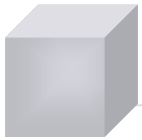
Einführung in die Logfile Analyse

Referent: Philipp Mayr

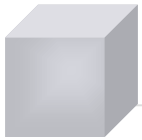
7. InternetSalon am 31. März 2004, pr-ide



Agenda

Dauer ca. 1,5 h

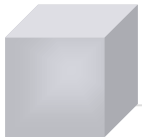
- Begrüßung
- Einführung, Grundlagen, Probleme, Potenziale
- Zehn typische Logfile Auswertungen
- Erweiterte Logfile Analyse (Beispiele)
- Zusammenfassung, Ausblick



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

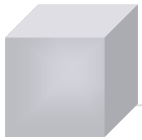
Fakten

- 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:

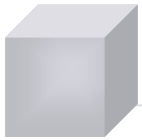
1. Tracking des Traffics zur Etablierung v. Benchmarks (Bezugsgrößen)
2. Identifikation von Verbesserungsmöglichkeiten und Änderung der Site
3. Messung der Auswirkungen bzw. Effektivität der Änderung
4. Wiederholung dieses Prozesses



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

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)"

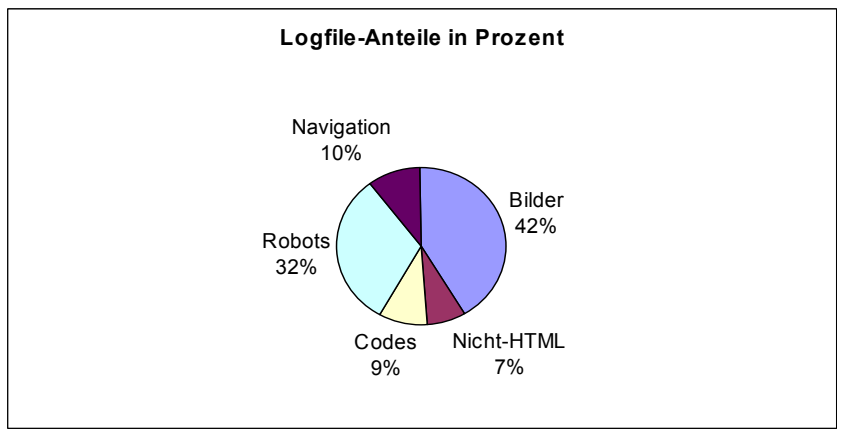
- Download (2 Hits, aber nur eine Datei)

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)"

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?

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, …)

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

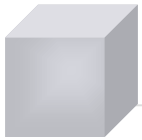
"http://www.**google**.**de**/search?q=**fernstudium**&**start=80**&sa=N

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/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

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

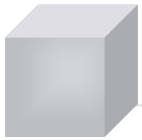
Charakteristische Zitate:

“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 …“

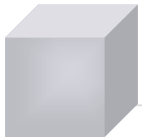
„The trouble, of course, is that there is no single measure of consumption and each measure has to be taken with a large dose of statistical salt.”

[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/