

Einführung in die Logfile Analyse

Referent: Philipp Mayr

7. InternetSalon am 31. März 2004, pr-ide

© 2004, HiSolutions AG



Agenda

Dauer ca. 1,5 h

- Begrüßung
- Einführung, Grundlagen, Probleme, Potenziale
- Zehn typische Logfile Auswertungen
- Erweiterte Logfile Analyse (Beispiele)
- Zusammenfassung, Ausblick



Einführung

Was ist ein Logfile?

- DEF.: *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]*
- Strukturiert aufgebaute, serverplattformabhängig Protokolldatei (ASCII)

Grundlagen

Fakten

- 50% des Verkäufe im Web sind verloren, weil die User die Inhalte nicht finden [Gartner Group]
- 40% der wiederkehrenden Besucher gehen aufgrund negativer Erfahrungen verloren [Zona Research]
- 85% der Besucher brechen den Besuch auf einer schlecht gestalteten Site ab [cPulse]



Grundlagen

Loganalyse Ablauf:

1. Tracking des Traffics zur Etablierung v. Benchmarks (Bezugsgrößen)
2. Identifikation von Verbesserungsmöglichkeiten und Änderung der Site
3. Messung der Auswirkungen bzw. Effektivität der Änderung
4. Wiederholung dieses Prozesses

Grundlagen

Allgemeine Fragen die Logfiles beantworten können

- Wie lauten IP-Adresse und Hostname des Nutzers?
- Welchen Browser hat der Besucher benutzt?
- Von welchen Seiten kommen die Besucher?
- Welche Suchmaschinen bzw. Suchwörter werden genutzt?
- Wie sichtbar ist die Website?
- Ist die Site bzw. die Seite gut bei Suchmaschinen gerankt?



Grundlagen

Fragen cond.

- Wie lange bleiben die Besucher auf einzelnen Webseiten?
- Wie viele Seiten rufen sie dabei auf?
- Auf welcher Seite verlassen die Besucher die Website?
- Welches Betriebssystem nutzen die User?
- Welche Internetseiten hat Mitarbeiter XY während der Arbeitszeit besucht?



Beispiel Logdaten

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&sa=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



Beispiel Logdaten cond.

Beispiel Logzeile - WER?

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&sa=N"

"Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0)"

- **Feld: Host**
- Beschreibung: die anfordernde Host-Adresse (IP-Nummer), der Computer
- Siehe oben: **120.0.0.7**
- Beispiele: numerische IP-Adressen (z.B. 123.22.33.444), aufgelöste Adressen (z.B. firewall-firma-xyz)



Beispiel Logdaten cond.

Beispiel Logzeile - WER genau?

```
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&sa=N"  
"Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0)"
```

- **Feld: ident, authuser**
- Beschreibung: Benutzeridentifizierung bzw. authentifizierung
- Siehe oben: - -
- Beispiele: pmayr, userxyz



Beispiel Logdaten cond.

Beispiel Logzeile - WANN?

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&sa=N"

"Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0)"

- **Feld: date**
- Beschreibung: genaue Zeitangabe des Zugriffs für jeden Hit, folgt dem Format [Tag/Monat/Jahr:Stunde:Minute:Sekunde Zone]
- Siehe oben: **[06/Jan/2002:11:14:34 +0100]**



Beispiel Logdaten cond.

Beispiel Logzeile - was?

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&sa=N"

"Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0)"

- **Feld: request**
- Beschreibung: Welche Datei wurde angefordert? HTTP-Methode (GET, HEAD, POST), Ressource und Protokoll
- Siehe oben: "**GET /~fern/ HTTP/1.1**"
- Beispiele: GET /robots.txt, GET /pics/pic.gif, GET /finanzen/index.htm



Beispiel Logdaten cond.

Beispiel Logzeile – WIE ist die Transaktion verlaufen?

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&sa=N"

"Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0)"

- **Feld: status**
- Beschreibung: Statusnummer der Transaktion, 200er = erfolgreich, 300er = weitergeleitet, 400er = fehlgeschlagen, 500er = Serverfehler
- Siehe oben: **200**
- Beispiele: 404, 304, 500



Beispiel Logdaten cond.

Beispiel Logzeile - WIEVIEL Daten?

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&sa=N"

"Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0)"

- **Feld: byte**
- Beschreibung: Menge der gesendeten Daten in Byte
- Siehe oben: **12872**



Beispiel Logdaten cond.

Beispiel Logzeile - WOHER?

```
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&sa=N"
```

```
"Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0)"
```

- **Feld: referer**
- Beschreibung: Von welcher URL wurde die Seite angefordert? Wird nur im „extended log format“ protokolliert!
- Siehe oben: **"http://www.google.de/search?q=fernstudium&start=20&sa=N"**
- Beispiele: Suchmaschine, externer Link, direkt (Bookmark, History, Eingabe), interner Link



Beispiel Logdaten cond.

Beispiel Logzeile - womit?

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&sa=N"

"Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0)"

- **Feld: useragent**
- Beschreibung: Informationen zum Browsertyp und Betriebssystem des zugreifenden Rechners
- Siehe oben: **"Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0)"**
- Beispiele: Googlebot/2.1, Slurp,

Nutzungsmaße

1. Einfache Nutzungsmaße (Hit, View, Impression, Visits)
2. Besuchercharakteristiken (Nationalität, Organisation, Vielkäufer, Subscriber)
3. Suchcharakteristiken (Themen, Trends, Suchbegriffe vs. Informationsbedürfnis)
4. Benutzungspfade (z.B. Muster)



