

Fachhochschule Stuttgart –
Hochschule der Medien

15.06.03

Referat zum Thema:
Logfile-Analyse und Webmining

Seminar: Website-PR
Dozent: Prof. Dr. Grudowski

1 Verfasser: Marc Kluge
Sven Szigeti
Monika Biedlingmeier

6. Semester IW

Inhalt:

| | | |
|------|--|----|
| 1 | Einführung; | 3 |
| 1.1 | Logfile: | 3 |
| 1.2 | Webmining: | 3 |
| 1.3 | Hit: | 3 |
| 1.4 | PageImpression bzw. PageView : | 3 |
| 1.5 | Visit: | 3 |
| 2 | Einordnung in die Thematik der Website-PR..... | 4 |
| 3 | Logfile-Analyse | 4 |
| 3.1 | Einführung | 4 |
| 3.2 | Aufruf einer Website..... | 4 |
| 3.3 | Inhalt und Formate von Logfiles | 5 |
| 3.4 | Bestandteile des CLF: | 5 |
| 3.5 | Bestandteile des ECL-Formats..... | 5 |
| 3.6 | Probleme bei der Analyse | 6 |
| 3.7 | Mögliche Lösungsansätze | 7 |
| 3.8 | Ausgewählte Analyse-Tools | 7 |
| 3.9 | Beispiele graphischer und numerischer Logfile-Statistiken..... | 8 |
| 3.10 | Quellen- und Literaturangaben:..... | 9 |
| 3.11 | Literaturangaben zum Teil Logfile-Analyse | 9 |
| 4 | Web Mining: Grundlagen..... | 10 |
| 5 | Die Datengrundlage | 11 |
| 5.1 | Datengrundlage für die Analyse | 11 |
| 6 | Web Usage Mining | 11 |
| 6.1 | Personalisiertes Web Usage Mining | 12 |
| 6.2 | 4.2 Nicht personalisiertes Web Usage Mining..... | 12 |
| 7 | Web Content Mining | 12 |

1 Einführung;

Zu Beginn des Referats werden zunächst einmal einige Begriffe definiert, die für das Verständnis der Thematik ausschlaggebend sind, und anschließend wird die Bedeutung der Thematik für die Website-PR erläutert.

1.1 Logfile:

Logfiles sind Dateien, die vom Webserver angelegt und verwaltet werden. Es handelt sich dabei um eine Art Webserverstatistik, in der jeder Besuch eines Online-Angebots protokolliert wird. Dabei werden eine Vielzahl von Informationen festgehalten, z.B. Datum und Uhrzeit des Zugriffs oder die Namen der angeforderten Dateien. Logfiles werden in der Regel monatlich neu angelegt und von Site Hostern an ihre Kunden entweder komplett übermittelt oder als Zusammenfassung im WWW angeboten.

1.2 Webmining:

Beim Webmining handelt es sich um eine spezielle Form des Data-Minings unter Einschluß von online erhobenen Daten, die in vielen Fällen ebenfalls aus Logfiles gewonnen werden. Das heißt, man versucht in umfangreichen Datenbeständen Regelmäßigkeiten, Auffälligkeiten oder komplexe Zusammenhänge zu finden. Eine mögliche Anwendung des Webmining wäre die Erstellung von Kundenprofilen, aufgrund der Daten aus dem Logfile.

Man unterscheidet zwischen Web-Content-Mining, bei dem es um die Analyse der im Netz bereits befindlichen Daten geht, im Gegensatz zu Web-Usage-Mining, bei dem im Zentrum die Interaktion des Nutzers mit dem Internet steht.

1.3 Hit:

Darunter versteht man jeden Zugriff auf einen Webserver, den ein Seitenaufruf erzeugt, d.h. jede Anforderung zum Laden einer Datei. Dadurch kann jede Website viele Hits erzeugen (beispielsweise wäre jeder Button, jede Graphik usw. ein Hit). Deshalb kann die Anzahl der Hits nicht als Indikator für die Zahl der Besucher auf der Website gesehen werden.

