Web Log Mining als Controllinginstrument der PR

Diplomarbeit
im Studiengang Informationswirtschaft
an der
Fachhochschule Stuttgart
Hochschule der Medien

vorgelegt von: Markus Leibold

Erstprüfer: Prof. Dr. Stefan Grudowski, M.A.
Zweitprüfer: Prof. Dr.-Ing. Peter Lehmann

Bearbeitungszeitraum: 22.08.2003 bis 22.12.2003

Stuttgart, Dezember 2003

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Diplomarbeit selbstständig angefertigt habe. Es wurden nur die in der Arbeit ausdrücklich benannten Quellen und Hilfsmittel benutzt. Wörtlich oder sinngemäß übernommenes Gedankengut habe ich als solches kenntlich gemacht.

Ort, Datum

Unterschrift

Kurzfassung

Gegenstand dieser Diplomarbeit ist Web Log Mining und dessen Einsatz als Controllinginstrument bei Public Relations. Der Ablauf des Web Log Mining wird beschrieben, dabei wird auf Logfile-Kennzahlen und ihre Ermittlung eingegangen. Weiter werden wichtige Data Mining-Methoden erläutert und Aspekte des Datenschutzes werden diskutiert. In Bezug auf Public Relations wird auf spezielle Merkmale der Online-PR und auf Zielgruppen der Online-PR eingegangen. Weiterhin wird ein Modell für PR-Controlling vorgestellt, in dem das Web Log Mining eingeordnet wird. Die Möglichkeiten der Erfolgsmessung von Online-PR werden ebenso betrachtet, wie der Vergleich von Kosten und Nutzen von Web Log Mining. Ein Beispiel für eine Data Mining-Anwendung zur Zielgruppenidentifikation erläutert den praktischen Nutzen von Web Log Mining.

Schlagwörter: Data Mining, Web Log Mining, Public Relations, Logfile, Controlling.

Abstract

Topic of this thesis is web log mining and its application as a controlling instrument in the public relation sector. The description of the procedure of Web Log Mining concentrates specifically on logfile key data and its identification. Important Data Mining methods are being described, followed by a discussion of specific aspects of privacy. Special characteristics referring to online public relations and their target groups are examined and discussed. Furthermore, the position of Web Log Mining will be shown in the context of a public relations-controlling model. Possibilities to measure success of online public relations will be closely looked at, followed by a cost-benefit examination of Web Log Mining. An example for a Data Mining application to identify target groups explains a practical usage of Web Log Mining.

Keywords: Data Mining, Web Log Mining, Public Relations, Logfile, Controlling.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

BDSG	Bundesdatenschutzgesetz
BPN	Backpropagation-Netze
CLF	Common Logfile Format
DNS	Domain Name System
DWH	Data Warehouse
ECLF	Extended Common Logfile Format
ERP	Enterprise Resource Planning
ETL	Extrahieren, Transformieren und Laden (Extraction, Transformation and Loading)
GMT	Greenwich Mean Time
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IVW	Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern e. V.
KDD	Knowledge Discovery in Databases
KNN	Künstliche Neuronale Netze
LSA	Latent Semantic Analysis
OLAP	On-Line Analytical Processing
OS	Operating System (Betriebssystem)
PR	Public Relations (Werbung, Öffentlichkeitsarbeit)
SCG	Scaled Conjugate Gradient
SOM	Self Organizing Maps
TDDSG	Teledienstedatenschutzgesetz

1 Einleitung

Für Unternehmen und Behörden, die im Wettbewerb mit anderen Einrichtungen stehen, aber auch für solche, die auf eine breite Akzeptanz angewiesen sind, ist es unerlässlich, ein positives Bild in der Bevölkerung und in der Geschäftswelt zu haben und es zu erhalten. Für dieses Ziel ist eine gute Öffentlichkeitsarbeit (PR) nicht mehr wegzudenken. Public Relations sind in Zeiten immer aggressiverer Webemaßnahmen keineswegs weniger wichtig für die Unternehmen, sondern haben sogar noch an Bedeutung gewonnen.

> *„Die steigende Bedeutung der Public Relations ergibt sich im Übrigen zuvorderst aus der zunehmenden Resistenz der Öffentlichkeit gegenüber der üblichen Massenwerbung. Berichte in der neutralen Presse über das Unternehmen oder seine Produkte, die durch Öffentlichkeitsarbeit erreicht werden, wirken hingegen wesentlich glaubwürdiger.“* [1]

Dieses wichtige Instrument der Unternehmenskommunikation wird heute zunehmend durch das Medium Internet ergänzt. Um einen möglichst effektiven Einsatz von Online-Public Relations zu gewährleisten, liegt es nahe zu überprüfen, welche Verbreitung die PR über das Internet erzielt hat.

