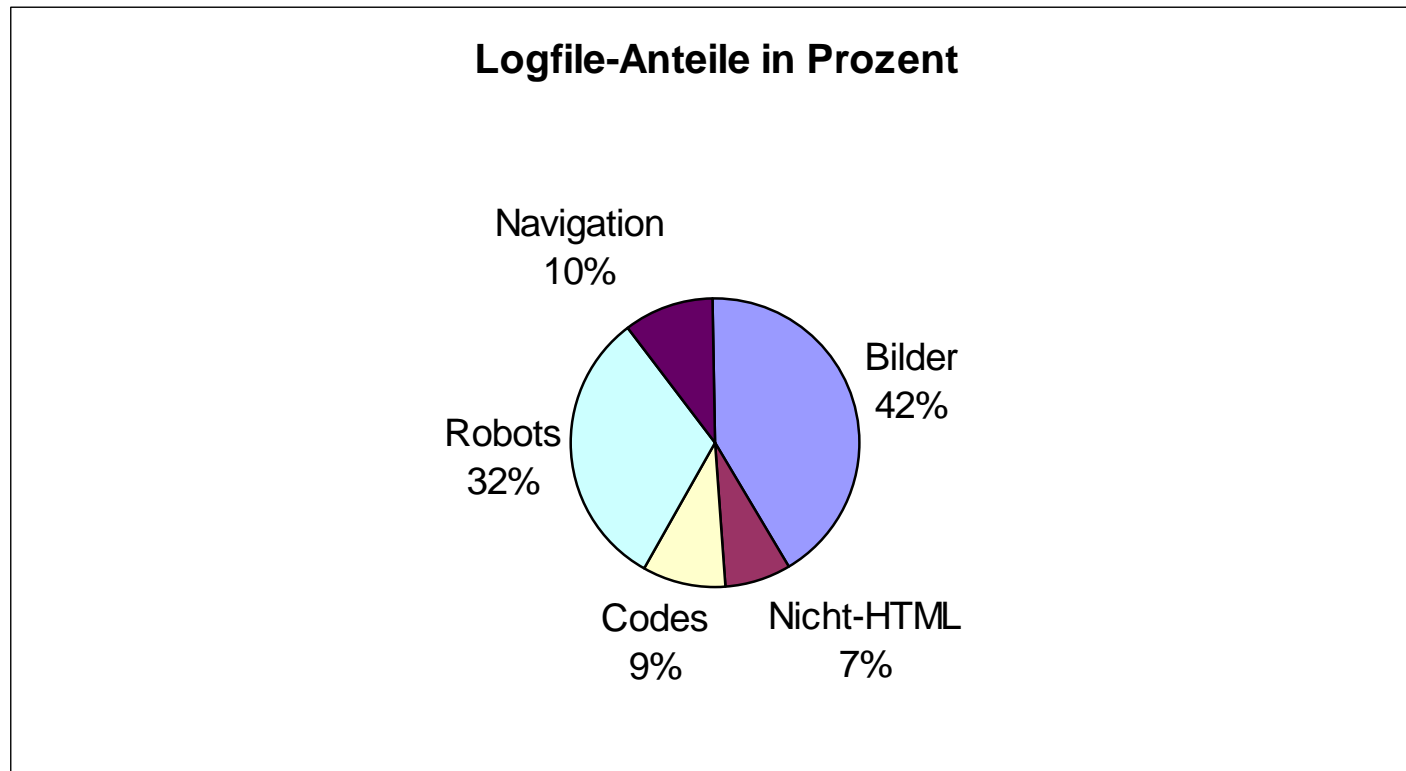
Magisterarbeit - Ablauf





Magisterarbeit - Ergebnisse

- Data-cleaning und Logfile-Anteile
 - Logfile 2000 = 4.715.037 Zeilen (449.727 Zeilen gefiltert)
 - Logfile 2002 = 6.739.902 Zeilen (820.678 Zeilen gefiltert)





Magisterarbeit - Ergebnisse

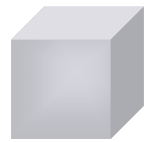
- Verteilung der Navigationsarten
 - 59% der Entries über Suchmaschinen Queries (2002)
 - 34% der Entries über direkte Navigation (2002)
 - 7% der Entries über Backlinks (2002)
- sehr ähnliche Werte für die Jahre 2000 und 2002
- Top 100 Entry pages (2002)
 - 58 text pages (text)
 - 21 organisational pages (orga)
 - 4 documentation (docu)
 - 6 database entries (db_entry) -> klassische Einstiegsseite
 - 10 homepages (home) -> klassische Einstiegsseite
- median WEF_SearchEngine = 0.81, median WEF_Direct = 0.15, median WEF_Backlink = 0.02 für die Top100



Magisterarbeit - Ergebnisse

Beispiel:
Ergebnis-
Tabelle
(Top 5,
2002)

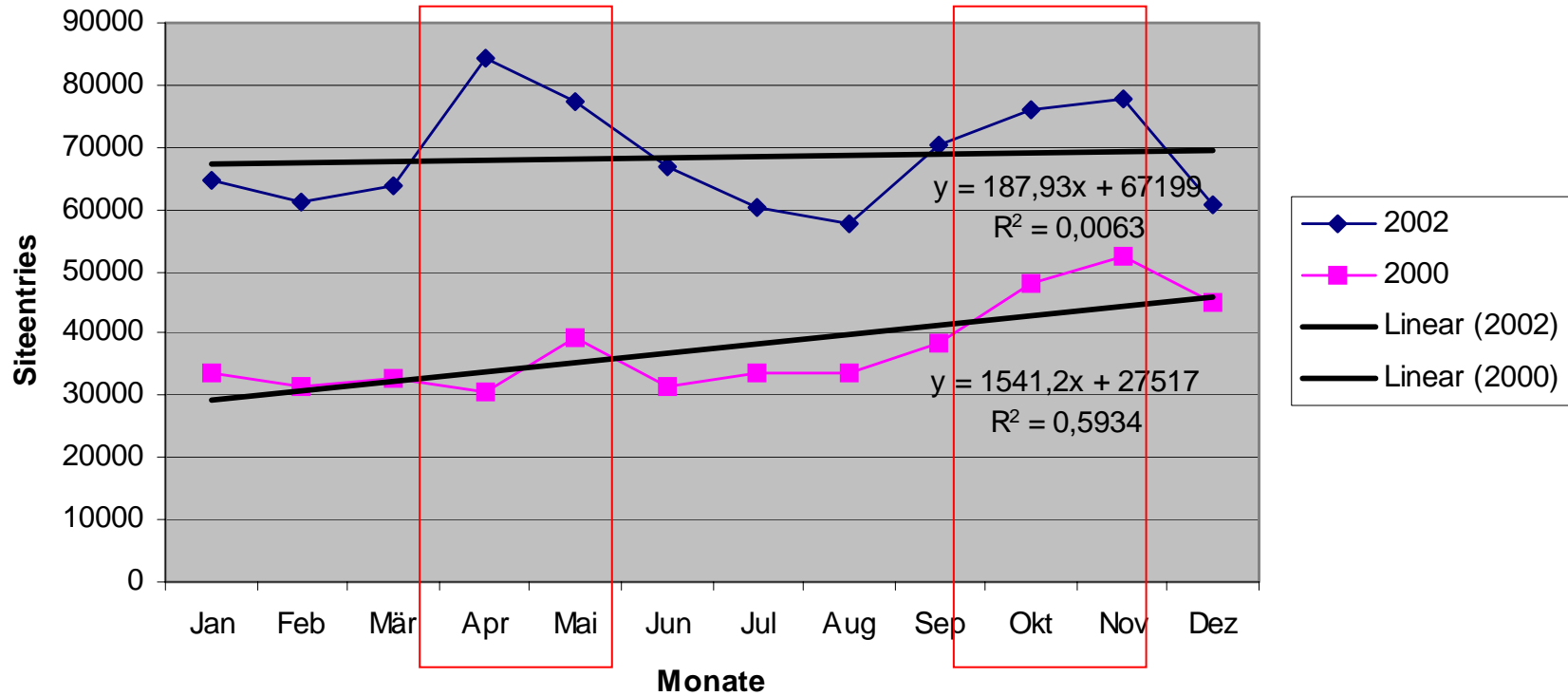
Rang	URL	Beschreibung	Kategorie	Größe	PR	Entry· S	Entry· D	Entry· R	WEF· S	WEF· D	WEF· R	Entry· total
1	/	Homepage des Instituts für Bibliothekswissenschaft (IB Homepage)	Home	av	6	6345	54369	6558	0,09	0,81	0,10	67272
2	/~mh/gedv/ascii. htm	Referenz der ASCII-Code Kodierung	Docu	av	4	19248	2399	187	0,88	0,11	0,01	21834
3	/~mh/projekte/m etaopac/	Startseite des „Meta-Opac Berlin- Brandenburg“	DB- Entry	av	5	2952	2490	8677	0,21	0,18	0,61	14119
4	/~is/computerku rs/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/script s/bwl/bwl.html	Vorlesungsscript zum Thema „Betriebswirtsch aftslehre“	Text	lg	4	7530	1656	719	0,76	0,17	0,07	9905