1.4 PageImpression bzw. PageView :

bezeichnet jeden Sichtkontakt des Benutzers mit der Seite. Der Wert gibt an, wieviele komplette Seitenaufrufe stattgefunden haben, unabhängig von der Anzahl der darin enthaltenen Elementen und Dateien.

1.5 Visit:

definiert die Anzahl der Besucher einer Website, ohne zwischen Neu- und Mehrfachkontakten zu differenzieren. Definiert wird ein Visit als Nutzungsvorgang, wenn zwischen dem letzten und dem aktuellen externen Seitenabruf mindestens 60 Sekunden liegen.

2 Einordnung in die Thematik der Website-PR

Das Ziel dieser beiden Werkzeuge aus der Sicht der PR ist die Websiteoptimierung durch eine Nutzungsanalyse.

Der Anbieter muss folgende Informationen über die Nutzer seiner Website wissen, um sie zu optimieren und Fehler darin zu erkennen und zu beseitigen

- wie die Nutzer zur Website gelangen
- über welche Seite sie das Angebot wieder verlassen
- welche Aktionen sie in der Zwischenzeit ausführen (dabei würden die Betreiber gerne wissen, ob der Benutzer seinen Weg durch das Angebot gut findet oder umherirrt, welche Seiten er sich warum länger ansieht usw.)
- welche Pfade sie durch die Website nehmen
- welche Suchbegriffe die Nutzer in Suchmaschinen eingeben, um zur entsprechenden Website zu gelangen (um zu erkennen, was den betreffenden Nutzer interessiert, was er sich davon erhofft usw.)

Welche dieser Informationen Logfiles den Anbietern geben können:

- Mit der Entwicklung der Anzahl der Besucher lässt sich die Wirkung von kurz- oder langfristigen Marketingaktivitäten verfolgen
- Es können Hauptbereiche des Online-Angebots ermittelt werden, die am meisten genutzt werden, um sie dann verstärkt zu verbessern und auszubauen
- Bereiche, in denen die Nutzer am längsten verweilen können herausgefiltert werden.
- Typische Pfade der Benutzer durch das Webangebot können ermittelt werden
- Begriffe, über die die meisten Suchmaschinentreffer erzielt wurden können als Hinweis dafür gesehen werden, welche Themengebiete den Nutzer am meisten interessieren.

3 Logfile-Analyse

3.1 Einführung

Web-Server speichern die Dokumente, Skripte und weitere Daten und Anwendungen einer oder mehrerer Websites, d.h. HTML-Text, Grafiken und andere Anwendungen.

Zum Zweck der Auswertung der Daten über Zugriffe auf den gespeicherten Webseiten, ist eine Funktion in der Betriebssystem-Software des Web-Servers implementiert, die die einzelnen Zugriffe automatisch und sequenziell in einer Datei als ASCII-Text speichern kann. Die dabei angelegte Datei wird "Logfile" oder "Logdatei" genannt. Schnell nehmen diese Dateien jedoch enorme Größen an. Obwohl die Daten als für den Menschen lesbarer Text vorliegen, erscheint die Interpretation, auch aufgrund der vielen Textzeilen, schwierig. Die Lösung des Problems wird durch den Einsatz von Analyseprogrammen erreicht, die diesen Teil des Website Monitorings zu leichter verständlichem Textmaterial verdichten und durch Grafiken erweitert darstellen. Funktionen und Problematiken solcher Tools werden anhand von Beispielen im Verlaufe des Texts erläutert.

3.2 Aufruf einer Website

Zum Aufruf einer Seite im World Wide Web (WWW) macht der Internet-Browser vom Übertragungsprotokoll HTTP (Hypertext Transport Protocol) Gebrauch. Durch den Aufruf des URLs (Uniform Resource Locator) wird eine Anfrage (Request) des Typs GET an den Web-Server gesendet. Der Server startet eine Anwendung oder sendet eine Antwort zurück (Response), und zwar im Falle einer korrekt vorliegenden Seite deren HTML- und sonstige Dateiinhalte. In der vorgenannten Hierarchie "User- Visit- Page View - Hit" wird nun für jeden Hit, also für jede noch so kleine übertragene Datei ein Log-Eintrag geschrieben.