Begriffe

- **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



Beispiele Hit

Beispiel 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, User Session

```

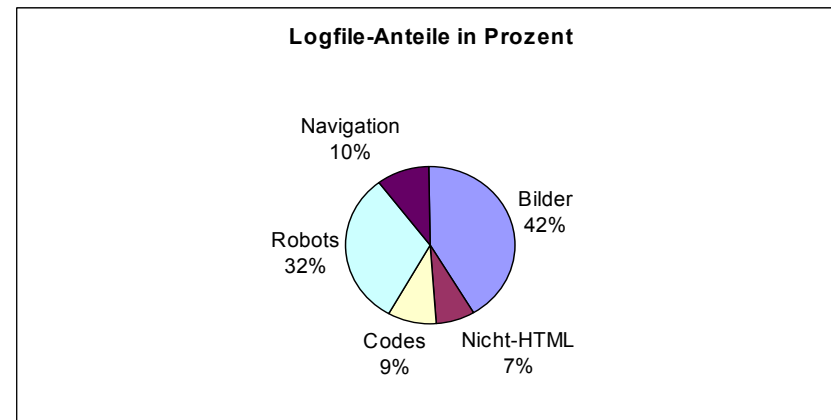
203.30.5.145 - - [01/Jun/1999:03:09:21 -0600] "GET /Calls/OWOM.html HTTP/1.0" 200
3942 "http://www.lycos.com/cgi-bin/pursuit?query=advertising+psychology-
&maxhits=20&cat=dir" "Mozilla/4.5 [en] (Win98; I)"
203.30.5.145 - - [01/Jun/1999:03:09:23 -0600] "GET /Calls/Images/earthani.gif
HTTP/1.0" 200 10689 "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

- Filtern und Säuberung der Daten (Hit ist nicht gleich Hit)
[10% der Daten verbleiben nach Data Cleaning zur Analyse]
- Firewall-Problematik und Caching – Welche Transaktionen wurden nicht aufgezeichnet? (under/over reporting)
- HTTP-Protokoll ist zustandslos – Registrierung von atomaren Transaktionen
- URL's häufig nichtssagend und müssen daher semantisch angereichert werden



Probleme cond.

- Virtual users ungleich end user
- Konzeption als technisches Protokoll, nicht zur Untersuchung von Onlinebehavior und -retrieval
- Datenschutz
- Größe der Logfiles (AOL.com ca. 400 Views/s)
- Zugang zu den Logfiles
- Anomalien



Potenziale

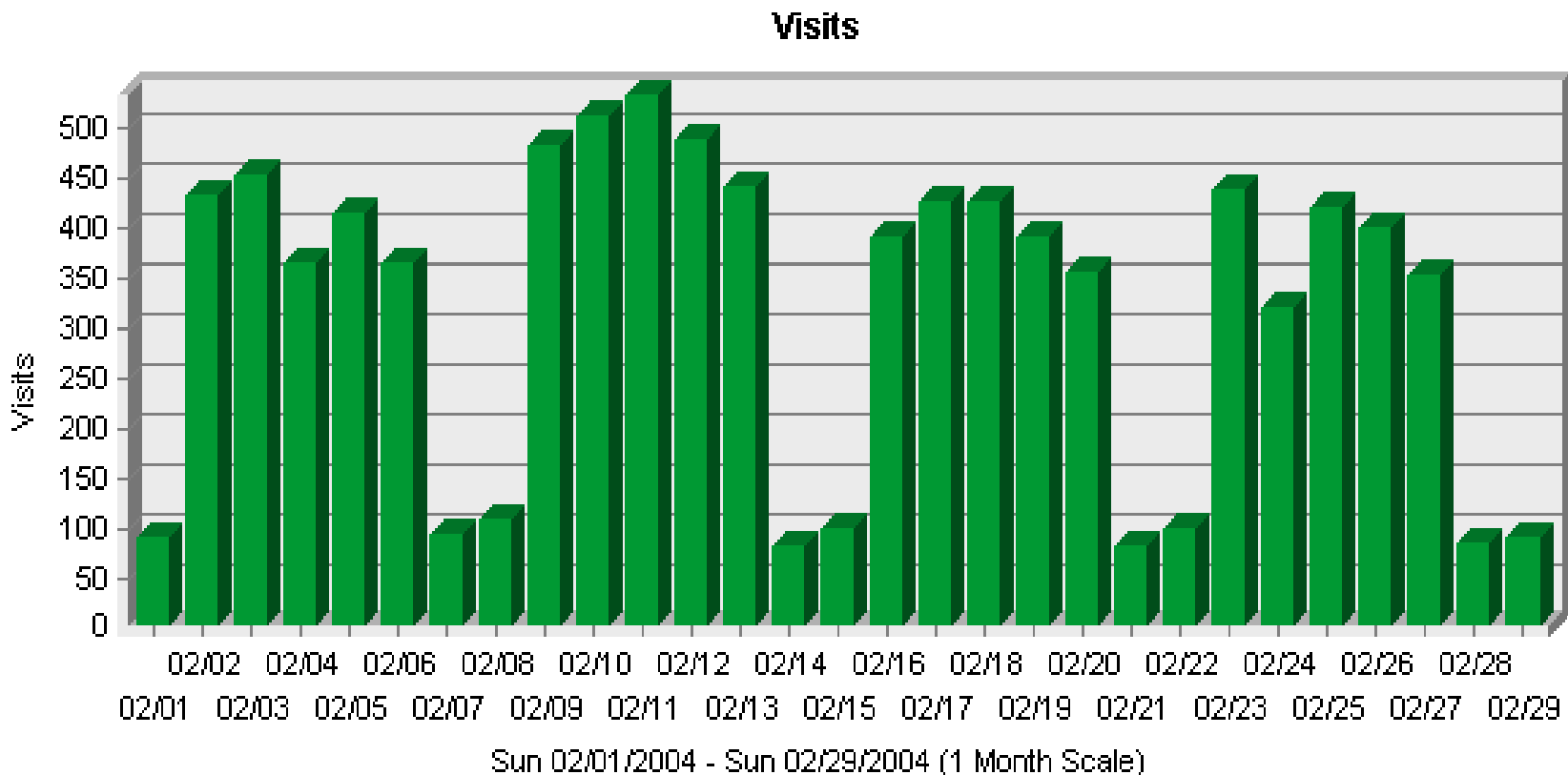
- große und heterogene Besucherschaft
- 24h und 7 Tage die Woche
- Sehr zeitnahe Analysen
- Untersuchung des Informationsverhalten, -
aufnahmeverhalten
- Hinweise zur Optimierung, Evaluation und Adaption
von Websites

Zehn typische Reports

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?
8. Wo sind Löcher in meiner Site? Single Access Pages
9. Kommen die Besucher zurück? Neue vs. alte Besucher
10. Gibt es technische Probleme?

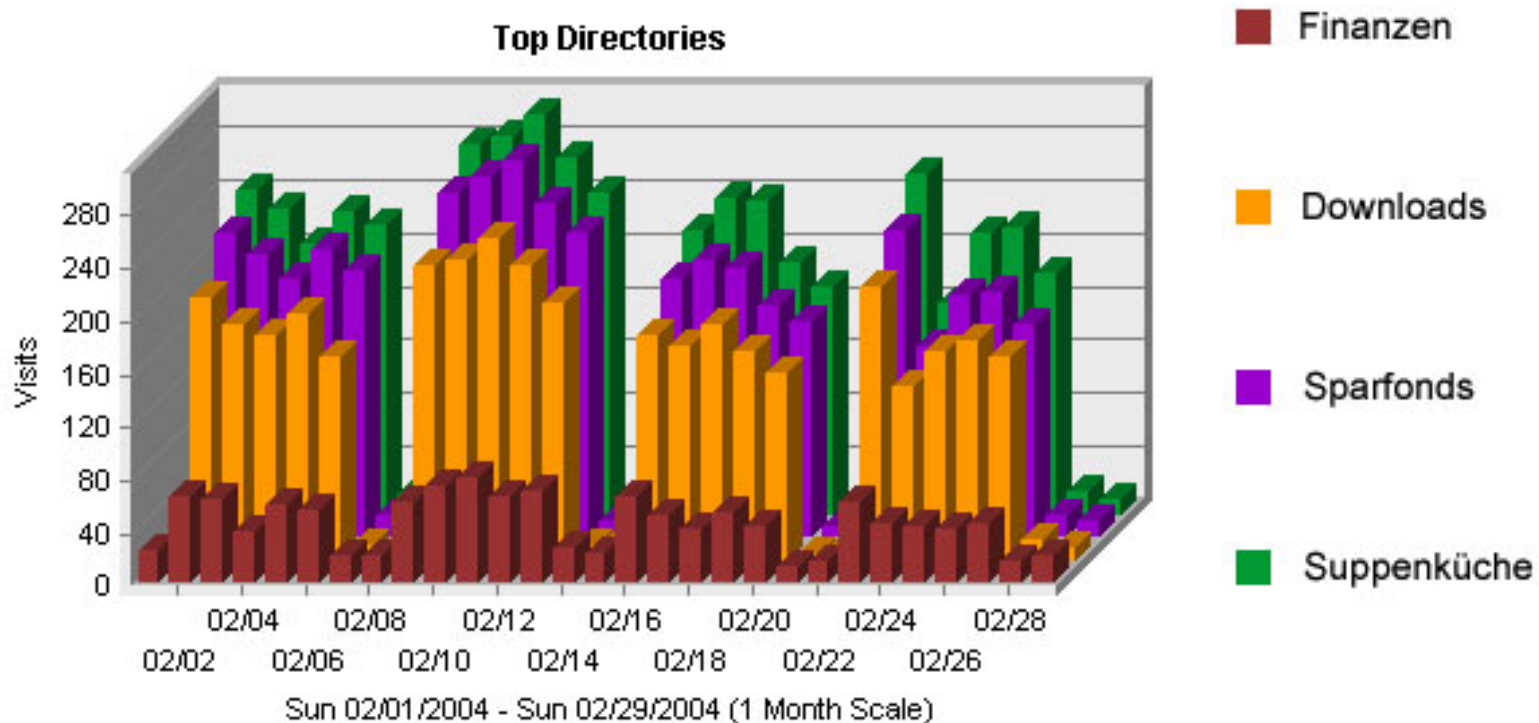
Report 01 - Traffic

- Wie viel Traffic erhält die Site, Directory, Page, ...?
- Kennzahlen (visits, page views, downloads) nach Tag, Stunde