Web Log Mining ist eine Methode, über welche sich Art und Umfang der Zugriffe auf eine Internetpräsenz auswerten lassen. Durch Web Log Mining lassen sich gesammelte Informationen über die Internetseitenbesucher und deren Verhalten auf der Internetpräsenz untersuchen und somit verborgene Zusammenhänge aufdecken. Daher ist Web Log Mining ein Instrument, Public Relations im Internet auf ihre Effektivität und Effizienz hin zu untersuchen. Bei Abweichungen von Soll-Vorgaben bieten die Ergebnisse eine gute Grundlage, angemessene Korrekturmaßnahmen einleiten zu können. Somit wird ein Regelkreis geschaffen, der, ausgehend von dem Ziel, ein positives Image eines Unternehmens in der Öffentlichkeit zu wahren und das Image zu verbessern, über die Kontrolle eingesetzter Online-PR-Maßnahmen bis hin zu aktiven Prozessoptimierungen den optimalen Einsatz von Online-PR gewährleisten kann.

Die vorliegende Arbeit zeigt die Möglichkeiten auf, welche Web Log Mining für das Controlling von Online-PR-Ressourcen bietet. Es werden sowohl die technischen Voraussetzungen und Möglichkeiten, als auch der praktische Nutzen herausgearbeitet. Wird in dieser Arbeit im Zusammenhang mit Public Relations der Bezug auf ein Unternehmen hergestellt, das Public Relations betreibt, ist dies exemplarisch und kann auch für Einrichtungen, Institutionen und Organisationen stehen.

---

[1] explido (2003). URL: http://www.promotionwelt.de/marketingmix_online_pr.htm – Zugriff am 15.10.2003.

Anhand einiger Grafiken wird im Anhang die Auswertung eines Logfiles skizziert, um dem Leser einen Eindruck von Logfile-Auswertungen zu vermitteln.

Wegen der Ausrichtung des Web Log Mining auf Logfiles bleiben die Ausführungen dieser Arbeit bezüglich der Informationsquellen ebenfalls auf Logfiles beschränkt. Andere Quellen für das Auswerten von Website-Nutzung, die beim Integrated Web Log Mining verwendet werden, wie Web-Formulare oder e-Mail, werden in dieser Arbeit nicht näher betrachtet.

1.1 Begriffsklärungen

An dieser Stelle werden einige Fachbegriffe geklärt und voneinander abgegrenzt. Somit soll eine einheitliche Verständnisgrundlage für den Leser geschaffen werden.

Public Relations (PR) ist der englischsprachige Begriff für Öffentlichkeitsarbeit. Nach Kotler et. al. hat Öffentlichkeitsarbeit die Aufgabe, *„[…] gute Beziehungen zu den verschiedenen Partnern des Unternehmens in der internen (Mitarbeiter, Geldgeber) und externen Öffentlichkeit zu erhalten und zu pflegen.“* [2]. Ziel der Öffentlichkeitsarbeit ist also, *„[…] dass über das Unternehmen gesprochen und geschrieben wird – dass das Unternehmen im positiven Sinn nicht in Vergessenheit gerät.* [3] “

Online-PR ist der Begriff für die Öffentlichkeitsarbeit, die über das Medium Internet (zum Beispiel über eine Website oder per e-Mail) umgesetzt wird.[4]

Controlling wird in der Literatur nicht einheitlich definiert, kann aber als Überwachung, Planung und Steuerung von Unternehmensprozessen beschrieben werden. Controlling ist gegenwarts- und zukunftsorientiert, anders als bei einer vergangenheitsorientierten Kontrolle.[5]

Ein *Controllinginstrument* ist eine Methode oder Vorgehensweise, die zur Bewältigung von Controllingaufgaben eingesetzt wird.

Das in dieser Arbeit angesprochene „PR-Controlling“ ist ein Begriff, der so in der Fachwelt nur selten auftaucht. PR-Controlling bezeichnet das Anwenden klassischer Controlling-Prinzipien auf den Bereich der PR.

