

*Hi*solutions

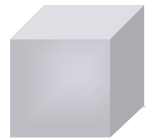
Logfile Analyse II

Referent: Philipp Mayr

8. InternetSalon am 26. Mai 2004, pr-ide

© 2004, HiSolutions AG

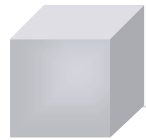
*Hi*solutions



Agenda

Dauer ca. 1,5 h

- Begrüßung
- Einführung, Grundlagen, Probleme, Potenziale
- Zehn Reports am Beispiel von vers. Logfiles
- Fragen
- Zusammenfassung & Ausblick
- Exkurs: Erweiterte Logfile-Analyse



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, serverplattformabhängige, konfigurierbare 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]

A 3D isometric cube icon, rendered in a light gray color with a slight shadow, positioned to the left of the main title.

Grundlagen

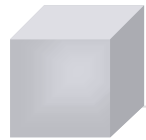
Loganalyse Ablauf:

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

A 3D isometric cube icon in shades of gray.

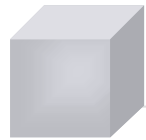
Analyse-Kriterien

1. Einfache Nutzungsmaße (Hits, Views, Impressions, Visits)
2. Besuchercharakteristiken (Nationalität, Organisation, Vielkäufer, Subscriber, ...)
3. Suchcharakteristiken (Themen, Trends, Suchbegriffe, Informationsbedürfnis)
4. Navigationspfade (z.B. Muster)



Begriffe

- **Hit:** kleinste Maßeinheit für einen Zugriff, relativ 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)"
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)"
```

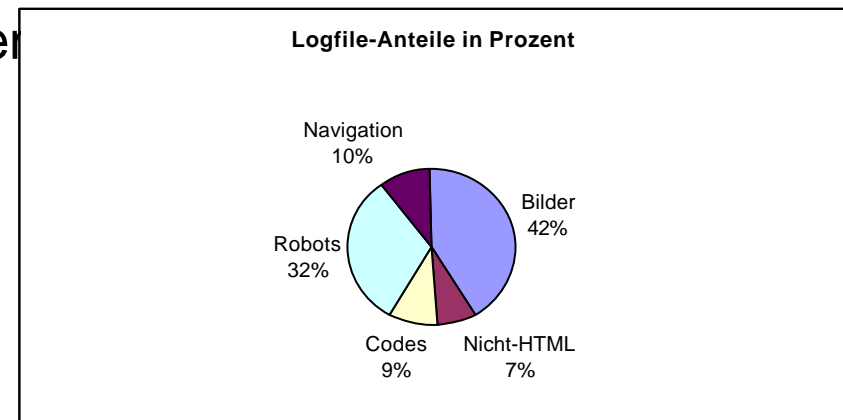
Weitere Beispiele

9 Hits, 4 Views, 2 Visit, 2 Visitors

```
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)"
```

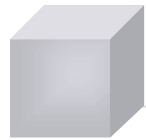
Probleme

- Filtern und Säuberung der Daten (Hit ist nicht gleich Hit)
[etwa 10% der Daten verbleiben nach Data Cleaning
(Preprocessing) 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 angereichert werden
(dynamische Inhalte)



Probleme (cont.)

- Virtual users ungleich end user
- HTTP: 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 (z.B. 1 Tag auf Seite 1 bei Google)



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

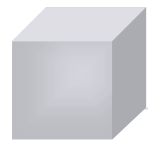
Beispiel Logdaten

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 (cont.)

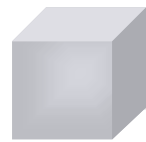
Beispiel Logzeile - WER?

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)

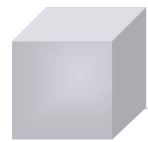


Beispiel Logdaten (cont.)

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

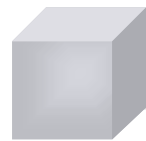


Beispiel Logdaten (cont.)

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 (cont.)

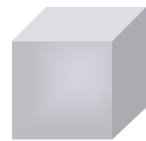
Beispiel Logzeile - was?

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

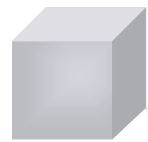


Beispiel Logdaten (cont.)

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 (cont.)

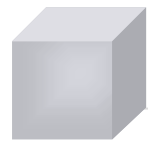
Beispiel Logzeile - WIEVIEL Daten?

```
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 (cont.)

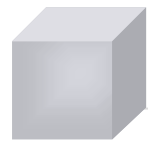
Beispiel Logzeile - WOHER?

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 (cont.)