Magisterarbeit - Ergebnisse

- Monatsverlauf der Entries 2000 und 2002

12-Monatsverlauf Siteentries



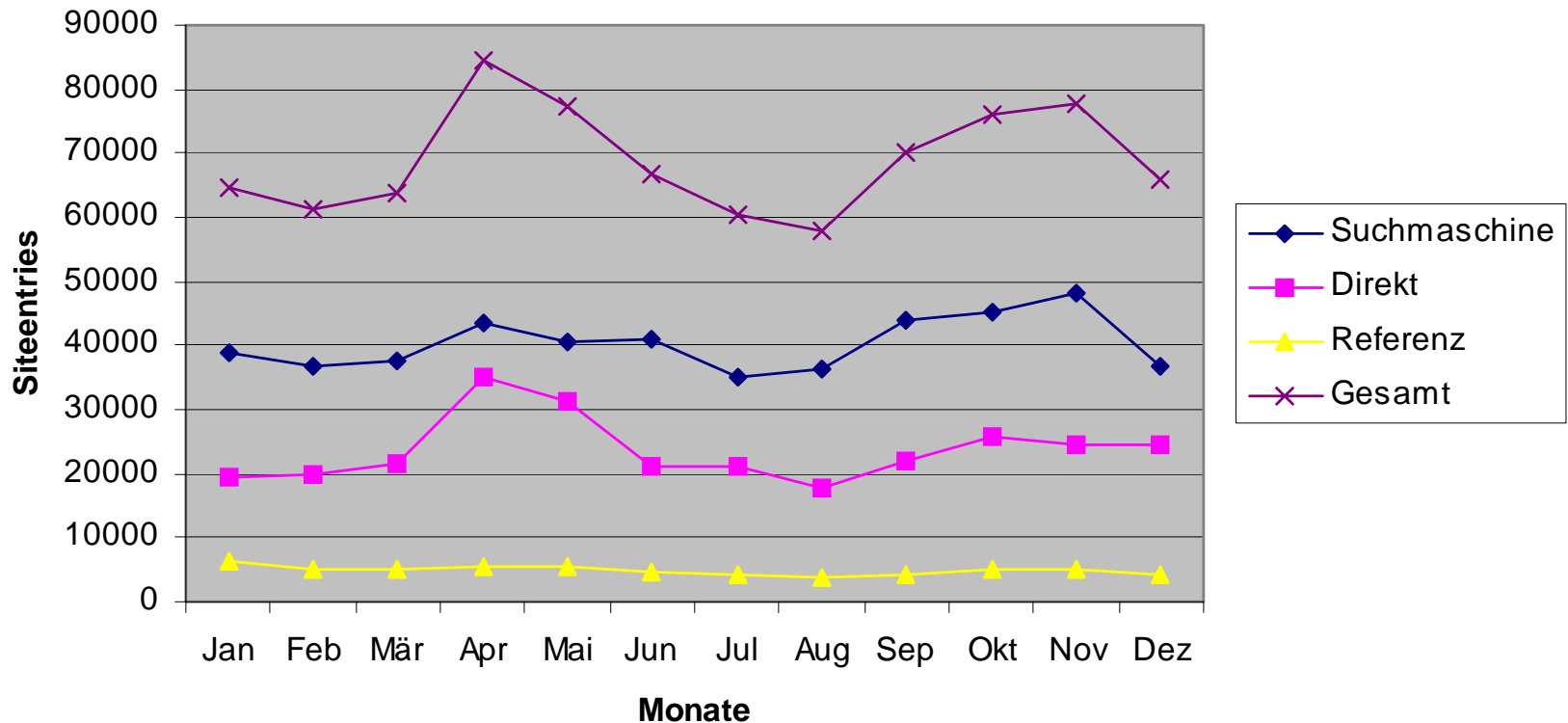
- ◆ 2002
- 2000
- Linear (2002)
- Linear (2000)

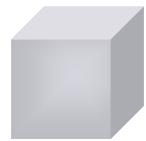


Magisterarbeit - Ergebnisse

- Verlauf der Entries über ein Jahr (2002), unterschieden nach Navigationsart
- Starke Schwankungen (siehe Semesterbeginn SS und WS)

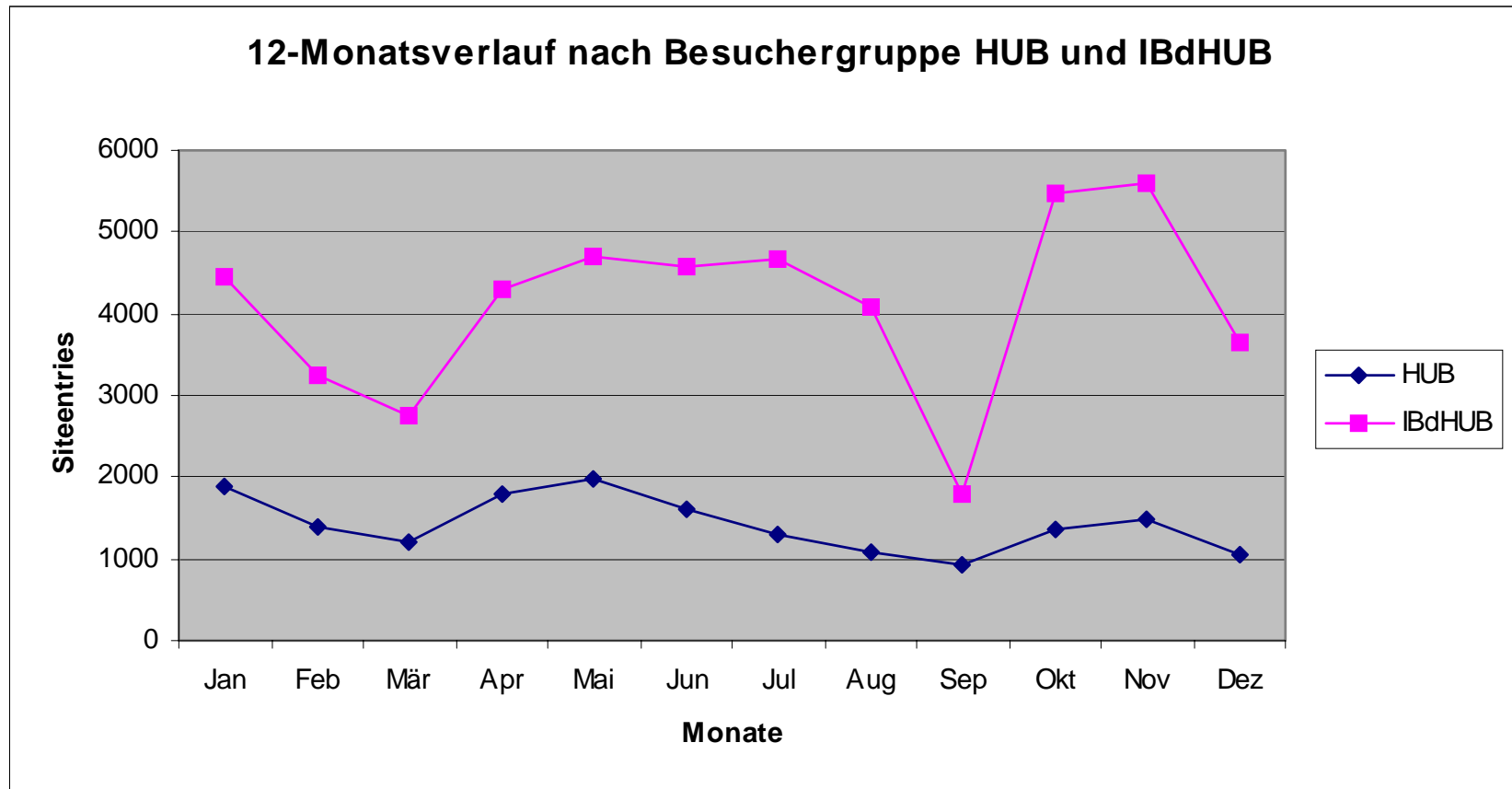
12-Monatsverlauf Siteentries nach Navigationsarten