Web Log Mining ist die Analyse des Verhaltens von Internetnutzern, bei der unter anderem Data Mining-Methoden (siehe Kapitel 2.5, S. 25 ff) auf die von Webservern generierten Logfiles angewendet werden, um Interessen und Verhaltensmuster von Onli-

---

[2] Kotler, P. et al. (2003), S. 946

[3] Kotler, P. et al. (2003), S. 946

[4] vgl. explido (2003). URL: http://www.promotionwelt.de/marketingmix_online_pr.htm – Zugriff am 15.10.2003.

[5] vgl. Schwickert, A. C. / Beiser, A. (1999) URL: http://wi.uni-giessen.de/gi/dl/showfile/Schwickert/1155/Apap_WI_1999_07.pdf – Zugriff am: 19.11.2003. – S. 4 f

ne-Kunden zu ergründen. Beim Web Log Mining bleibt die Datenquelle auf Logfiles beschränkt.[6]

1.2 Aufbau der Arbeit

Im ersten Kapitel wird ein Überblick über die Grundlagen, die Ausgangssituation und die Ergebnisse vermittelt.

Das zweite Kapitel der Arbeit stellt das Web Log Mining näher vor und thematisiert sowohl technologische und methodische als auch rechtliche Gesichtspunkte des Web Log Mining.

Kapitel drei beschreibt anschließend Web Log Mining im Rahmen der Online-PR, wobei dem PR-Controlling besondere Aufmerksamkeit geschenkt wird.

Das vierte Kapitel hat zum Inhalt, wie Web Log Mining in der Praxis als Controllinginstrument der Online-PR zum Einsatz kommen kann. Unter anderem wird hierbei auf die Auswertungsmöglichkeiten und den Nutzen für die PR eingegangen.

Schließlich wird eine Zusammenfassung der angesprochenen Themen gegeben.

[6] vgl. Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-a), S. 7

2 Web Log Mining

2.1 Web Mining

Web Mining ist ein Anwendungsfeld des *Data Mining*, wobei die Datenbasis Nutzungsdaten einer Website sind, die vom Webserver als Logfile aufgezeichnet werden. In manchen Fällen wird die Datenbasis mit weiteren Daten angereichert oder ergänzt. Data Mining selbst ist ein Bestandteil von *Knowledge Discovery in Databases* (KDD), ein Prozess, der in Abbildung 1 dargestellt ist.