Im HTTP-Protokoll der Version 1.0 wurde die Verbindung zwischen dem Browser (auch Client genannt) nach erfolgreicher Interaktion wieder getrennt, um die Bandbreite der Internet-Verbindung nicht unnötig zu belasten. Diese Trennung ist der Grund dafür, dass

Logfiles aus vielen einzelnen Textzeilen für jeden einzelnen Hit bestehen.

Hieraus kann man leicht erkennen, dass es sich lediglich um eine chronologische Auflistung von aufeinanderfolgenden Aufrufen handelt, die über die Nutzung der Seite und ihren Wert für den Nutzer noch keinen Aufschluss bietet.

3.3 Inhalt und Formate von Logfiles

Das von allen Betriebssystemen unterstützte Standardformat CLF (Common Logfile Format) definiert die Werte, die mindestens erfasst werden. Dieses Format wurde durch die W3C definiert (2). Weitere vor allem für das Marketing relevante Daten geben jedoch die zwei Bestandteile des ECLF (Expanded CLF) an, wie weiter unten dargestellt. Beide Formate kombiniert nennen sich CbLF (Combined Log File Format). Darüber hinaus existieren proprietäre Formate des Microsoft IIS, IIS Extended, Lotus Domino, Apache extended, NCSA Combined und andere.

3.4 Bestandteile des CLF:

- remotehost: Statische und dynamische IP-Adresse des zugreifenden Servers (seltener nach Domain aufgelöst)
- rfc 931 (oder ident): Remote-Log-Name des Users. (diese Angabe wird nur selten verwendet)
- Authuser: Authentifizierter Benutzername, sofern eine Anmeldung (Login) auf der Seite notwendig war. Hierbei ist hervorzuheben, dass es sich hier nicht um die Login-Daten beim Internet Service Provider handelt. Eine rekursive Identifikation des Besuchers ließe sich in der BRD nur unter besonderen rechtlichen Umständen anhand der IP-Adresse und der Uhrzeit vornehmen. Durch eine (erzwungene) Anmeldung auf einer Website versucht die dahinter stehende Organisation leichter zuzuordnende und verlässlichere Daten eines Besuchers bzw. eines Nutzers zu gewinnen, z.B. den exakten Clickstream eines Visits.
- Date: Datum und Uhrzeit im Format tt/mm/jj:hh/mm/ss
- Time zone: Angabe der Zeitzone im Format „(+ oder -01:00)“. Diese Angabe beziffert die Abweichung vom Greenwich Mean Time, bezogen auf den Standort des Servers.
- Request: Methode, Protokoll und Dokument des Zugriffs
- Status: Antwortstatus des Web-servers als 3stellige Codenummer aus einer der fünf Klassen.
Beispiel: „200“ – für erfolgreiche Übertragung, „404“ – für Datei nicht gefunden
- Bytes: Gesamtzahl übertragener Bytes (bei fehlerfreier Übertragung gleich der Dateigröße)

Alle Felder werden durch ein Leerzeichen getrennt. Sollten keine Informationen vorhanden sein, wird im Logfile ein Bindestrich "-" eingetragen.

Beispieleintrag:

```
209.185.253.185 - - [08/Oct/1999:05:07:41 +0200] "GET /html/link.htm HTTP/1.0" 200 21575
 "-" "Googlebot/1.0 (googlebot@googlebot.com http://googlebot.com/)"
```

In diesem Fall handelte es sich um einen Zugriff eines Such-Roboters der Suchmaschine Google (www.google.com), erkennbar an der Adresse „...googlebot.com“.

3.5 Bestandteile des ECL-Formats

1) Referrer: Hierbei handelt es sich um den URL der referenzierenden Seite einer internen oder externen Seite. Bei externen URLs wird festgestellt, ob der Besucher seinen Weg zur Seite über einen Banner-Link, einen Suchmaschinen-Eintrag, ein Lesezeichen in seinem Browser, ein Eintrag in der Historieliste oder einen Link etwa auf einer privaten Website gefunden hat. Bei internen URLs kann der Navigationspfad durch die Webpräsenz, der sogenannte Clickstream, nachvollzogen werden. Von besonderem Interesse ist hier, auf welcher Seite der gesamten Website der Nutzer seinen Besuch begonnen hat.