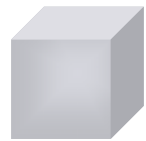




Magisterarbeit - Ergebnisse

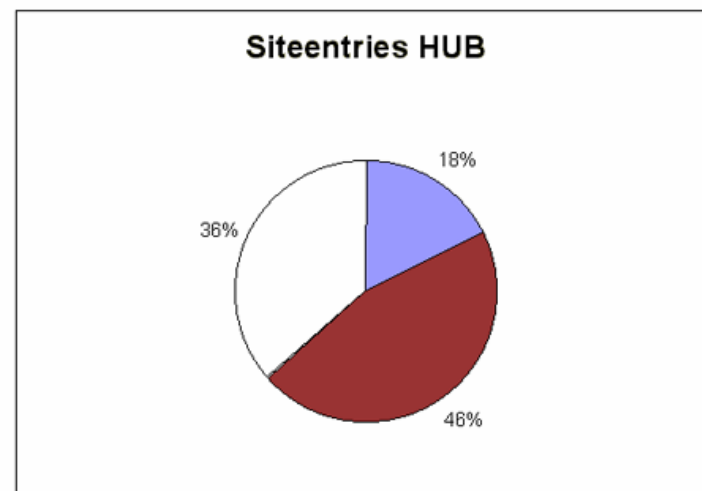
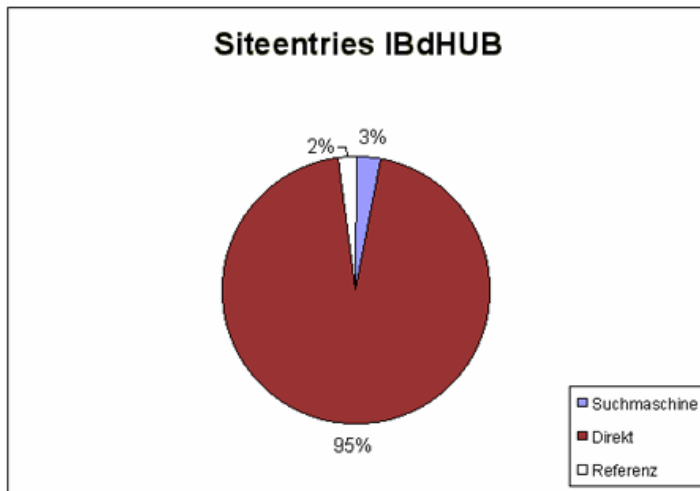
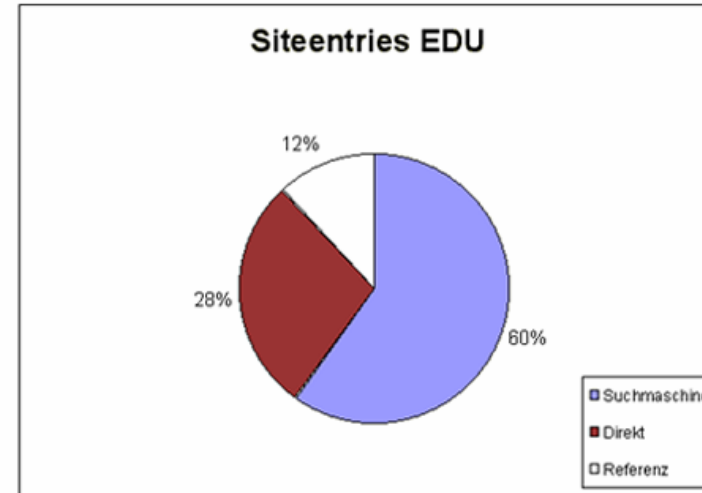
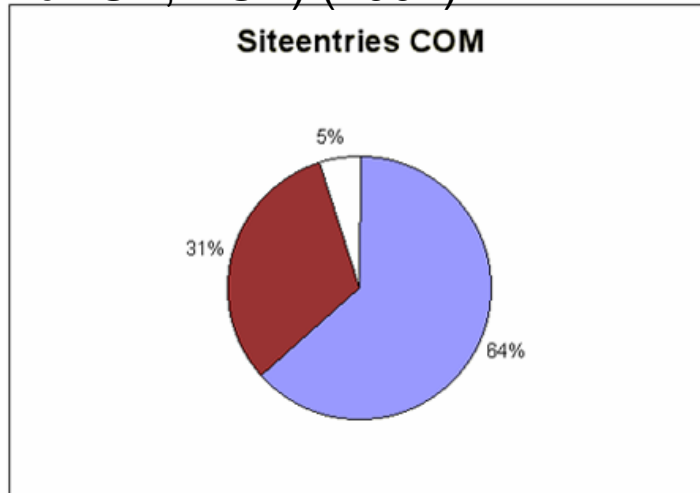
- HUB und IBdHUB Nutzer tendenziell ähnlich (2002)
- IBdHUB schwanken stärker





Magisterarbeit - Ergebnisse

- Verteilung der Einstiegsarten nach Besucher-Typ (COM, EDU, IBdHUB, HUB) (2002)