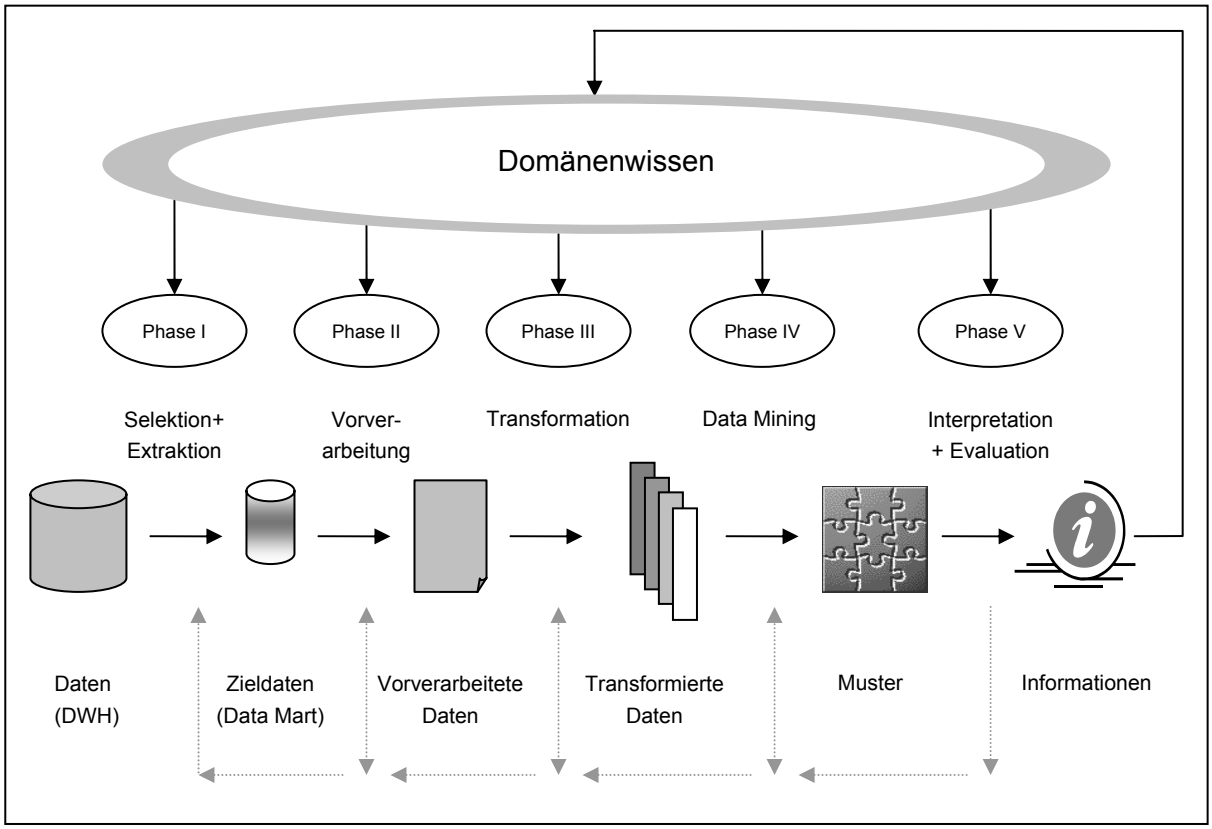

Abbildung 1: Aufbau des KDD-Prozesses[7]

Ausgehend von Daten, die zum Beispiel in einem Data Warehouse (DWH) abgelegt sein können, werden in der ersten Phase des KDD-Prozesses die relevanten Daten selektiert und nach so genannten Data Marts extrahiert. Data Marts sind kleinere Datenbanken, welche diejenigen Daten eines DWH enthalten, die für eine bestimmte Anwendung benötigt werden. Sie sind leichter handhabbar als die komplexe Datenstruk-

[7] vgl. Bensberg, F. (2001), S. 72 [aufbauend auf Fayyad, U. M. et al. (1996), S. 10] und vgl. Fayyad, U. M. / Piatetsky-Shapiro, G. / Smyth P. (1996), – URL: http://www.aaai.org/Library/Magazine/Vol17/17-03/Papers/AIMag17-03-002.pdf – Zugriff am 29.11.2003. – S: 41

tur eines DWH.[8] Entscheidungsgrundlage für die Selektion der Daten ist stets ein konkret verfolgtes Ziel, das mit dem Prozess erreicht werden soll, zum Beispiel Kundengruppen zu identifizieren. Daran schließt sich in Phase zwei eine Vorverarbeitung an. In dieser Phase werden mögliche Fehlerquellen beseitigt, welche die beabsichtigte Untersuchung verfälschen könnten. In der dritten Phase werden die Daten transformiert. Das ist wichtig, um die Daten in die gewünschte Struktur zu bringen, die für die beabsichtigte Data Mining-Methode vorliegen muss. Phase vier ist der Abschnitt, in dem mit Data Mining Mustererkennung betrieben wird. Nahezu jede Data Mining-Methode erfordert eine Vorbereitung, die hierbei eingerechnet werden muss. In Phase fünf werden gefundene Muster evaluiert und interpretiert, was zu einem Informationsgewinn führt, der dem Domänenwissen (=relevantes Fachwissen) zugute kommt. Da das Domänenwissen mit jeder Interpretation von Ergebnissen zunimmt, ist der KDD-Prozess rekursiv und wird deshalb auch als „dynamisch“ bezeichnet.[9]

Web Log Mining ist ein abgrenzbarer Bereich des Web Mining und zeichnet sich dadurch aus, dass als primäre Datenquelle das Logfile eines Webservers verwendet wird. Bedingt durch die Datenquelle „Logfiles“ wird beim Web Log Mining vor allem untersucht, wie das Angebot eines Webservers genutzt wird. Im Gegensatz zum „Integrated Web Usage Mining“ wird beim „Web Log Mining“ auf zusätzliche Datenquellen, die direkt Informationen über den Besucher beinhalten, verzichtet, allein das Logfile wird untersucht.[10]

Eine Übersicht über die Disziplinen des Web Mining und die Einordnung des Web Log Mining gibt folgende Darstellung:

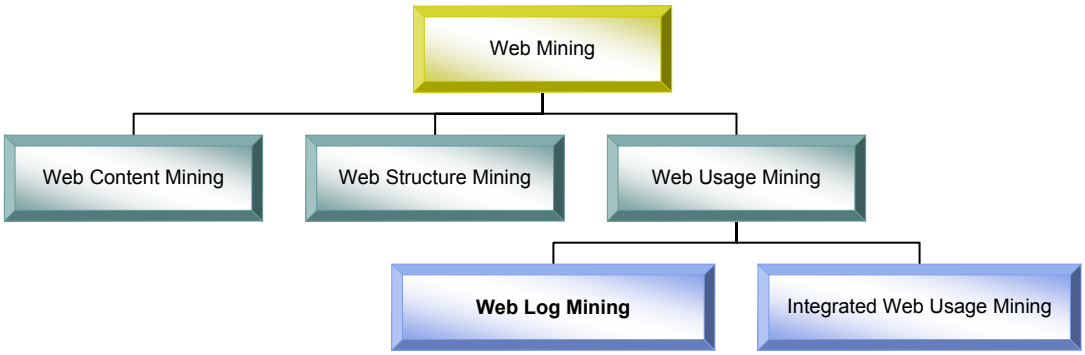

Abbildung 2: Einordnung des Web Log Mining[11]

Das Web Mining kennt neben Web Usage Mining noch zwei weitere Disziplinen: Das Web Content Mining hat die Suche nach Informationsinhalten zum Gegenstand und

[8] vgl. Brosius, G. (2001), S. 33
[9] Bensberg, F. (2001), S. 72
[10] vgl. Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-a), S. 7
[11] vgl. Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-a), S. 7 f

das Web Structure Mining den Aufbau und die Verlinkung von Ressourcen. Bei diesen anderen Web Mining-Disziplinen steht im Vordergrund, ursprünglich unübersichtliche Informationsmengen im Internet zu erfassen und so leichter zugänglich zu machen.

Die Produktvielfalt an Web Mining-Software ist mittlerweile groß. Einige Websites zum Thema KDD und Data Mining haben Übersichten zu Web Mining-Software erstellt, die einen Einstieg in das Softwarespektrum erleichtern.[12] Eine sehr gute Übersicht bietet die Site www.kdnuggets.com .[13]

2.2 Logfiles

Die Datenquelle einer Web Log Mining–Untersuchung ist ein Logfile eines Webservers. Jeder Webserver erstellt während des Betriebs Logfiles zu unterschiedlichen Zwecken, die sich im Format[14] unterscheiden. Das von nahezu allen Webservern generierte Logfile-Format ist das so genannte „Common Logfile Format“ (CLF). Häufig wird es durch zusätzliche Informationen erweitert und dann als „Extended Common Logfile Format“ (ECLF) bezeichnet. Tabelle 1 zeigt die Datenfelder, die durch diese Formate erhoben werden.

Tabelle 1: Common Logfile Format und Extended Common Logfile Format[15]

		Common Logfile Format	Extended Common Logfile Format
Remotehost	IP-Adresse des zugreifenden Servers		
Ident	Identifikation (falls vorhanden, sonst “-“)		
Authuser	Benutzername (nur bei Passwortabfragen, sonst “-“)		
Date	Datum und Uhrzeit des Zugriffs		
Timezone	Abweichung von der Greenwich Mean Time (GMT) in Stunden		
Request	Methode, Dokument und verwendetes Protokoll des Zugriffs		
Status	Antwortstatus als Code[16]		
Bytes	Gesamtzahl der übertragenen Bytes		
Referrer	URL der Seite, die den Link zur angefragten Ressource enthielt		
Agent	Systemumgebung (Browser, OS[17]) des anfragenden Browsers		

[12] vgl. Bolz, C. (2001), URL: http://www.bolz.org/Vergleich_Web_Mining_Software.PDF – Zugriff am 06.12.2003. – S. 6

[13] vgl. KDnuggets (2003): URL: http://www.kdnuggets.com/software/web.html – Zugriff am 06.12.2003.

[14] Dies sind zum Beispiel Zugriffsprotokolle, Fehlerprotokolle oder Anwendungsprotokolle.

[15] vgl. Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-a), S. 10

[16] Siehe auch Anhang C: HTTP Status Codes auf Seite 58

[17] „OS“ steht für Operating System (Betriebssystem).