2) Betriebssystem + Browserversion: Diese Angaben geben Aufschluss über die technische Ausrüstung der Nutzer. Besonders bei regional spezifischen Websites kann dies relevant sein, z.B. wenn sich das angesprochene Kundenpotenzial in Ländern informationstechnisch unterentwickelter Infrastruktur befindet.



Abb. 1) Datenkette im ECLF-Format

Darüberhinaus lässt sich feststellen, mit welchen technischen Feinheiten die Site weiterentwickelt werden kann bzw. wie häufig vorhandene genutzt werden, beispielsweise die Anzahl der Besucher mit Flash-Plugin (Browser-Erweiterung zur Darstellung bestimmter Animationen). Flash-Animationen werden häufig auf der Home-Seite der Webpräsenz eingesetzt, weisen jedoch für viele Nutzer lange Ladezeiten auf, was Interessenten bereits beim Einstieg in die Seite abschrecken könnte.

Die eigentlichen Protokolldateien ergänzend werden je nach Web-Server eine Referrer-Protokolldatei (die aufgerufenen und verweisende Adresse angebeude) und eine Error-Protokolldatei generiert, die zur Fehlersuche verwendet werden kann.

Weitere Details über die nutzerspezifischen Informationen können unter <http://www.anonymizer.com> abgerufen werden. Unter <http://www.datenschutz.ch> kann der interessierte Nutzer darüber hinaus einen Test der Integrität und Datensicherheit seines Browsers durchführen.

3.6 Probleme bei der Analyse

Durch den Aufbau des Internet und die simple Konzeption der Logfile-Formate ergeben sich Unschärfen, die selbst das beste Analysewerkzeug nicht ausgleichen kann. Hierbei handelt es sich um

- Zeitliche Eingrenzung von Visits: Da nur sequenziell gesammelte statische Daten vorliegen, ist der zeitliche Ablauf des Nutzers mit seinem Clickstream nicht ersichtlich. Zur Abgrenzung eines Visits muss daher ein realistisches Zeitfenster gesucht werden, da Hits nach Ablauf dieses Zeitraums bereits von einem anderen Nutzer erzeugt werden konnten.
- Eindeutige Identifikation des Nutzers: Durch die Vergabe von dynamischen IP-Adressen bei grossen ISPs ist es insbesondere bei schnellem Wechsel einer IP-Adresse zu einem anderen Nutzer nicht möglich, beide Nutzer voneinander abzugrenzen.
- Caching: Durch das Zwischenspeichern von Seiten auf dem Server des ISP oder des Clients werden erneute Seitenabrufe oft nicht bis zur Server-Website gesendet, sondern aus dem Speicher aufgerufen.
- Proxies: Proxy-Server agieren ähnlich einem Cache. Als zwischengeschaltete Server halten sie eine Auswahl an Websites vor, die der Nutzer oft zwingend zunächst von diesem Server abfragen muss. Die aufrufende IP-Adresse gibt jedoch nur über den Proxy-Server Aufschluss, nicht über den Endnutzer.
- SSL: Beim Aufruf solcher verschlüsselter Seiten, z.B. bei der Übertragung von Formulardaten (Bsp. Onlinebanking) wird generell kein Logfile-Eintrag generiert.
- Robots: Programme einer Suchmaschine, die die Website zum automatischen Indexieren aufrufen, Offline-Reader-Programme u.a. erzeugen eine weitere Unschärfe.
- ...

3.7 Mögliche Lösungsansätze

Um die Verweildauer korrekt messen zu können, kann sich der Administrator der Website für den Einsatz von JavaScript oder eines bestimmten Java-Applets entscheiden.

Für die Abgrenzung der Visits liegt die Lösung in einer Anpassung des Zeitfensters, in der Praxis sind dies meist fünf bis 30 Minuten.

Zur eindeutigen Identifikation des Nutzers liesse sich ein Registrierungszwang einführen. Solange der Nutzer unter seinem Login auf der Seite aktiv ist, lässt sich genau protokollieren, wie er sich durch die Seite navigiert hat und welche Ein- und Ausstiegspunkte er benutzt hat.