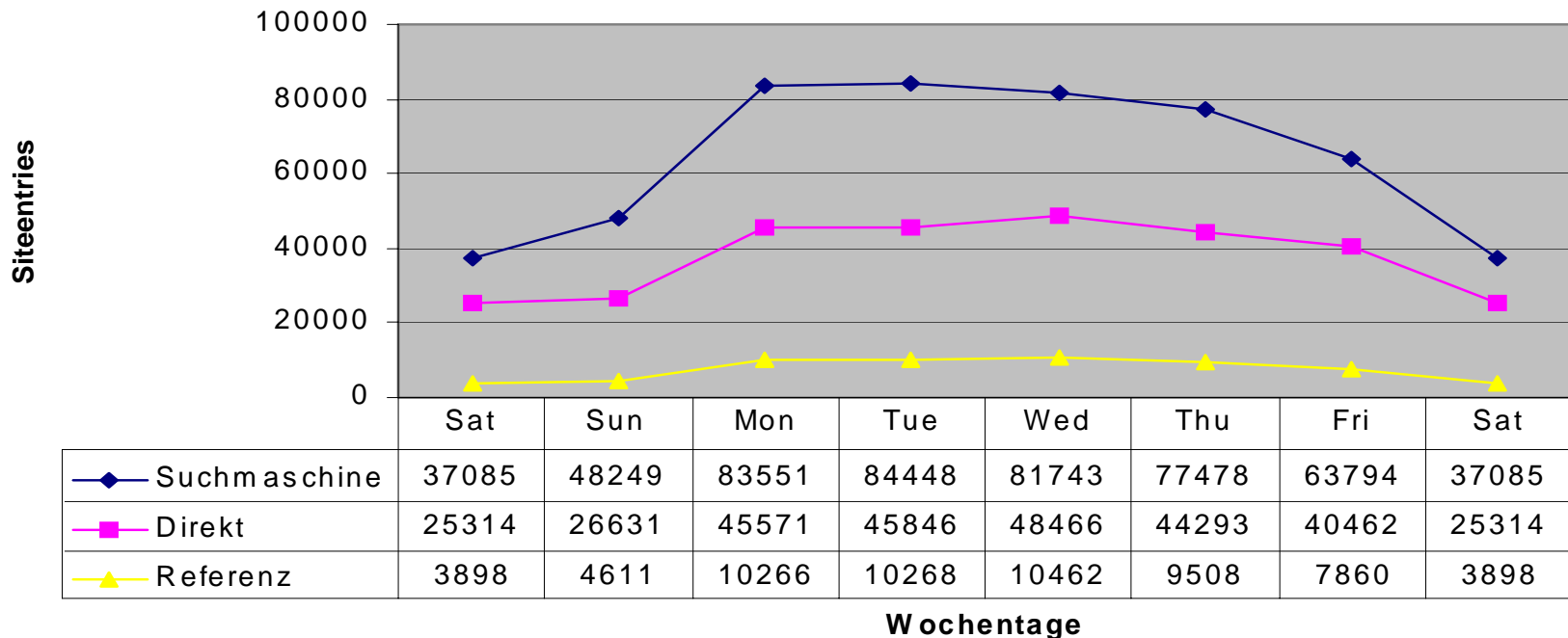




Magisterarbeit - Ergebnisse

- Suchmaschinen sind die wichtigsten Traffic-Lieferanten
- Entries unterschieden nach Wochentag

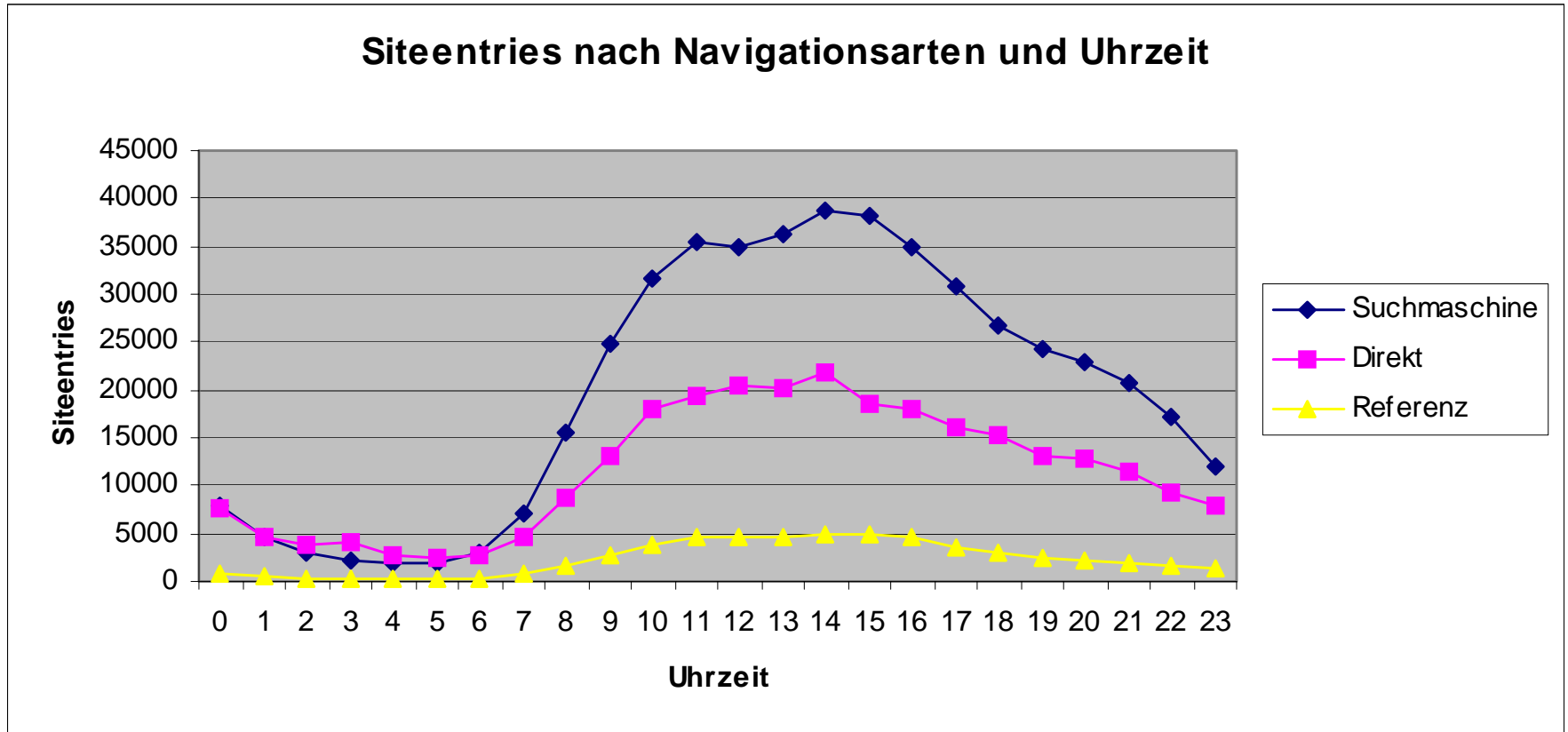
Siteentries nach Navigationsarten und Wochentag





Magisterarbeit – Ergebnisse

- Suchmaschinen sind die wichtigsten Traffic-Lieferanten
- Entries unterschieden nach Uhrzeit

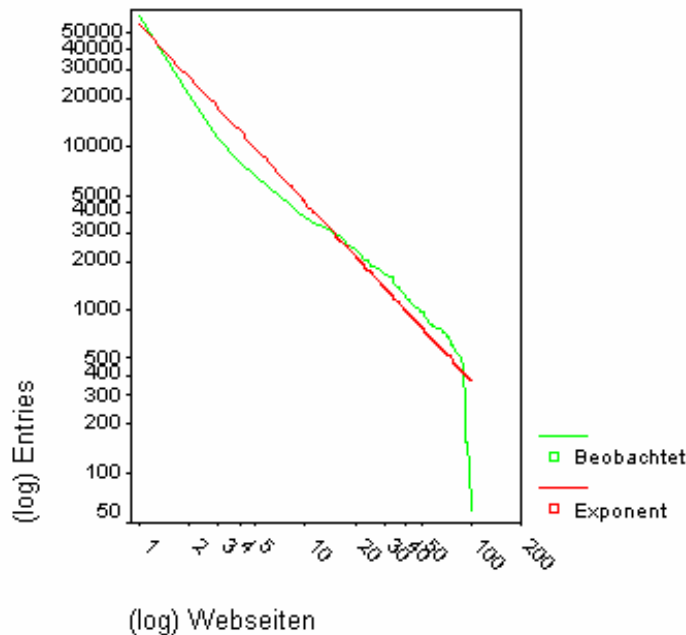




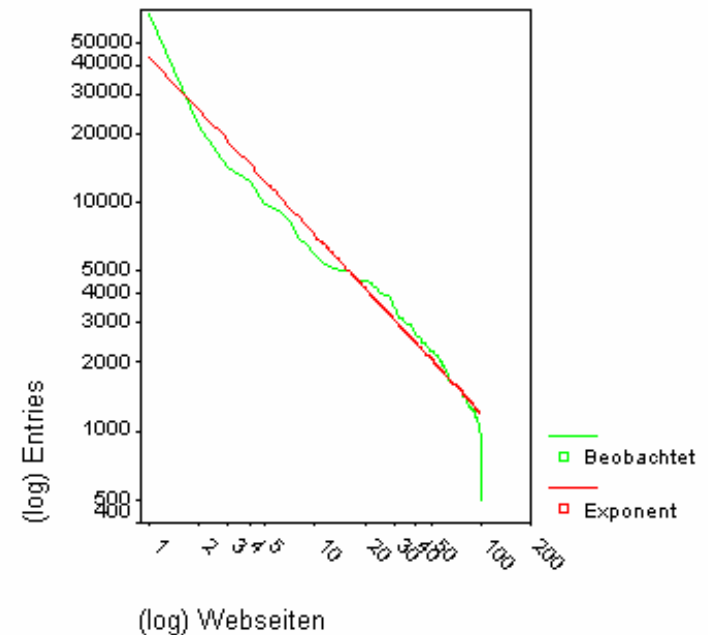
Magisterarbeit - Ergebnisse

- Power Law Verteilung (Top 100)
- Wenige Seiten erhalten sehr viele Entries, sehr viele Seiten erhalten wenige Entries