In der Praxis hängt es aber von der Konfiguration des Webservers ab, welche Daten im Logfile protokolliert werden. Logfiledateien nehmen nicht selten innerhalb kurzer Zeit umfangreiche Dateigrößen an. Deshalb gibt es auch Webserver, die ein „abgespecktes“ Logfileformat haben, um den erforderlichen Speicherplatzbedarf in Grenzen zu halten. Werden für regelmäßige Auswertungen der Logfiledaten lückenlose und detaillierte Datenbestände benötigt, muss das Logfile in regelmäßigen Abständen (in der Regel täglich oder wöchentlich) archiviert werden.[18]

2.3 Der Web Log Mining Prozess

Innerhalb des Web Log Mining sind einige Schritte notwendig, um aussagekräftige Ergebnisse zu erhalten. Diese bauen größtenteils aufeinander auf, manche Schritte können nur ausgeführt werden, wenn vorhergehende abgeschlossen sind. Web Log Mining ist damit ein Prozess, der von einigen Autoren analog des KDD gestaltet wurde.[19]

Da bei jedem Zugriff auf eine Web-Ressource die in Tabelle 1 angeführten Informationen im Logfile des jeweiligen Webservers protokolliert werden, können diese Informationen später für Auswertungen herangezogen werden. Der Inhalt des Logfiles ist Datenbasis und Grundlage jeder weiteren Untersuchung beim Web Log Mining.

Die Logfiledaten beinhalten Kennzahlen, die in Kapitel 2.4 ab Seite 19 näher erläutert werden. Die direkt bestimmbaren, da aus einem einzigen Feld im Logfile ableitbaren Kennzahlen, zählen hierbei zu „einfachen Auswertungen“, Kennzahlen, die sich aus anderen zusammensetzen, werden unter der Überschrift „fortgeschrittene Auswertungen“ behandelt.

2.3.1 Ablauf des Web Log Mining

Der in folgender Abbildung dargestellte Ablauf ist analog zu den Phasen des KDD-Prozesses aufgebaut. Im Web Log Mining kommen aber nicht allein Data Mining-Techniken zum Einsatz, sondern auch so genannte Logfile-Analysen, die sich mit der Ermittlung von Logfile-Kennzahlen beschäftigen.

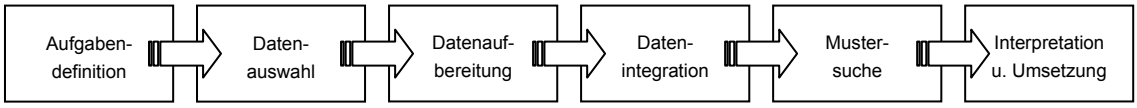

Abbildung 3: Ablauf der Web Log Mining Analyse[20]

[18] vgl. Bürlimann, M. (1999), S. 225

[19] vgl. Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-a), S. 8
sowie Bensberg, F. (2001), S. 133

[20] Aus: Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-a), S. 8 f

Bei der *Aufgabendefinition* wird festgelegt, welche Ziele in einer konkreten Untersuchung verfolgt werden sollen. Die definierten Ziele bestimmen entscheidend, wie in den folgenden Schritten weiter vorgegangen wird. Beispiele für Abhängigkeiten zum späteren Web Log Mining-Prozess sind:

- Start- und Endpunkt der Logfile-Erhebung (beziehungsweise zu untersuchende Zeitspanne)
- Art der Datenaufbereitung (ETL-Maßnahmen[21])
- Art und Umfang der Mustersuche: Techniken, Aufwand, Dauer, etc.

Die *Datenauswahl* ist der Schritt, bei dem die zur Untersuchung notwendigen Daten erhoben beziehungsweise herangezogen werden. Sollen vergangene Zeiträume untersucht werden, sind die bestehenden Datenstrukturen in der Regel fixiert. Wenn die Daten erst noch erhoben werden sollen, hat dies den Vorteil, dass das Format des Logfiles angepasst werden kann. Es können Datenfelder hinzugefügt werden und die Syntax der Datenfelder kann beeinflusst werden.

Der Prozessschritt der *Datenaufbereitung* ist eng verknüpft mit der *Datenintegration* und aus mehreren Gründen besonders wichtig. Vorrangig werden im Rahmen der gegebenen Möglichkeiten unerwünschte, verfälschende Einflüsse aus dem Datenbestand herausgefiltert und, wenn nötig, eliminiert. Zum Beispiel verfälschen Zugriffe des Administrators einer Website die Aussagekraft des Logfiles und sollten im Logfile gelöscht werden. Nicht alle verfälschenden Einflüsse können jedoch eliminiert oder behoben werden, es ist aber wichtig einschätzen zu können, welche Aussagekraft ein Logfile hat, um Ergebnisse besser bewerten zu können. In Tabelle 2 (auf Seite 19) werden verfälschende Einflüsse in Logfiles zusammengefasst.