Eine andere, jedoch vagere, Lösungsmöglichkeit zur User-Identifikation wären Cookies, kleine Texte oder Textdateien, die an eine bestehende Datei im Client angefügt werden bzw. als einzelne Datei gespeichert werden, durch die der Nutzer „wieder erkannt“ werden kann. Häufig werden sie jedoch vom Client abgelehnt oder vor dem nächsten Besuch auf der Site wieder gelöscht.

Die Probleme, die die dynamische IP-Adressvergabe im Internet mit sich bringt, kann von Seiten des Website-Betreibers nicht gelöst werden, wie auch die Verwendung von Proxies und weitere Aktionen des Nutzers wie etwa das Vor- und Zurücknavigieren im Browser.

Das Caching kann jedoch mit der sogenannten „Pixelmethode“ unterbunden werden: Durch das Einfügen einer unsichtbaren Grafikdatei, die nicht gecached werden kann, werden zu protokollierende Seiten immer wieder neu geladen und erneute Seitenaufrufe erzwungen (von dieser Methode macht das IVW Gebrauch. Die blabla misst PageImpressions und Visits mit einem eigens entwickelten Tool zur Messung der Reichweite der an den Verband angeschlossenen Online-Publikationen (weitere Informationen siehe Handout).

3.8 Ausgewählte Analyse-Tools

| NetIQ Logfile Analyzer 8.0 | Exody WebSuxess 4 | Gekko Analog 5.32 |
|---|--|---|
| <ul style="list-style-type: none"> ■ Website: www.netiq.com ■ Preis: ab US-\$ 499 (für 1 User) ■ Kostenlose Testversion verfügbar | <ul style="list-style-type: none"> ■ Website: www.exody.net ■ Preis: ab € 750 (für 1 User) ■ Kostenlose 30-Tages-Testversion verfügbar | <ul style="list-style-type: none"> ■ Website: www.gekko.net/analog ■ Freeware ■ „The most popular logfile analyser in the world“ ■ In 30 Sprachen erh. ■ Kostenloser Support |

Abb. 2) Kurzüberblick über drei gängige Logfile-Analyse-Programme

Bei den Programmen Logfile Analyzer (aktuelle Version: 8.0, Hersteller: NetIQ) und WebSuxess (aktuelle Version: 4, Hersteller: Exody) handelt es sich um aufwändig gestaltete Werkzeuge, die die Ergebnisse in ansprechender und übersichtlicher Form graphisch in HTML darstellen und durch ihre Datenexportmöglichkeiten (z.B. in ein Data Warehouse) auffallen. Die Lizenzierung dieser Programme ist bereits für einen Einzelanwender im Unternehmen kostenaufwändig (ab US-\$ 499). Kostenlose Testversionen sind verfügbar. Jedoch sind auch Freeware-Programme im Internet erhältlich. Hierzu

lohnt sich ein Blick auf die Website des Herstellers Gekko (URL s. Abb.), der für sein Aufbereitungswerkzeug „Analog“ sogar kostenlose Hilfe anbietet.

3.9 Beispiele graphischer und numerischer Logfile-Statistiken

Unter <http://www.statslab.cam.ac.uk/webstats/reportmagic/> kann man darüber hinaus mit Report Magic generierte Statistiken und Grafiken betrachten. Einige Auszüge hier:

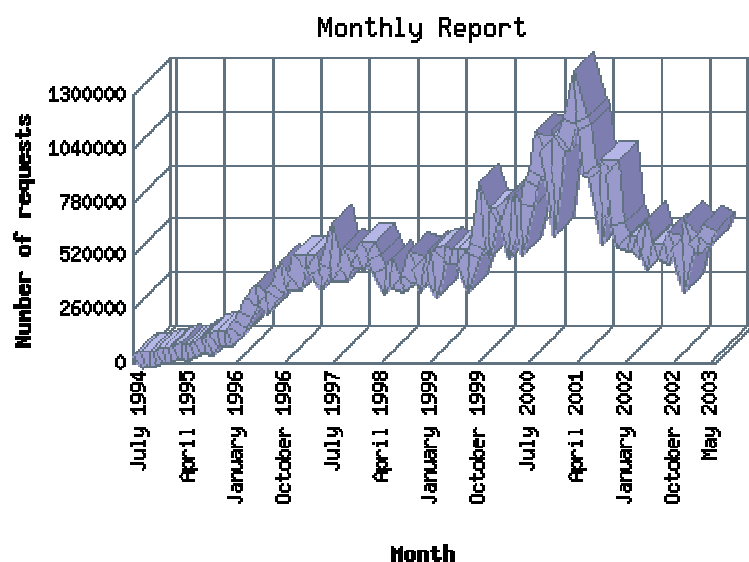


Abb. 2) Monatliche bzw. tägliche „requests“ auf der Seite

| Search Word | | Number of re-quests |
|-------------|------------|---------------------|
| 1. | backgammon | 145,032 |
| 2. | analog | 73,745 |
| 3. | nethack | 34,009 |
| 4. | of | 26,850 |
| 5. | and | 24,869 |
| 6. | log | 23,971 |
| 7. | web | 23,665 |
| 8. | markov | 21,550 |
| 9. | stochastic | 19,907 |
| 10. | statistics | 14,793 |

Abb. 3) Auszug aus der Suchwort-Statistik

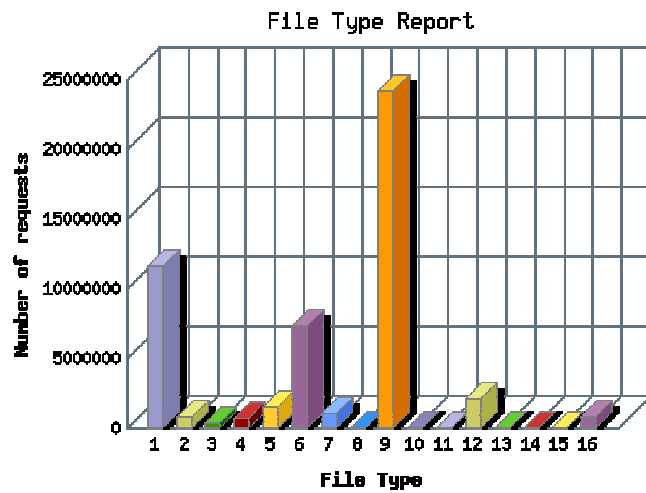


Abb. 4) Auszug aus der Dateitypen-Statistik

3.10 Quellen- und Literaturangaben:

Für die Einführung verwendete Websites:

<http://www.onlinemarketer.de>

<http://www.contentmanager.de>

3.11 Literaturangaben zum Teil Logfile-Analyse

(1)Vgl. Heinrich, Lutz J.; Roithmayr, Friedrich: Wirtschaftsinformatik-Lexikon, 6., vollständig überarbeitete und erweiterte Auflage, München: Oldenburg 1998, S. 194

(2)W3C: <http://www.w3.org>

Statistiken: <http://www.statslab.cam.ac.uk/webstats/reportmagic/>

4 Web Mining: Grundlagen

Das World Wide Web hat in den letzten Jahren als Vertriebskanal dramatisch an Bedeutung gewonnen. Das Internet ermöglicht Unternehmen den unmittelbaren Zugriff auf die globalen Märkte. Oftmals ist das Internet dabei der erste "point of contact" zwischen einem Unternehmen und seinen Kunden. Dementsprechend ist ein überzeugender Internetauftritt ein wichtiges Glied in der Wertschöpfungskette einer Unternehmung. Umso notwendiger ist es diesen Auftritt erfolgreich, zu gestalten. Aufgrund dessen müssen Bewertungsmaßstäbe entwickelt werden, welche die Effektivität und Effizienz eines Internetauftritts abbilden. Data Mining bzw. Web Mining Technologien ermöglichen es das Verhalten der Webseitenbesucher zu verstehen und die Angeboten Inhalte zu optimieren

Prinzipiell verfolgt Web Mining zwei Ziele:

1. Es ermöglicht dem Benutzer/Betreiber eine Bewertung der Effizienz und Effektivität eines Webauftritts. Dabei geht es darum grundlegende Informationen über die Benutzer und über die Benutzung des Internetauftritts, zu erhalten. Dies bezeichnet man als **Web Usage Mining**.
2. Ein weiteres Anwendungsfeld ist das **Web Content Mining**. Die Web Content Mining Methodik beschäftigt sich mit der Analyse der Struktur der im Netz befindlichen Daten. *„Die aus dem Web Content Mining gewonnenen Informationen können z.B. verwendet werden, um zu begreifen, wie man Daten Internet am effizientesten auffinden kann“.*¹

Die verschiedenen Ausprägungen von **Web Usage Mining** und **Web Content Mining** werden in den nachfolgenden Abschnitten näher erläutert

¹ Vgl. <http://www.unet.univie.ac.at/~a9604058/wsinf2/netmin1/content.html>,
abgerufen am 20.Juli 2003

5 Die Datengrundlage

Prinzipiell kann man das Web Mining als Anwendungsfeld des Data Minings be- greifen.

Als Data Mining bezeichnet man die, softwaregestützte Ermittlung bisher unbe- kannter Zusammenhänge, Muster und Trends aus dem Datenbestand sehr großer Datenbestände bzw. des Data Warehouse. Dabei kann man der Benutzer be- stimmte Ziele vorgeben, für die das System angemessene Beurteilungskriterien ableitet und damit die Datenobjekte der Datenbank (en) analysiert. Eine andere Möglichkeit besteht darin, das das System auf eine vage Frage hin eine bestimm- te Menge von Datenobjekten automatisch in Cluster aufteilt, für die bestimmte Zusammenhänge existieren".²

5.1 Datengrundlage für die Analyse

Die Daten der Besucher einer Web Seite werden in **Hitlog** Files gespeichert, in denen der entsprechende Web Server alle Anforderungen in Dateien (**Logfiles**) speichert.

*„Eine **Logfile** ist eine automatisch erstellte Protokolldatei, die alle Anfragen an einen Webserver und deren Ergebnisse aufzeichnet. In einem Logfile werden zu jeder Anfrage die IP-Nummer des zugreifenden Rechners, Datum und Uhrzeit des Zugriffs, die angeforderten Dateien und das Ergebnis der Übertragung gespei- chert. Auch die zuvor besuchte Adresse sowie Betriebssystem und Browser des Benutzers können im Logfile erfasst werden. Logfiles sind damit die Basis für eine Bewertung und Optimierung einer Website in Bezug auf Effizienz und Nutzerver- halten“.*³

Bezüglich der gewonnenen Daten werden nun Datamining Techniken angewendet um aus den gewonnenen Daten klare Aussagen über das Benutzerverhalten, zu erhalten. Dies kann durch Algorithmen erfolgen. Die Algorithmen suchen alle Sequenzen von Seitenabfragen nach wie- derkehrenden Mustern ab. Daraus resultieren zumeist sehr zahlreiche Regeln. Die generierten Regeln beschreiben das Benutzerverhalten mittels so genannter „WENN – DANN“ Regeln. Z.B, wenn Benutzer Produktseite aufruft, dann zu 33 % Wahrscheinlichkeit : Bestellvorgang! Zur Definition solcher Regeln muss die Abfolge der Seitenanfragen des Benutzers und der Zeiträume, in dem sich ein Benutzer auf bestimmten Seiten bewegt identifiziert werden.

6 Web Usage Mining

² Quelle: H.R. Hansen; G.Neumannn Wirtschaftsinformatik 1, 8.A S. 474, 475 Lucius & Lucius Stuttgart 2001

³ Quelle : <http://www.osthus.de/Service/Glossar/Logfile> abgerufen am 20. Juli 2003

Die Web Usage Mining Analyse wird mit dem Ziel durchgeführt bestimmte Muster innerhalb der Seitenanfragen und Aktionen aufzudecken.

Die Kernfragen der Web Usage Mining Analyse lauten:

- Wie interagiert der Benutzer mit dem Internet bzw. wie interagiert der Benutzer mit „einer“ Webseite?
- Welche Inhalte bewegen den Benutzer zu welchem Handeln?