Entries pro Webseite (2000)



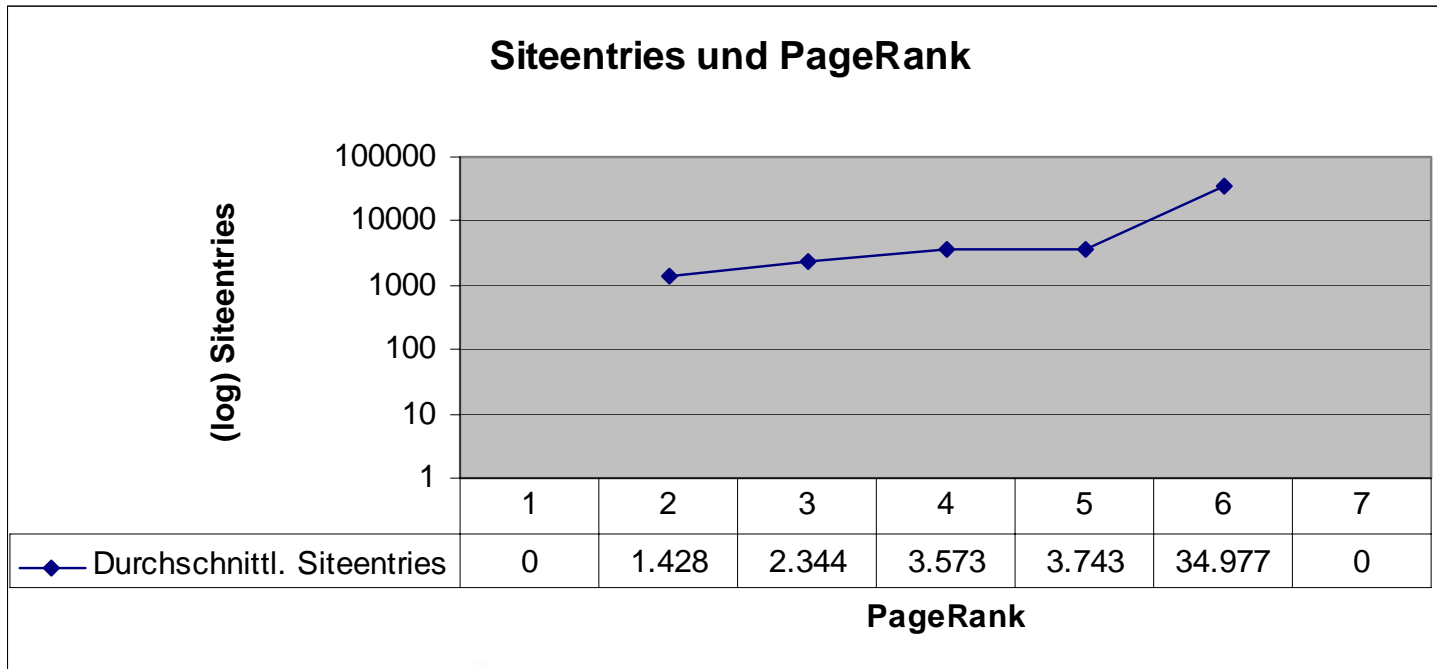
Entries pro Webseite (2002)





Magisterarbeit - Ergebnisse

Positive Korrelation zwischen d. durchschnittlichen Siteentries pro Seite und deren PageRank-Werten



Positive Korrelation zwischen durchs. WEF und durchs. PageRank

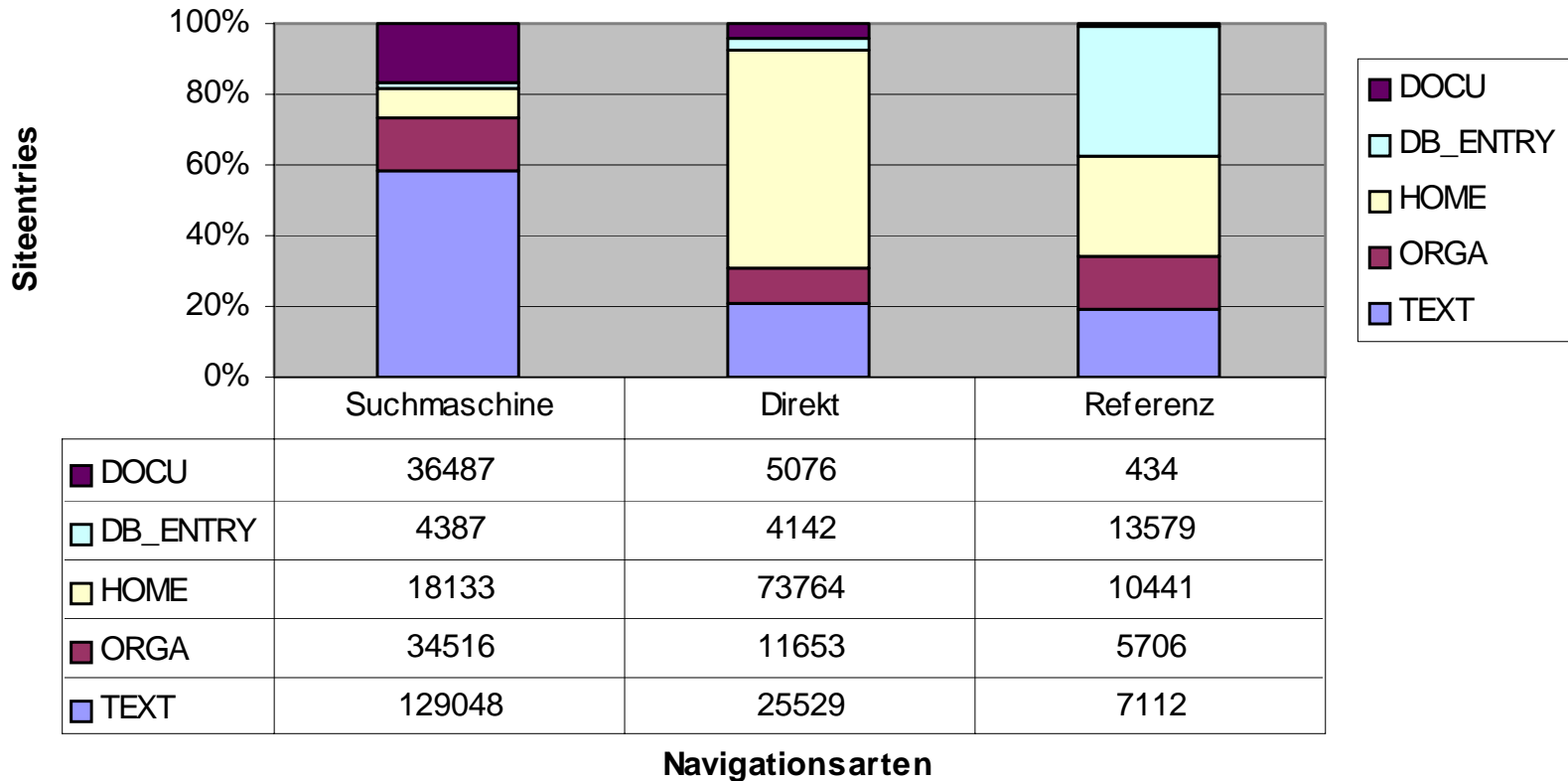
Inhaltsklasse	PageRank	WEF_R	WEF_D	WEF_S
DB_ENTRY	5,25	0,57	0,22	0,21
HOME	4,90	0,14	0,55 · (0,52)	0,32
ORGA	3,90	0,10	0,26	0,65
TEXT	3,71	0,04	0,16	0,80
DOCU	3,17	0,02	0,12	0,86