Neben der Bereinigung muss ein Logfile in den meisten Fällen auch einer Umformatierung unterzogen werden. Um Logfile-Daten in ein Analyseprogramm einlesen zu können, müssen sie im entsprechend geeigneten Format vorliegen. Für diese Aufgabe eignet sich die aus dem Data Warehousing bekannte Technologie des „Extrahieren, Transformieren und Laden“ (ETL). Beispielsweise kann es sein, dass Zeilen im Logfile zu lang sind (bedingt durch sehr lange Request- oder Referrer-Einträge) um eingelesen werden zu können. Ohne Datenaufbereitung ist kein sinnvolles Verarbeiten eines Logfiles möglich.

> *„Data Mining Projekte setzen nicht nur Datenvielfalt, sondern auch qualitativ einwandfreie Daten voraus. Sind die Daten nicht redundanzfrei und konsistent, so wird jedes Data Mining Projekt scheitern.“* [22]

Im Prozessschritt der *Mustersuche* findet das eigentliche „Mining“ (engl.: graben) nach wertvollen Informationen statt. Es wird nach interessanten und noch unbekannten Häu-

[21] ETL steht für „Extrahieren, Transformieren und Laden“
[22] Rapp, R. / Guth, S. (2003), S. 175

figkeiten, Anhängigkeiten, Mustern und weiteren Kenngrößen gesucht. Kapitel 2.5 stellt ab Seite 25 wichtige Techniken detailliert vor.

Der letzte Schritt im Web Log Mining, der Prozess der *Interpretation und Umsetzung* setzt die gewonnenen Erkenntnisse der Mustersuche dazu ein, den Webauftritt selbst und damit auch seine Wirkung zu verbessern. Durch diesen Prozessschritt tritt die Controllingeigenschaft in den Vordergrund.

2.3.2 Negative Einflussfaktoren bei der Datenerhebung

Es gibt eine Vielzahl denkbarer Einflüsse, angefangen von Suchmaschinen-Robots bis hin zu Anonymisier-Diensten, welche die Daten im Logfile verfälschen. Die Kenntnis über diese störenden Faktoren versetzt den Webmaster in die Lage, sich darauf einzustellen und das Logfile vor einer näheren Betrachtung zu bereinigen. Im Folgenden wird eine Übersicht der wichtigsten Faktoren gegeben, welche die Datenerhebung beeinflussen. Verfälschende Logfile-Einflüsse und mögliche Gegenmaßnahmen werden in Tabelle 2 auf Seite 19 zusammenfassend dargestellt.

Das Zwischenspeichern und Vorhalten von Teilen oder vollständigen Dateien einer Website wird als *Caching* bezeichnet. Als Cache kann der *Browsercache* fungieren, wenn dieselbe Ressource erst kürzlich genutzt wurde, und auch ein Proxyserver, der in der Regel von Providern und auch Firmennetzwerken eingesetzt wird. Proxyserver (im Folgenden auch als Proxy, pl:: Proxies, bezeichnet), werden in Netzwerken eingesetzt und speichern von Benutzern angefragte Dateien für eine bestimmte Zeit, um sie einem Nutzer, der die Dateien nochmals anfragt, dann schneller zur Verfügung stellen zu können. Eine besondere Form des Caching stellen so genannte Mirror-Sites dar. Häufig gefragte Ressourcen werden dabei von einem oder mehreren, von der eigentlichen Website unabhängigen Server bereitgestellt. Zugriffe auf einen Mirror werden im Logfile des Webservers, der die Website beheimatet, nicht registriert.

Neben der schon angesprochenen Cache-Funktion, welche *Proxyserver* einnehmen, entsteht durch Proxies ein weiteres Problem, und zwar werden Anfragen, die über den Umweg eines Proxys gestellt werden, mit der (externen) IP-Adresse des Proxys, und nicht mit der IP-Adresse des eigentlich anfragenden Clients im Logfile festgehalten.