- Benchmarks, Trends, Anomalien



Report 02 – Beliebteste Bereiche

- Die beliebtesten Bereiche der Site
- Welche Inhalte ziehen die Besucher an?
- Gruppierung von Seiten – Inhaltsgruppen
- Kanalisierung von Traffic



Report 03 – Einstiegsseiten

- Die wichtigsten Einstiegsseiten (Entry Pages) der Site
- Der erste Eindruck der Besucher ist meist entscheidend!
- Es landen nicht alle Besucher auf der Homepage!
- Optimierungshinweise, Informationsbedürfnisse, Adaption
- Zugangsarten: Suchmaschine, Bookmarks, Backlink (Referer)

Top Entry Pages			
	Page	% of Total	Visits
1	http://www.ib.hu-berlin.de/	5.35%	229
2	http://www.ib.hu-berlin.de/~hab/arnd/Start.html	3.74%	160
3	http://www.ib.hu-berlin.de/~mh/gedv/ascii.htm	2.26%	97
4	http://www.ib.hu-berlin.de/~mh/css/css2/fonts.html	1.96%	84
5	http://www.ib.hu-berlin.de/~mh/projekte/metaopac/	1.75%	75
6	http://www.ib.hu-berlin.de/jaw/Html/studwohn.html	1.26%	54
7	http://www.ib.hu-berlin.de/~hab/christine/gaudi1.html	1.16%	50
8	http://www.ib.hu-berlin.de/~pbruhn/russgus.htm	1.14%	49
9	http://www.ib.hu-berlin.de/~wumsta/rehm8.html	1.14%	49
10	http://www.ib.hu-berlin.de/~wumsta/rehm4.html	1%	43

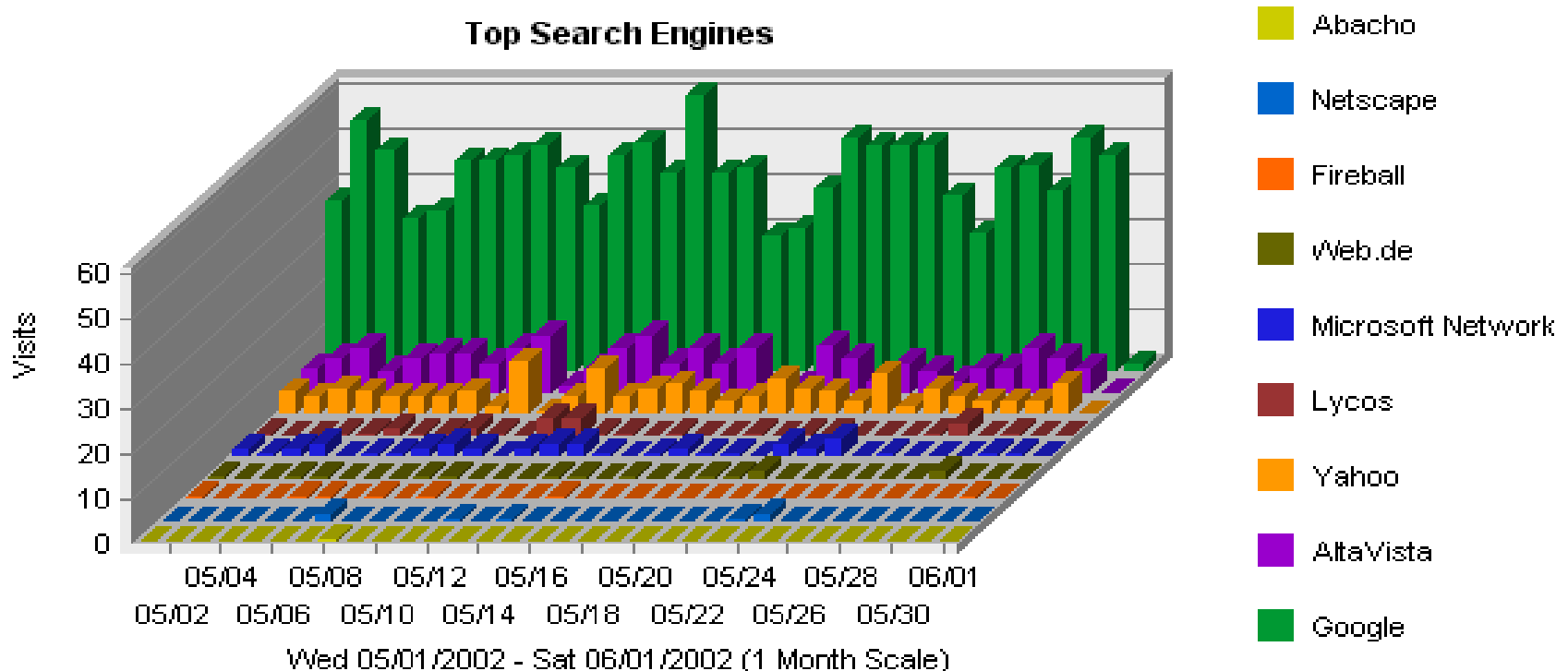
Report 04 - Sichtbarkeit

- Woher kommen die Besucher?
- Wer bringt den meisten Traffic? (Partnerseiten, Suchmaschinen, Directories)
- „No Referrer“ – ein Maß für Bekanntheit, direkte Navigation

Top Referring Sites		Top Referring Sites		Visits
1	http://www.ub.hu-berlin.de/	1	No Referrer	2,360
2	http://www.hu-berlin.de/	2	http://www.google.com/	1,251
3	http://www.physik.fu-berlin.de/	3	http://www.ib.hu-berlin.de/	992
4	[unknown origin]	4	http://www.google.de/	830
5	http://de.dir.yahoo.com/	5	http://www.altavista.com/	670
6	http://www2.hu-berlin.de/	6	http://www.google.ch/	397
7	http://www.sewanee.edu/	7	http://google.yahoo.com/	314
8	http://www.dbi-berlin.de/	8	http://hub.ib.hu-berlin.de/	308