Magisterarbeit - Ergebnisse

- Verteilung der Siteentries auf die 5 Inhaltelassen nach Haas & Grams, 2000

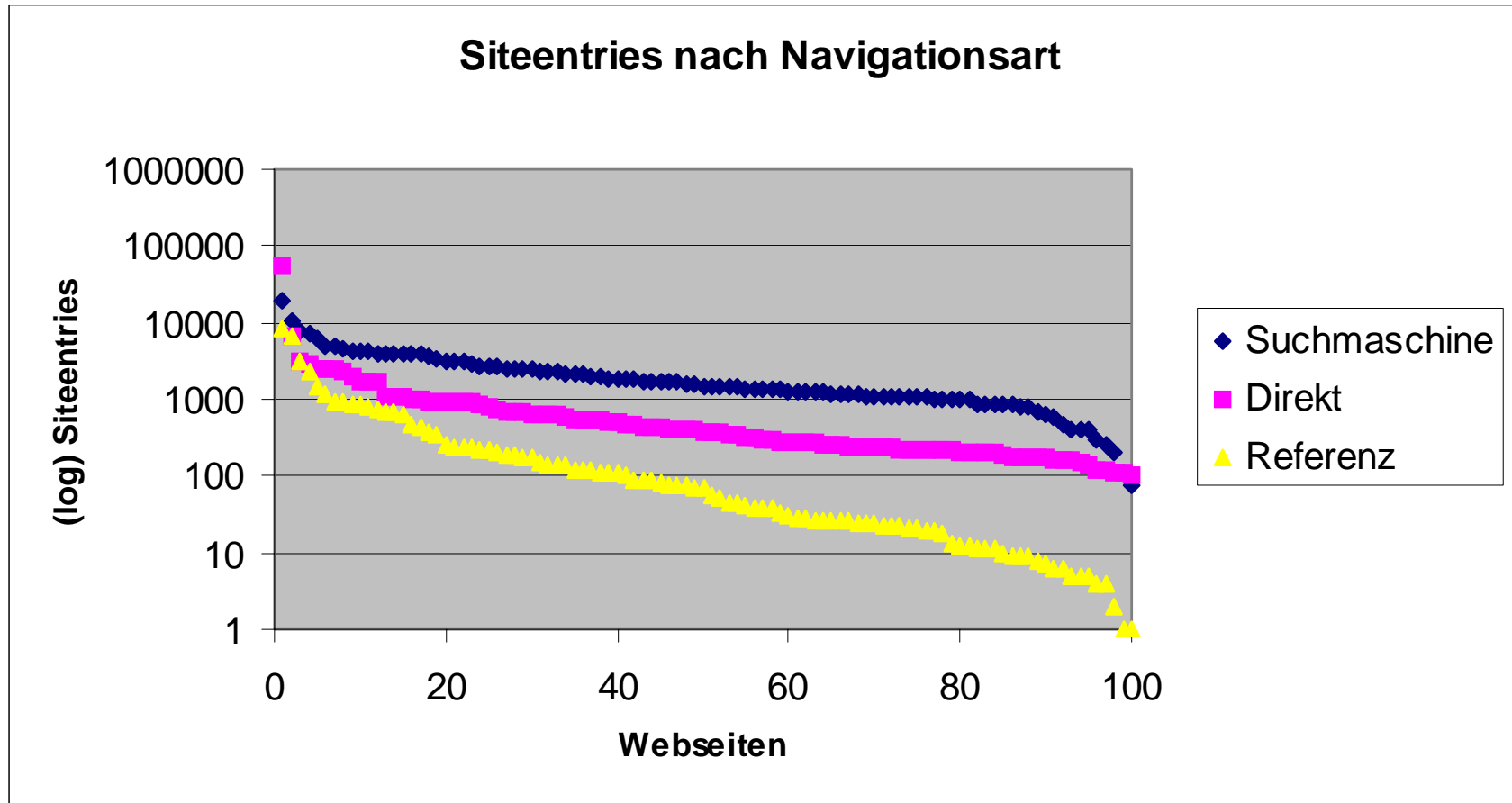
Siteentries der Inhaltelassen in % nach Navigationsart





Magisterarbeit - Ergebnisse

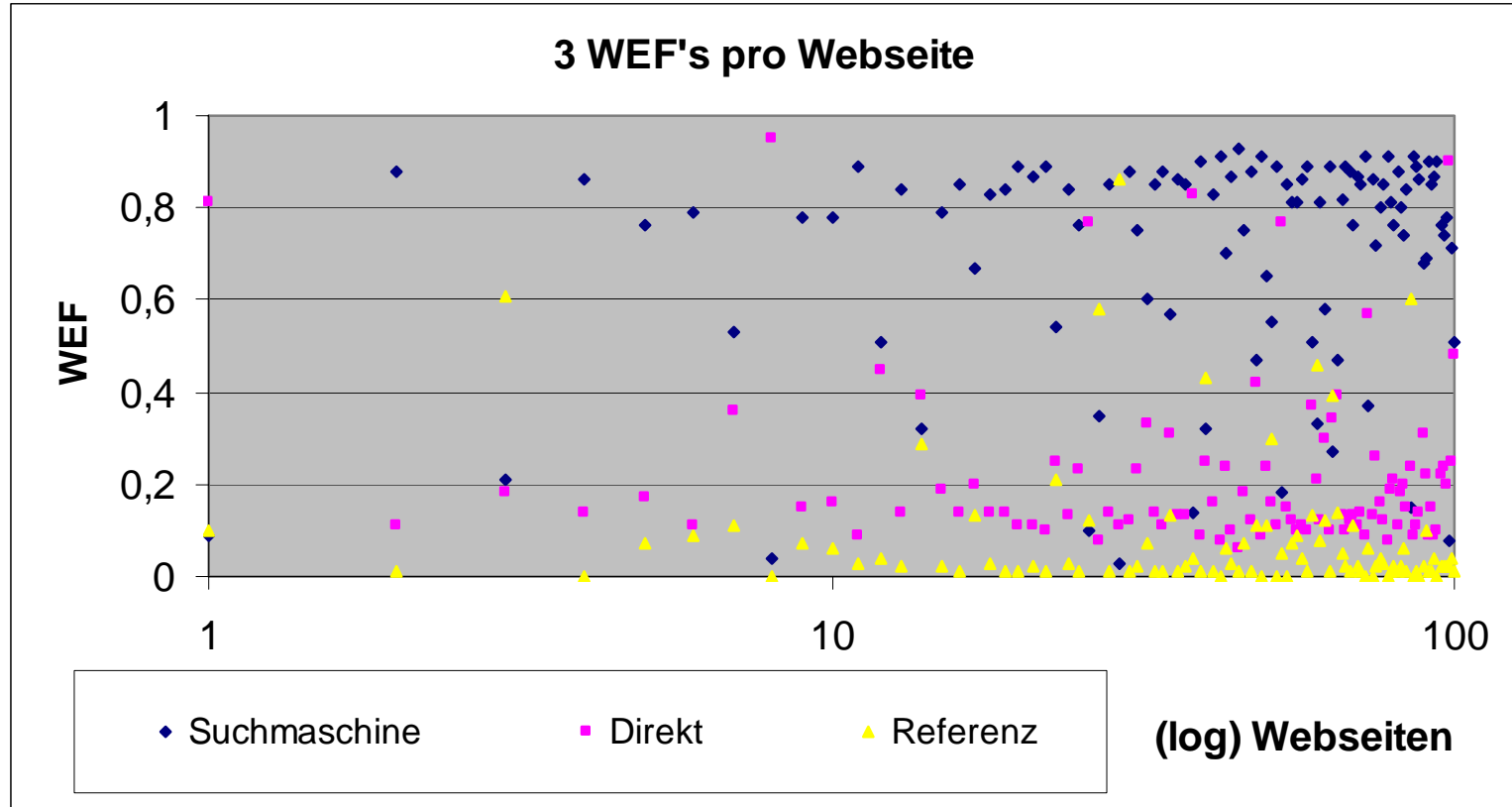
- Verteilung der Siteentries für die Top 100 Webseiten in Rangdarstellung