Die Untersuchung des Benutzerverhaltens erfolgt über Bewegungspfadanalysen, welche Zusammenhänge im Navigationsverhalten der Webseitenbesucher aufzeigen. Die Identifikation von Schwachstellen der Websitetopologien ist die Hauptaufgabe dieses Verfahrens d.h. es muss analysiert und protokolliert werden, wie sich der User auf einer Webseite verhält, welche Seiten er aufruft. Die Verweildauer auf einer bestimmten Seite ist ebenfalls von Interesse. Beim Web Usage Mining unterscheidet man zwischen nicht personalisiertem und personalisiertem Web Usage Mining

6.1 Personalisiertes Web Usage Mining

Hierbei ist das Sammeln von Informationen zu den jeweiligen Benutzern interessant. Der Benutzer gibt seine Identität durch eine Authentifikation bekannt z.B. durch einen Login.

„Unter dem Begriff der Authentifikation (engl. Authentication) versteht man die nachweisliche Identifikation eines Benutzers oder eines Kommunikationspartners“⁴

Der Benutzer ist somit namentlich bekannt. Entsprechend der Aktionsmöglichkeiten der Webseite werden die Aktionen des Benutzers in einem Benutzerprofil gespeichert und können anschließend zu einer **Clusteranalyse** genutzt werden um z.B. Personen mit gleichen Interessen zusammenzufassen bzw. zu klassifizieren.

Clusteranalyse : *"deskriptive Methode der Multivarianten Statistik zur Strukturierung der beobachteten Elemente, durch Bildung in sich möglichst homogener untereinander möglichst unähnlicher Gruppen oder Cluster ..."*⁵

Die durch die Clusteranalyse gewonnenen Gruppen sind der Ausgangspunkt für weitere Maßnahmen. Beispielsweise können die bisherigen Webinhalte durch personalisierte Werbeangebote für die einzelnen Gruppen/Cluster optimiert werden.

6.2 4.2 Nicht personalisiertes Web Usage Mining

Beim nicht personalisiertem Web Usage Mining bleibt der Benutzer anonym.

Der Nutzer hinterlässt seine Seitenanfragen dabei im so genannten Hitlog. Demzufolge können die Aktionen, welche ein einzelner Benutzer auslöst, keiner bestimmten Person oder Gruppe zugeordnet werden

7 Web Content Mining

Web Content Mining bzw. Web Structure Mining befasst sich mit der Strukturierung der sich im Netz befindenden Daten. Es wird bei diesem Verfahren versucht eine Segmentierung der im Netz befindlichen Inhalte zu erreichen, indem z.B. statistische Verfahren, Verfahren des maschinellen Lernens, künstliche neuronale Netze angewendet werden. Ziel dieser Verfahren ist es, die sich im Internet befindenden Daten zu klassifizieren und dementsprechend Daten schnell und Effizienten aufzuspüren zu können. *„Kurz gesagt beschäftigt es sich mit der Art*

⁴ Quelle: HR Hansen; G.Neuman Wirtschaftsinformatik I, 8.A, Stuttgart S. 175 Lucius & Lucius Stuttgart 2001

⁵ Quelle: Gabler Wirtschaftslexikon, 15. A S. 631 Gabler Berlin 2000

und Weise, wie Suchmaschinen am effizientesten Informationen über eingegeben Suchbegriffe liefern." ⁶

Jedoch sind diese Verfahren technisch noch nicht ausgereift bzw. schwer auf das World Wide Web übertragbar. Die Bedeutung des Web Content Minings wird jedoch in den nächsten Jahren drastisch zunehmen.

„Durch die rasante Entwicklung des Internets und seine Popularität sind zunehmend mehr Informationen zu einem Thema erhältlich und es werden intelligente Systeme zur Suche, Filterung und Katalogisierung von Information aus Webdokumenten notwendig". ⁷

⁶ Quelle: <http://www.unet.univie.ac.at/~a9604058/wsinf2/netmin1/content.html>, abgerufen am 21.Juli 2003

⁷ Quelle: <http://www.unet.univie.ac.at/~a9604058/wsinf2/netmin1/content.html>, abgerufen am 21.Juli 2003