9	http://www.fh-potsdam.de/	9	http://images.google.com/	301
10	http://www.hbz-nrw.de/	10	http://www.google.at/	272
11	http://www.niester.de/			269

Report 05 - Suchmaschinen

- Welche Suchmaschinen bringen Traffic? Über welche Begriffe wird die Site sichtbar?
- Erfolg von Suchmaschinen-Marketing (Suchbegriffe, Phrases)
- Die Position auf der SE-Trefferseite ist entscheidend!





Report 05 – Suchmaschinen cond.

Top Search Phrases with Engines Detail			
Phrases	Engines	Searches	% of Total
florenz	Google	625	1.52%
	Yahoo	9	0.02%
	AltaVista	6	0.01%
	Web.de	5	0.01%
ascii code	Google	451	1.09%
	Microsoft Network	54	0.13%
	Yahoo	24	0.05%
	AltaVista	18	0.04%
	Lycos	15	0.03%
	Web.de	9	0.02%
	Ask Jeeves	3	0%
	dogpile	3	0%
	Look Smart	1	0%
	Acoon	1	0%
	MegaSpider	1	0%
fernstudium	Web.de	172	0.41%
	Google	166	0.4%
	Lycos	38	0.09%
	AltaVista	20	0.04%
	Microsoft Network	8	0.01%
	Yahoo	7	0.01%
	Fireball	4	0%
	Abacho	1	0%



Report 05 – Suchmaschinen cond.

Top Search Phrases			
	Phrases	Phrases found	% of Total
1	florenz	645	1.57%
2	ascii code	580	1.41%
3	fernstudium	416	1.01%
4	römische zahlen	239	0.58%
5	opac berlin	200	0.48%
6	ascii-code	199	0.48%
7	hildebrandslied	161	0.39%
8	rÄmische zahlen	111	0.27%
9	sagrada familia	105	0.25%
10	www.humboldt-uni.de	104	0.25%
11	ascii	104	0.25%
12	druckerei berlin	98	0.23%
13	theaterkassen berlin	92	0.22%
14	darwinismus	91	0.22%
15	bernsteinzimmer	89	0.21%
16	russia	88	0.21%
17	halbwertszeit	85	0.2%
18	second hand berlin	81	0.19%
19	kolumbus	81	0.19%
20	gaudi	79	0.19%

Report 06 - Navigation

- Wird die Navigation (interne Links) innerhalb der Site richtig eingesetzt? Navigationspfade der Besucher? Click-by-Click Analyse
- Gehen die Besucher die „richtigen“ Pfade? Vollenden sie ihre Aufgaben, Ziele? Mögliche Hindernisse?
- Werden die Hauptziele der Site erfüllt?
Z.B. Newsletter bestellen, Whitepaper downloaden