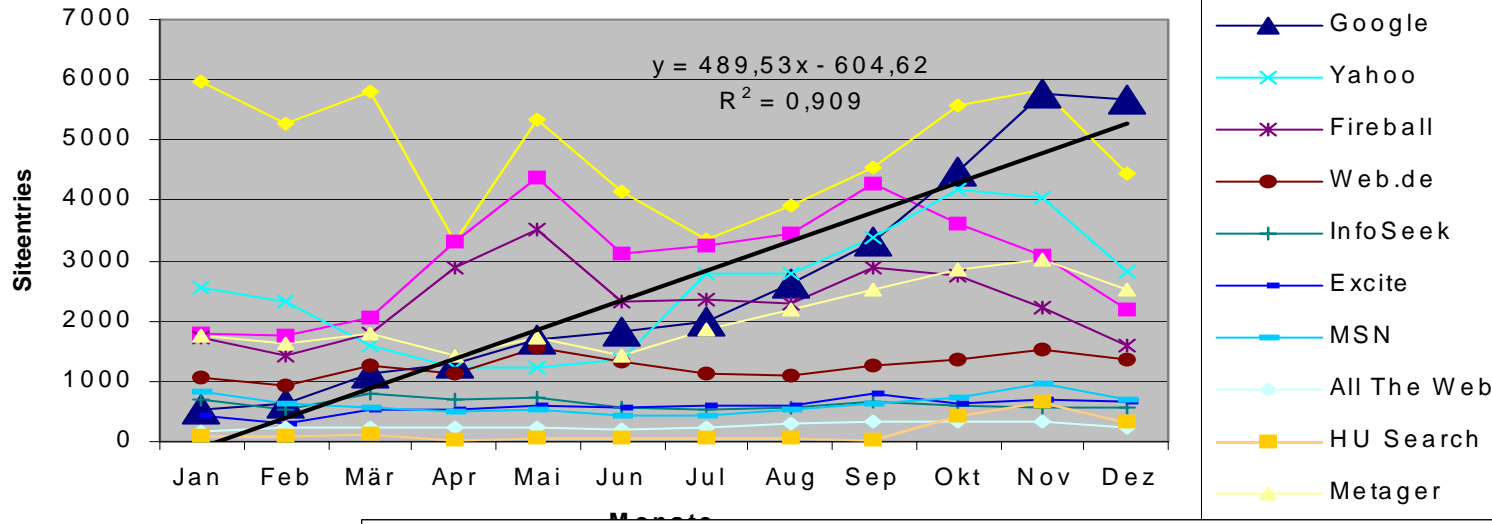
Magisterarbeit - Ergebnisse

- Verteilung der 3 WEF's für die Top 100 Webseiten
- Visualisierung der Zugänglichkeit
- Ausreißer bzgl. der WEF-Werte