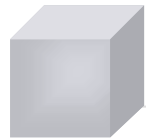
Beispiel Logzeile - womit?

```
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,

Zehn typische Reports

1. Wie viel Traffic erhält die Site, Directory, Page, ...?
2. Die beliebtesten Bereiche, Sparten & Themen der Site ...
3. Die wichtigsten Einstiegsseiten der Site ...
4. Woher kommen die Besucher?
5. Welche Suchmaschinen bringen Traffic über welche Begriffe?
6. Wird die Navigation auf der Site richtig eingesetzt? Finden die Besucher die gewünschte Information?
7. Steigen die Besucher zu früh aus? Exit-Seiten
8. Wo sind „Löcher“ in meiner Site? Single Access Pages
9. Kommen die Besucher zurück? Neue vs. alte Besucher
10. Gibt es technische Probleme mit der Site?



Analyse der Logfiles

- Tool: Webtrends Reporting Center 6.1
- Logdaten:
 - www.protektor-ag.de
Protektor Lebensversicherungs-AG
 - www.cdh.de
Verband Handelsvertreter Vertrieb Deutschland
 - www.ape-berlin.de
Ultrafast Laser Pulse Measurement
 - www.zab-brandenburg.de
Zukunftsagentur Brandenburg

Profiles

 View Reports
 New
 Edit
 Copy
 Delete

Analysis

 Analyze Now
 Stop Analysis
 Re-Analyze
 Clear

Status

 Profile Status
 Stop Auto Status

Archives

 List Archives
 Archive Now

Profile Name	Log Data Sources	Last Analysis
internetsalon_ape	internetsalon_ape	05/27/2004 10:00
internetsalon_cdh	internetsalon_cdh	05/27/2004 10:00
internetsalon_disinfo	internetsalon_disinfo	05/27/2004 10:00
internetsalon_ib	internetsalon_ib	05/27/2004 10:00
internetsalon_protekt	internetsalon_protect	05/27/2004 10:00
internetsalon_zab	internetsalon_zab	05/27/2004 10:00
IS_ape	internetsalon_ape	05/26/2004 14:05
IS_cdh	internetsalon_cdh	05/26/2004 14:04
IS_ib	internetsalon_ib	05/26/2004 03:58
IS_Protektor.de	internetsalon_protect	05/26/2004 14:04
IS_zab	internetsalon_zab	05/26/2004 09:57
sample: Streaming Media (RealServer)	sample: RealServer log	05/26/2004 16:53
TC_Intranetweb	TC_Intranetweb	05/27/2004 10:00

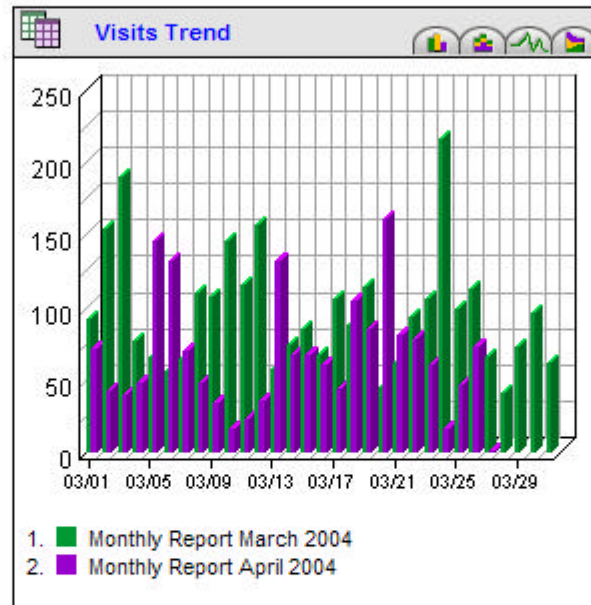
Report Calendar	http://io:1099/viewreport.pl?profileid=4.wlp
Profile File Name	4.wlp
Web Site URL	http://www.ape-berlin.de/
Next Analysis	11:00 05/27/2004
Status	Waiting for next scheduled analysis (Analysis completed)



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%

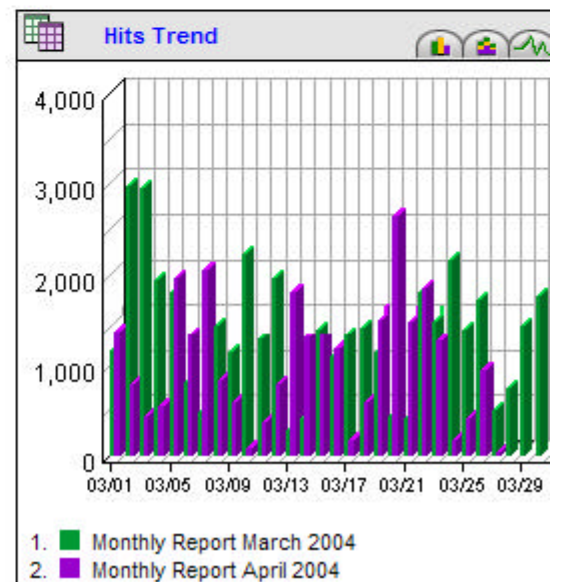