- Ziel: verbesserter und schnellerer Zugang zu den „eigentlichen“ Informationen



Report 06 – Navigation cond.

Top Paths Through Site			
Starting Page	Paths from Start	Visits	%
Homepage	1. Welcome Information /home.asp 2. Store Information /store/ 3. Wireless phones View /store/wireless_phones.asp 4. W1000 Information /store/wireless_phones/view.asp 5. Add Product to Cart /store/add.asp	41	1.80%
	1. Welcome Information /home.asp 2. Store Information /store/ 3. Wireless phones View /store/wireless_phones.asp 4. W1000 Information /store/wireless_phones/view.asp	30	1.32%

Report 07 - Exit

- Wo steigen die Besucher aus?
- Steigen die Besucher zu früh aus? (z.B. auf Seiten, die nicht zum Aussteigen konzipiert sind)
- Die Exit-Seiten „erzählen“ über die Interessen und das Besucherverhalten
- Die Exit-Seiten liefern Verbesserungshinweise (Inhalte, Struktur)

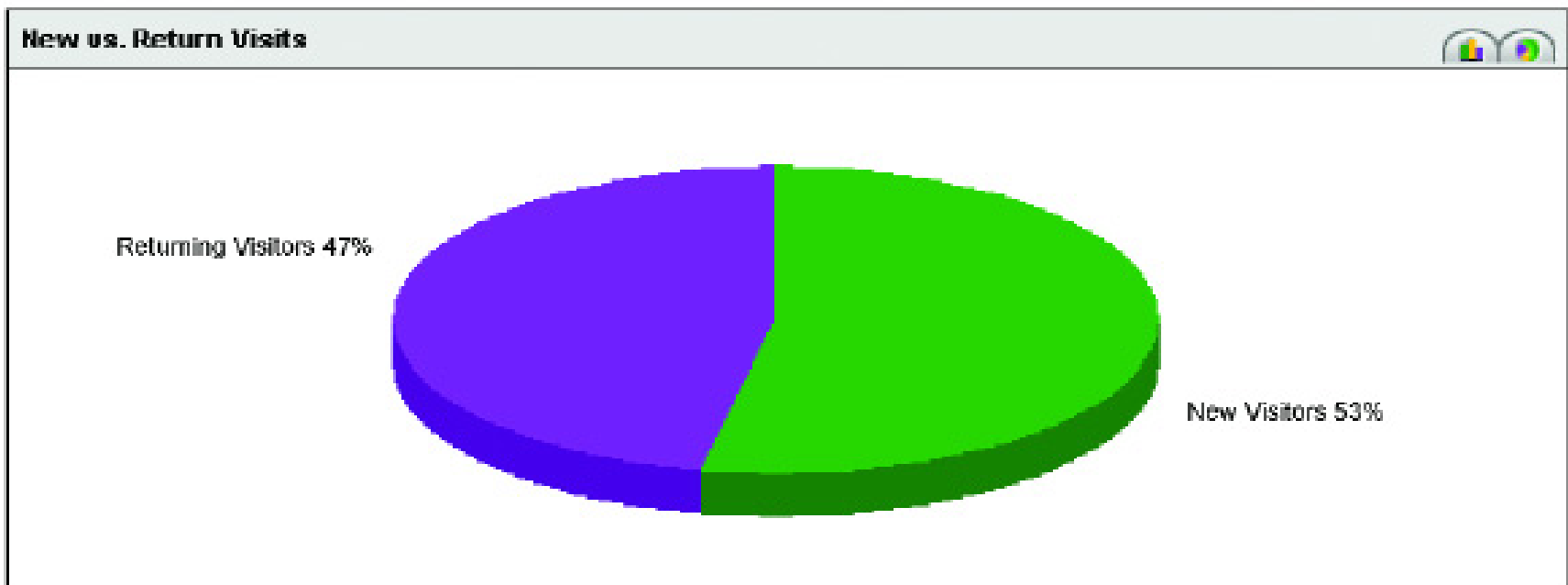
Report 08 – Single Access

- Wo sind „Löcher“ in meiner Site? Single Access Pages
- Werden die „richtigen“ Besucher angezogen?
- Ziel: die neuen Besucher sollen die Website kennenlernen und daher mehrere Seiten aufrufen
- Seiten mit Navigationsproblemen werden sichtbar (worst case: keine oder schlecht lesbare Links)

Report 09 – Alt vs. Neu

- Kommen die Besucher zurück? Welche Organisationen liefern dauerhaft die meisten Besucher?
- Alte vs. neue Besucher (Neuakquisition vs. Bewahrung)
- Verhältnis von alt/neu – ein Maß von Treue und Zufriedenheit
- Anreize für Besuche schaffen (Newsletter, Seminare, Gewinne, neue und aktuelle Inhalte)

Report 09 – Alt vs. Neu cond.

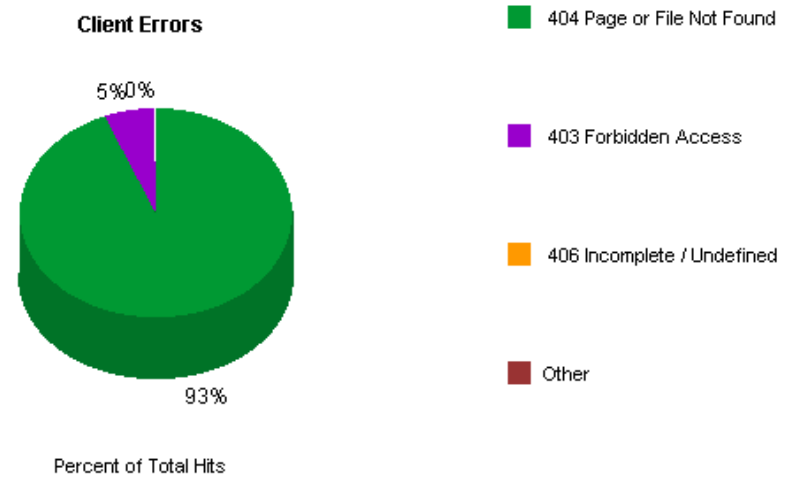


New vs. Return Visits

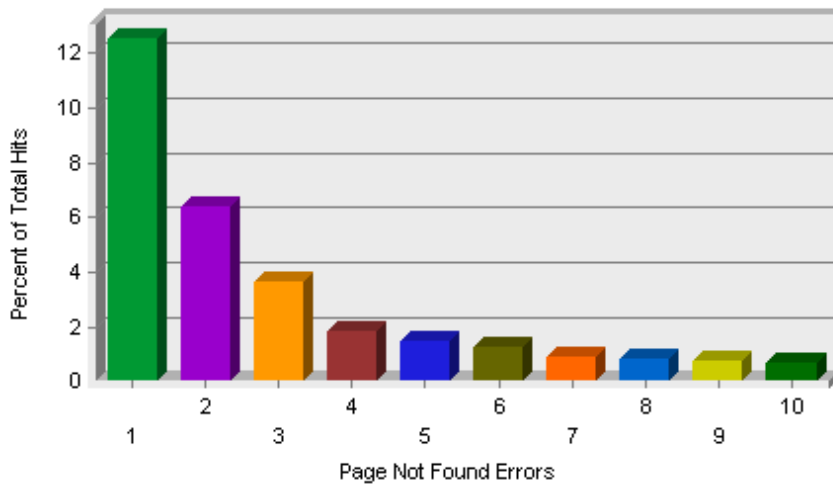
Visitor Type	Visits	%
1. New Visitors	8,781	53.12%
2. Returning Visitors	7,748	46.87%
3. Visitors Not Accepting Cookies	2	0.01%
Total	16,531	100.00%

Report 10 - Probleme

- Gibt es technische Probleme?
- Z.B. Client Errors, Server Error