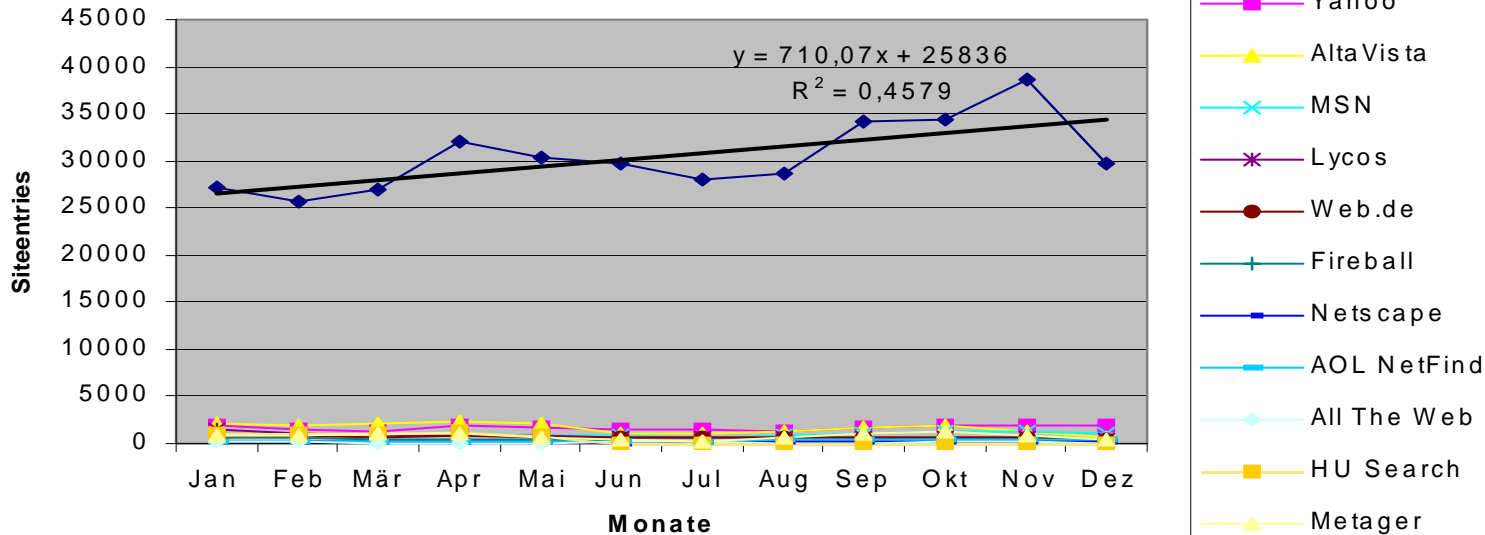




Top Suchmaschinen 2000



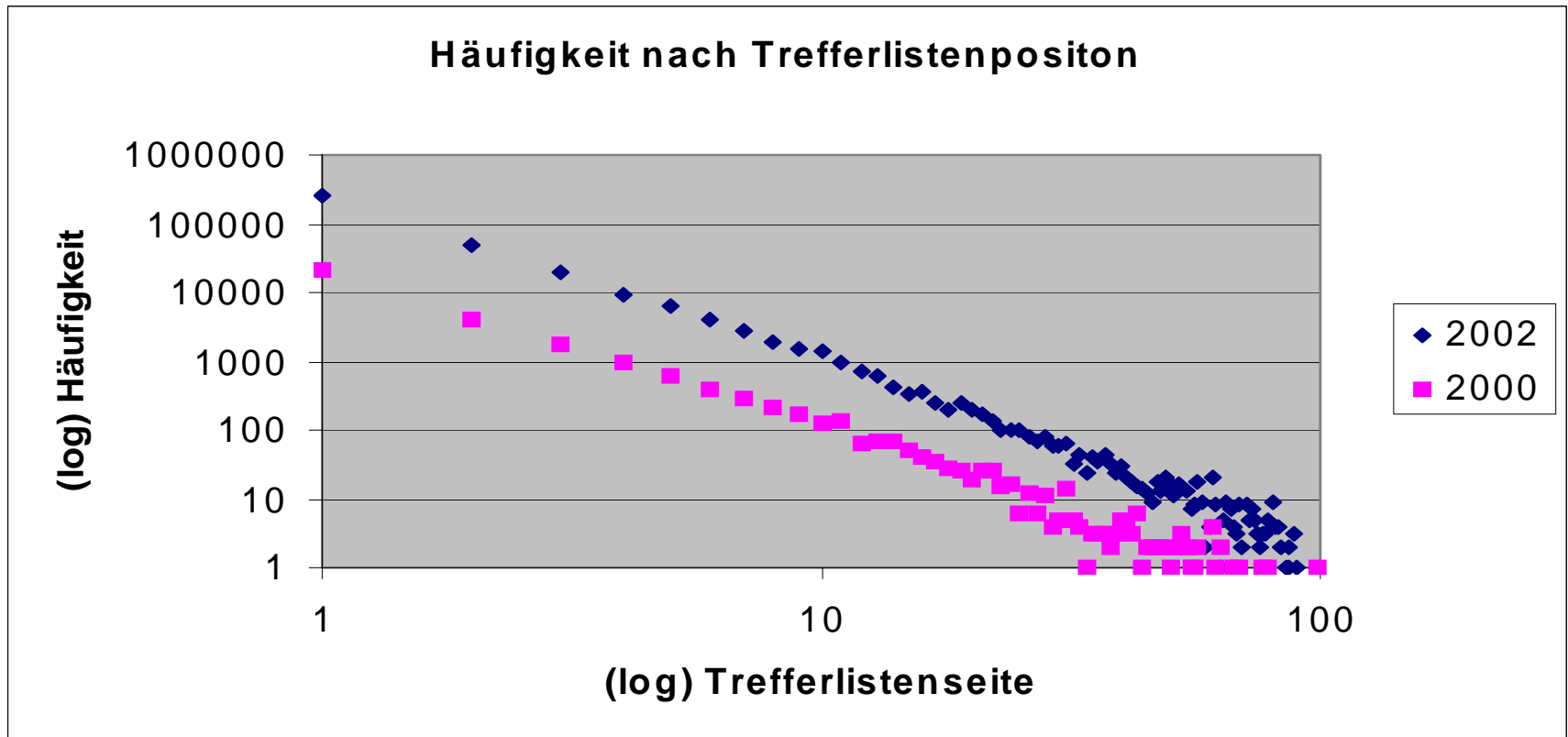
Top Suchmaschinen 2002





Magisterarbeit - Ergebnisse cont.

- Ergebnisse Queryanalyse: Eintritte nach Trefferlistenseite (doppelt logarithmierte Skala)
- „Die ersten beiden Trefferlistenseiten sind entscheidend!“





Magisterarbeit - Ergebnisse cont.

- Top Queries (April 2000, 2002)

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)



Magisterarbeit - Ergebnisse cont.

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

- Top Backlinks (2002)



Zusammenfassung

- Suchmaschinen sind die wichtigsten Trafficlieferanten
- Gründe:
 - Inhaltliche Zusammensetzung (große Textseiten) und Struktur der Site
 - Alter und Verlinkung der Site (link popularity)
- Text-Seiten sind die wichtigsten Einstiegsseiten
 - Suchmaschinen finden Text
 - Homepages werden meist direkt oder durch Backlink gestartet
- Ausreißer befinden sich in allen Seitenklassen
- WEF als nützliches Konzept zur Visualisierung und Messung der Zugänglichkeit, Sichtbarkeit und Verlinkung
 - Liefern neue Website-Informationen
- Analyse der Queries und Backlinks als Schlüssel zur Nutzung der Website



Ausblick

- Analyse von Web Logfile ist keinesfalls ausgereizt
 - Z.B. Neuere Untersuchungen von Nicholas et al.
- Im Gegenteil quantitative Nutzungsmessung könnte an Bedeutung und Anwendungsvielfalt gewinnen
- Analyse weiterer Websites wäre interessant
 - Regionaler Vergleich
 - Inhaltlicher Vergleich
 - Vergleiche in unterschiedlichen Sprachräumen
- Tiefere Analyse der Queries und Backlinks fruchtbar



Fazit

- Die „richtigen“ Daten erheben
- Eigene Anwendungen/Reports sind Standardtools meist überlegen
- Kombinierte Untersuchungen heben den Wert der Aussagen
 - z.B. Backlink-Zählung über Suchmaschine + Logfile
- Methoden Mix – quantitative und qualitative Analysen (z.B. Logfile, Fragebogen, Voting, andere Datensammlungen, ...)



Live-Demo

- Web Entry Miner (WEM) hat das Konzept der Magisterarbeit implementiert (Unterscheidung nach Navigationsarten)

Web Entry Miner | by elogic.ch

File

Top 100 Pages | Top 100 Directories | Other rankings

URL	SE-Entries	D-Entries	R-Entries	Total
/	287	5925	630	6842
/~mh/gedw/ascii.htm	1561	150	95	1806
/~fern/	229	488	162	879
/start.html	0	832	2	834
/~mh/projekte/metaopac/index.html	4	275	506	785
/inf/i_suche.htm	17	615	35	667
/~rfunk/lw/scripts/bwl/bwl.html	474	67	88	629
/amerika.html	514	36	25	575
/~wumsta/rehm1.html	362	64	85	511
/inf/bbbform.html	3	46	451	500
/~wumsta/rehm8.html	317	50	55	422
/~wumsta/rehm9.html	304	44	58	406
/~wumsta/rehm11.html	273	47	66	386
/~mh/gedw/romzs.htm	200	46	124	370
/~wumsta/rehm.html	218	39	109	366
/~pbruhn/russgus.htm	110	100	154	364
/~wumsta/rehm71.html	268	30	60	358
/~wumsta/rehm6.html	259	45	39	343
/~wumsta/rehm4.html	254	44	44	342
/~kumlau/handreichungen/h64/	170	57	99	326
/~mh/css/css2/fonts.html	272	20	23	315
/~hab/arnd/Start.html	64	35	214	313
/~kumlau/handreichungen/h58/	162	89	61	312
/inf/studium.htm	5	55	223	283
/~hab/arnd/	105	16	158	279

Entries' Type | Detailed report | Resume

86.0% Search engines

8.0% Direct access

5.0% External

/~mh/gedw/ascii.htm



Mehr Informationen

- Magisterarbeit –
<http://www.ib.hu-berlin.de/~mayr/magisterarbeit/>
- WEM – Web Entry Miner
<http://www.ib.hu-berlin.de/~mayr/wem/>

Der WEM steht für akademische Logfile-Analyse zur Verfügung!



Tools

Open Source

- Analog <http://www.analog.cx/>
- Webalizer <http://www.webalizer.org/>
- LogReport <http://logreport.org/>

Kommerzielle Programme

- WebTrends <http://www.netiq.com/webtrends>
- NetTracker <http://www.sane.com/>
- Funnel Web http://www.quest.com/funnel_web/analyzer/
- LogFileAnalyse Pro <http://www.lfa-pro.de/>

[vgl. Wikipedia]



Literatur

- Analysen von Webserver-Logfiles zur Kategorisierung des Navigationsverhaltens von Nutzern / von Heike Oldenburg, Magisterarbeit am Institut für Bibliothekswissenschaft, 2003.
- Cracking the Code: Web Log Analysis / von David Nicholas et al., in: Online & CD-ROM Review, 1999, Vol. 23, No. 5
- Developing and testing methods to determine the use of websites: case study newspapers / von David Nicholas et al., in: Aslib Proceedings, 1999, Vol. 51, No. 5
- Methodische Anmerkungen zur Auswertung der WWW-Log-Dateien des Servers www.gesis.org / von Wolf-Dieter Mell, 2002, IZ-Arbeitsbericht Nr. 26
- Micro-mining and segmented log file analysis: a method for enriching the data yield from Internet log files / von David Nicholas & P. Huntington, in: Journal of Information Science, 29 (5) 2003, pp. 391-404
- Web log file analysis: backlinks and queries / von Mike Thelwall, in: Aslib Proceedings, 2001, Vol. 53, No. 6



Abschluss

Charakteristische Zitate:

“Unfortunately the logs turn out to be good on volume and (certain) detail but bad at precision and attribution.” ...

“The research, in fact, turned out to be the type of research where the journey itself proved to be more important than the destination ...“

„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.”

[Nicholas et al., 1999]

Abschluss



Vielen Dank für Ihre Aufmerksamkeit!

Kontakt



Philipp Mayr M.A.

email: mayr@informatik.hu-berlin.de,

philippmayr@web.de

www: <http://www.ib.hu-berlin.de/~mayr/>