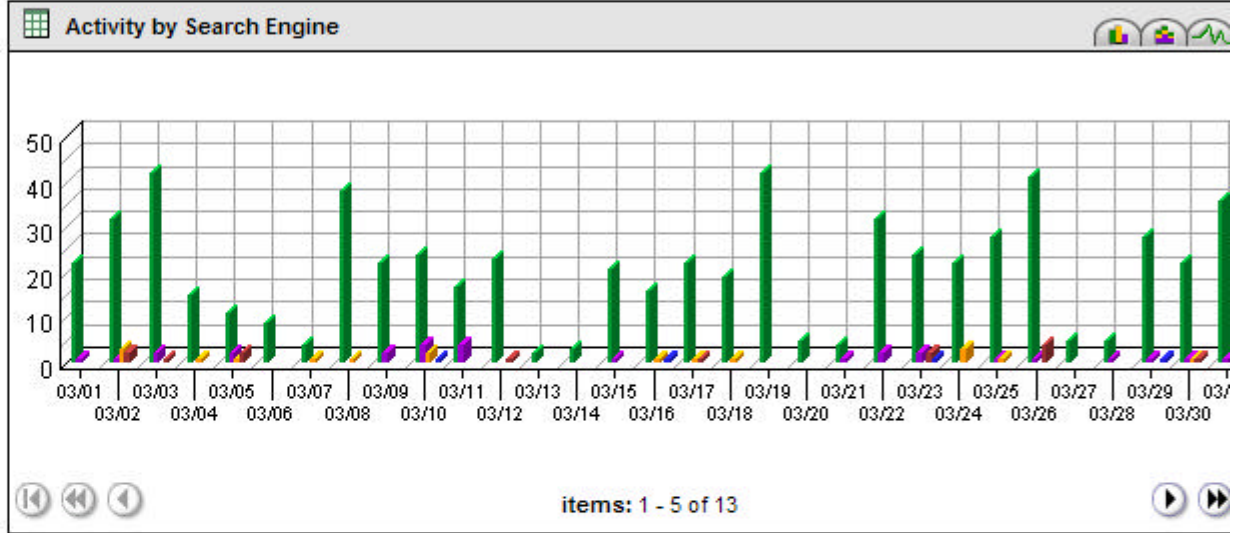




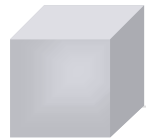
Table of Contents

- Overview
- Commerce
- Marketing
 - Marketing Dashboard
 - Campaigns
 - Referrers
 - Referrers Dashboard
 - Activity by Referring Site
 - Activity by Referring Domain
 - Activity by Referring Page
 - Search Engines
 - Search Engines Dashboard
 - Activity by Search Engine**
 - Activity by Search Phrase
 - Activity by Search Keyword
- Products



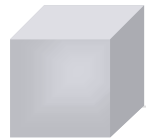
Activity by Search Engine

Engines	Referrals	
1. google germany	636	89.4
2. google	28	3.9
3. yahoo germany	17	2.3
4. msn	14	1.9
5. google austria	4	0.5
6. web.de	3	0.4
7. google france	2	0.3
8. google uk	2	0.3
9. google japan	2	0.3
10. altavista	1	0.1
11. ixquick	1	0.1
12. ...	1	0.1



Fragen

- Was kann mit einer Logfile-Auswertung erreicht werden?
- Wie sicher sind die Angaben und wo gibt es Risiken?
- Wie oft sollte ich eine Logfile-Auswertung vornehmen und ab wann kann ich von einem bestimmten Trend im Nutzerverhalten feststellen?
- Welchen Zeitraum benötige ich, um nach Veränderungen auf der Webseite auch Veränderungen im Nutzerverhalten festzustellen?
- ...



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

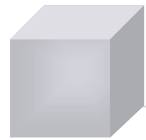
Open Source

- 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

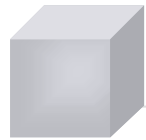
- Methodische Anmerkungen zur Auswertung der WWW-Log-Dateien des Servers www.gesis.org / von Wolf-Dieter Mell, 2002, IZ-Arbeitsbericht Nr. 26, available: http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/#ab26
- Cracking the Code: Web Log Analysis / von David Nicholas et al., in: Online & 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>

A 3D isometric cube icon, light gray with darker shading on the top and right faces.

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

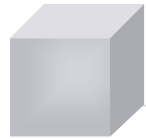
Charakteristische Zitate:

“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 ...“

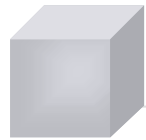
„The trouble, of course, is that there is no single measure of consumption and each measure has to be taken with a large dose of statistical salt.”

[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: <http://www.hisolutions.com>