File Not Found Errors



Page Not Found (404) Errors



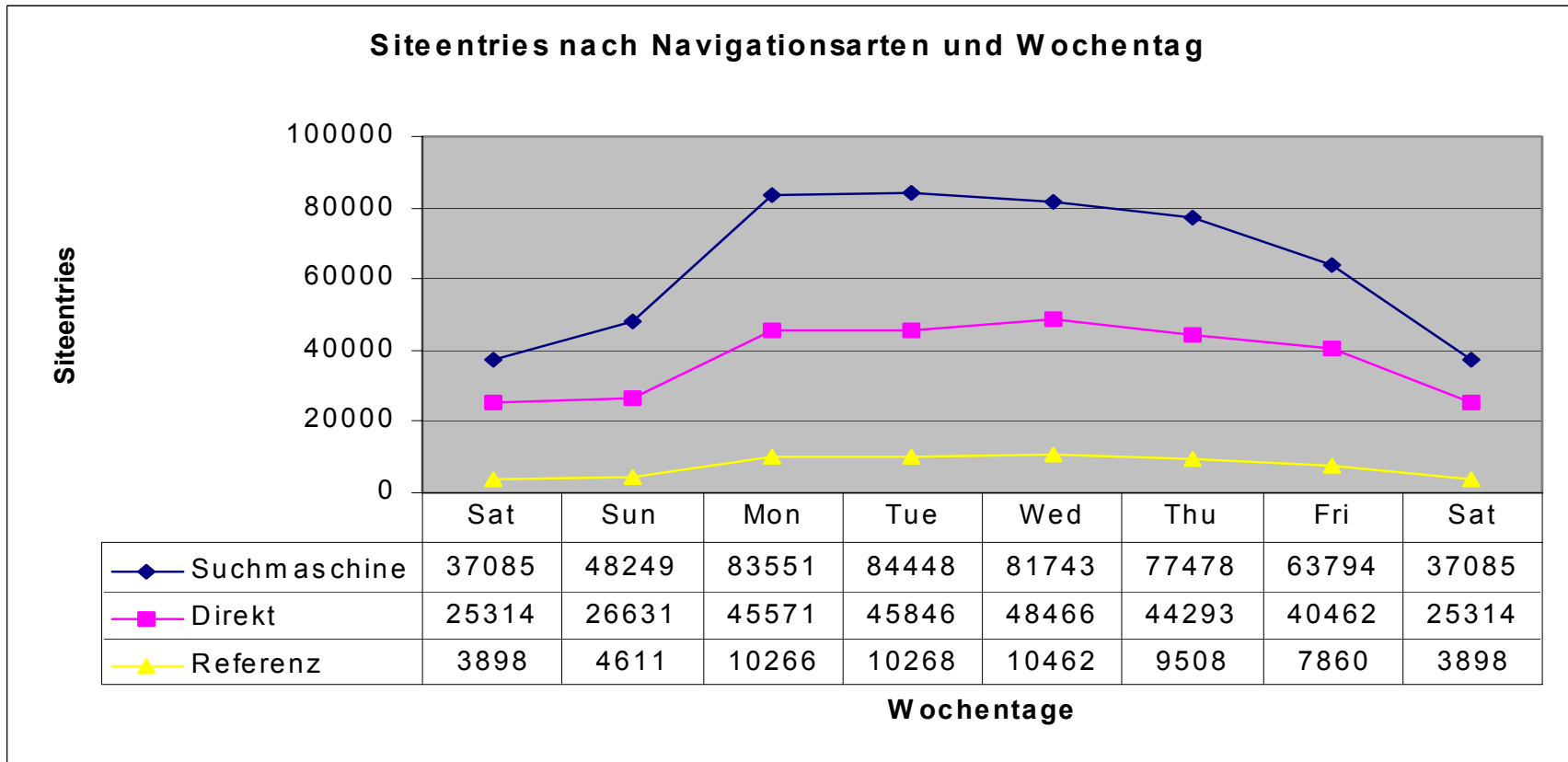
- 1- /robots.txt
- 2- /deutsch/presse/bilder/bi
- 3- /favicon.ico
- 4- /deutsch/presse/archiv_
- 5- /presse_index.html
- 6- /aktuell_index.html
- 7- /<Rejected-By-UrlScan>
- 8- /deutsch/presse/archiv_
- 9- /foerdernews.html
- 10- /branchen_index.html

Erweiterte Loganalyse

- Untersuchungsfokus der Studie: Sichtbarkeit von Websites (über die 3 Zugangsarten)
- Spezielle Anwendung: Web Entries am Beispiel einer großen deutschsprachigen akademischen Website
- Korrelation unterschiedlichster Parameter (PageRank und Traffic, Seitentyp und Traffic bzw. Zugangsart, ...)

Erweiterte Loganalyse cond.

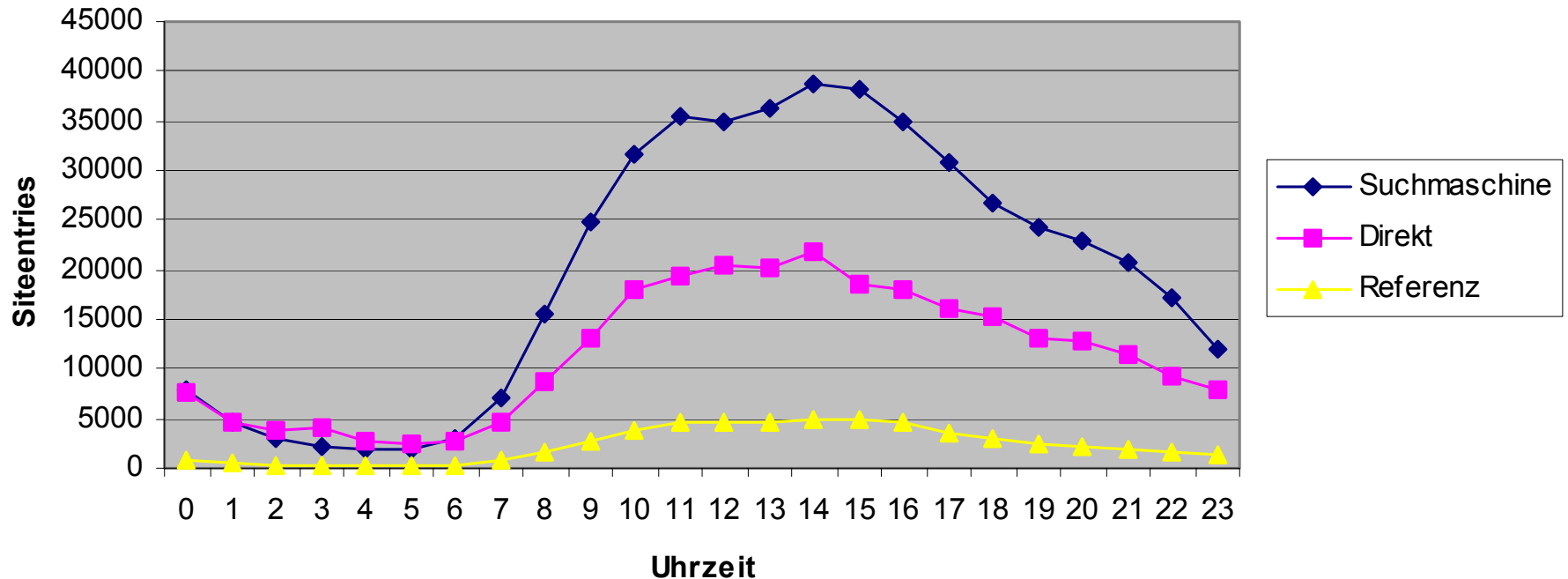
- Ergebnisse: Suchmaschinen sind die wichtigsten Traffic-Lieferanten (insb. neue Besucher)
- Eintrittsarten unterschieden nach Tag



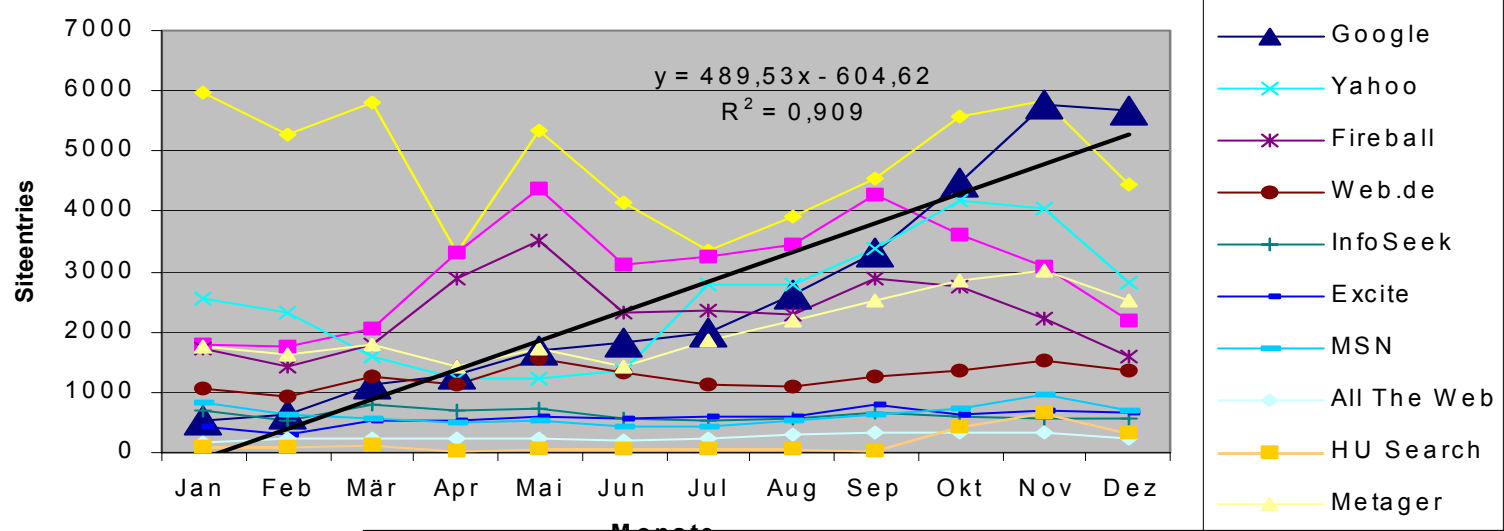
Erweiterte Loganalyse cond.

- Ergebnisse: Eintritte über Suchmaschinen, Bookmarks und externe Links, unterschieden nach Uhrzeit

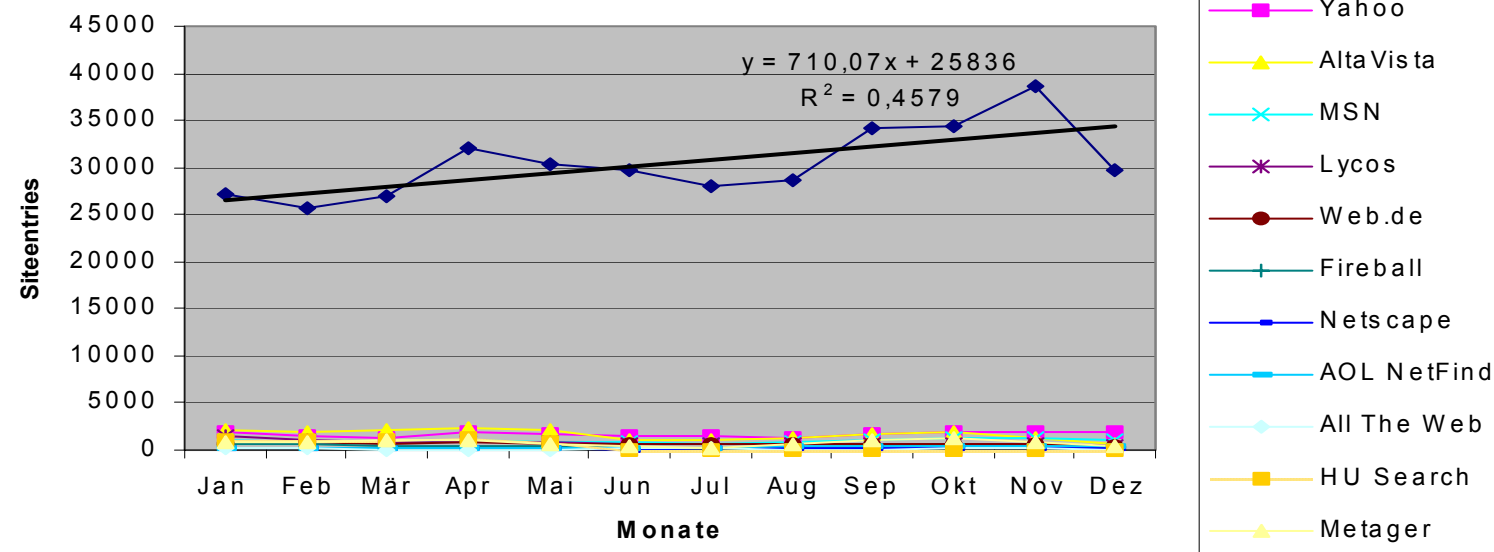
Siteentries nach Navigationsarten und Uhrzeit



Top Suchmaschinen 2000

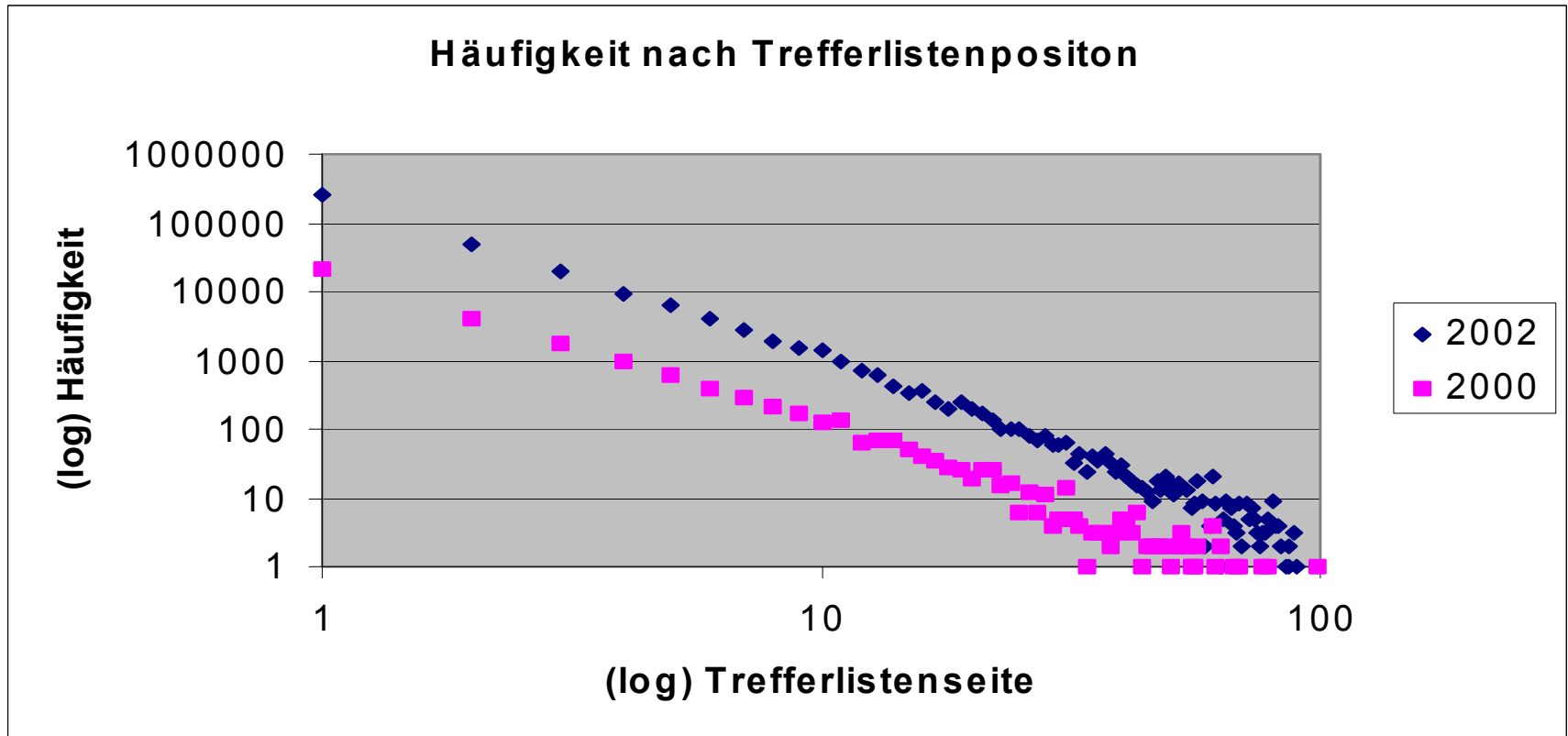


Top Suchmaschinen 2002



Erweiterte Loganalyse cond.

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





Erweiterte Loganalyse cond.

Eigene Anwendungen – Schritte

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

Erweiterte Loganalyse cond.

[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

Web Entry Miner - WEM

Web Entry Miner
_ □ ×

About WEM

Configuration

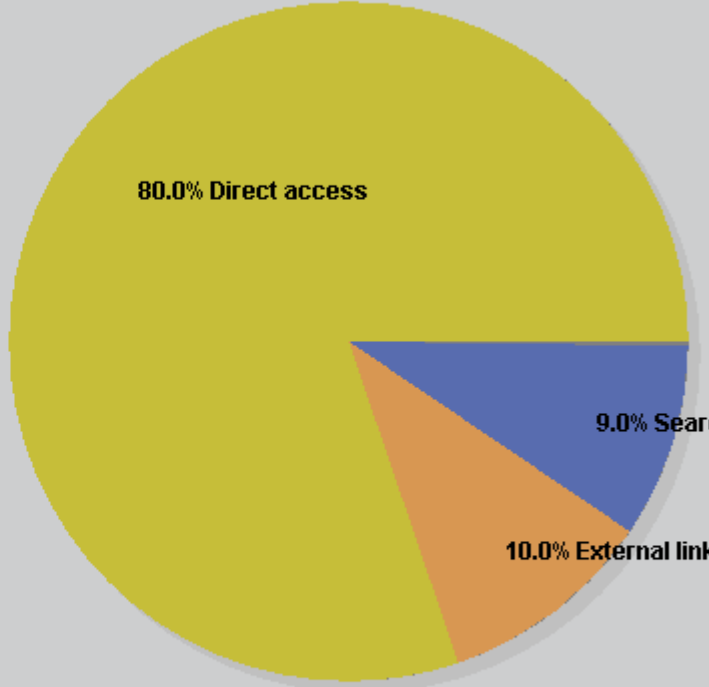
Enter domain:

Entries' Type

Top 100

URL	SE-Entries	D-Entries	R-Entries
/	133	1156	15
/issue2_4.htm	1	6	38
/index2.html	3	1	19
/free.htm	2	18	
/content_0203.htm	1	17	
/toc.htm	0	10	
/hilights.htm	4	5	
/index.html	0	6	
/internship.htm	2	1	
/issue1_2.htm	3	2	
/authors.htm	1	4	
/news2_1.htm	0	3	
/issue1_1.htm	1	2	
/papers.htm	4	0	
/issue2_3.htm	0	3	
/issue1_4.htm	2	1	
/about-us.htm	0	2	
/issue1_3.htm	3	0	
/content_0103.htm	0	2	
/issue2_1.htm	0	1	
/content.htm	0	1	
/index.htm	0	1	
/thanks_v.htm	0	0	

by chris.na



80.0% Direct access

9.0% Search engines

10.0% External link

Configure and start analyse first...

Website Auditierung

- Indexierungsfortschritte, -probleme durch die Suchmaschinen Robots (Roboterverhalten)
- Anteil der von Suchmasch. indexierten Seiten
- Positionsreporting, Kontrolle des Suchmaschinen-Rankings

"<http://www.google.de/search?q=fernstudium&sa=N>

"<http://www.google.de/search?q=fernstudium&start=10&sa=N>

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

"<http://www.google.de/search?q=fernstudium&start=80&sa=N>

Erweiterte Loganalyse

- Webmining = Anwendung von Data-Mining-Techniken auf Inhalt, Struktur und Nutzung von Webressourcen



The screenshot shows the Amazon.com product page for the book "Data Mining: Building Competitive Advantage" by Robert Groth. The page includes a search bar, navigation tabs, and a "Customers who bought this book also bought:" section. A callout box highlights the "Customers who bought this book also bought:" section.

Customers who bought this book also bought:

- [Building Data Mining Applications for CRM](#); Alex Berson, et al
- [Mastering Data Mining: The Art and Science of Customer Relationship Management](#); Michael J. A. Berry, Gordon
- [Data Mining Your Website](#); Jesus Mena
- [The Data Webhouse Toolkit : Building the Web](#); Ralph Kimball, Richard Merz



Zusammenfassung

Reportnummer und -name	Beschreibung	Priorität
01 Traffic	Wie viel Traffic erhält die Site?	A - B
02 Beliebteste Bereiche	Welche Seiten ziehen die Besucher an?	A
03 Einstiegsseiten	Die wichtigsten Einstiegsseiten	A
04 Sichtbarkeit	Woher kommen die Besucher?	A
05 Suchmaschinen	Welche Suchmaschinen bringen den Traffic? Über welche Begriffe?	A - B
06 Navigation	Einsatz der Navigation	B
07 Exit	Wo steigen die Besucher aus?	A - B
08 Single Access	“Löcher” innerhalb der Site	B
09 Alt vs. Neu	Alte vs. Neue Besucher auf der Website	B
10 Probleme	Welche technischen Probleme gibt es?	B



Ausblick

- Die „richtigen“ Daten erheben
- Eigene Anwendungen/Reports sind Standardauswertungen meist überlegen
- Loganalyse bzw. Webcontrolling wird zunehmend akzeptiert und für wichtig erachtet
- Kombinierte Untersuchung heben den Wert der Aussagen (z.B. Referer via Suchmaschine + Logfile)
- Methoden Mix – quantitative und qualitative Analysen (z.B. Logfile, Fragebogen, Voting, andere Datensammlungen, ...)

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

Artikel

- Methodische Anmerkungen zur Auswertung der WWW-Log-Dateien des Servers www.gesis.org / von Wolf-Dieter Mell, 2002, IZ-Arbeitsbericht Nr. 26, available: http://www.gesis.org/Publicationen/Berichte/IZ_Arbeitsberichte/#ab26
- 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
- Web log file analysis: backlinks and queries / von Mike Thelwall, in: Aslib Proceedings, 2001, Vol. 53, No. 6
- Web-Statistik – Potenziale und Grenzen / von Simone Fühles-Ubach, in: b.i.t. online, 2001, 4, available: <http://www.b-i-t-online.de/archiv/2001-04/fach1.htm>

Literatur

Bücher

- Statistische Anwendungen im Internet. In Netzumgebungen Daten erheben, auswerten und praesentieren / von Dietmar Janetzko, 1999, Addison-Wesley, München, ISBN 3827314313
- Perl for Web Site Management / von John Callender, 2001, O'Reilly, ISBN 1-56592-647-1, 528 S.



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

HiSolutions AG

Philipp Mayr

Bouchéstrasse 12

D - 12435 Berlin

Tel.: +49-(0)30 / 533289 – 0

email: mayr@hisolutions.com,

mayr@informatik.hu-berlin.de (privat)

www: www.hisolutions.com/