Eine spezielle Ausprägung von Proxies sind Anonymisier-Dienste im Internet. So genannte „Anonymizer“ sind spezielle Dienste im Internet, die sich der Proxy-Technologie bedienen. Sie bieten interessierten Nutzern an, einen frei zugänglichen Proxyserver für den Internetzugriff zu verwenden. Diese teils kommerziellen Dienste bieten im Gegensatz zu einem üblichen Proxy nicht nur die Verschleierung der IP-Adresse, sondern verfälschen und verbergen gezielt weitere Informationen wie zum Beispiel den Referrer und den Zeitstempel und machen so eine Erfassung im Logfile unmöglich.[23]

[23] vgl. Marschall, N. (2002), S. 82

Das *dynamische* Zuweisen von *IP-Adressen* sowohl durch Internetprovider an ihre Kunden als auch in häufigen Fällen in Firmen macht es einem außen Stehenden unmöglich, eine bestimmte IP-Adresse einer einzigen Person zuzuordnen. Die Unkenntnis darüber, welche IP-Adressen fest vergeben werden, und welche dynamisch zugewiesen sind, weitet diesen problematischen Effekt der dynamischen Vergabe auf die fest vergebenen IP-Adressen aus. Dennoch ist über Anfragen an DNS-Server (DNS: Domain Name System) über spezielle Dienste im Internet ein Eingrenzen und Zuordnen von IP-Adressen zumindest auf Firmen und Provider möglich. So kann zum Beispiel mit Sicherheit eine Aussage darüber getroffen werden, über welchen Provider oder von welchem Unternehmen ein Logfile-Eintrag stammt.

Computer, die von mehreren Personen für die Internetnutzung verwendet werden, verfälschen etwaige Logfileauswertungen dadurch, dass unter Umständen angenommen wird, ein und dieselbe Person tätigt die von dem Computer ausgehenden Zugriffe. Familien-PCs und Internetcafé-Computer sind klassische Vertreter solcher *Multi-User-Computer*.

Suchmaschinen und andere Server, die Web Content Mining betreiben, setzen so genannte *Robots* oder *Spider* ein, die automatisch im Internet nach Inhalten suchen. Dabei entstehen bei den besuchten Internetpräsenzen Logfile-Einträge, die nicht durch einen Aufruf eines Website-Besuchers entstanden sind und damit für eine Web Log-Analyse irrelevant sind. In aller Regel werden solche Einträge anhand der Host-ID entfernt und gegebenenfalls einer gesonderten Auswertung zugeführt.

2.3.3 Website-Architektur

Die Website-Architektur weist – vor allem bei größeren Web-Auftritten – sowohl hardware- als auch softwareseitig an einigen Stellen Besonderheiten auf, die von der standardmäßigen Logfile-Erstellung abweichen. Bei grossen Internetpräsenzen und auch bei Online-Shops werden beispielsweise mehrere Server dazu eingesetzt, Daten für die Website zur Verfügung zu stellen.

Durch geeignete Verfahren kann sichergestellt werden, dass die Logfiledaten möglichst wenig durch externe Einflüsse verfälscht werden und somit die größte erzielbare Aussagekraft und Datenqualität erhalten. Es gibt verschiedene Maßnahmen hierzu, die nun vorgestellt werden.

Cookies dienen hauptsächlich dem Wiedererkennen von Benutzern (Usern), die eine Website zuvor schon einmal besucht haben. Sie werden aber auch dazu eingesetzt, zusammenhängende Besuche auf einer Website zu identifizieren. Das erleichtert auch das Identifizieren von Sessions (auch Sitzungen oder zusammenhängende Besuche genannt). Generell werden zwei Arten von Cookies unterschieden: persistente und transiente Cookies. Persistente Cookies sind mittel- oder längerfristig auf einem Computer gespeichert. Transiente Cookies befinden sich nur für die Dauer einer Sitzung im Arbeitsspeicher des Computers und werden nicht dauerhaft gespeichert.

Netzwerk- oder *Server-Monitore* kommen bei größeren Internetpräsenzen zum Einsatz. Sie sind eigene Server im Netzwerk des Website-Servers und schreiben Logfiledaten in Echtzeit in eine Datenbank. Der Server-Monitor erledigt das Aufzeichnen von Logfiledaten für einen Webserver, Netzwerkmonitore können den Datenverkehr von mehreren Webservern gleichzeitig erfassen. Eine spezielle Variante des Netzwerkmonitors stellt der *Reverse-Proxy-Monitor* dar. Er ist zwischen den Webservern und der Internetanbindung angesiedelt und protokolliert Logdaten.[24]

Das *Pixel-Verfahren* ist ein Messverfahren des IVW (Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern e. V.)[25] zur Bewertung von Werbeträgern im Internet.

> *„Dies wird dadurch versucht, dass alle Zugriffe auf potenziell werbeführende Seiten gezählt werden, um einen objektiven Vergleich zwischen unterschiedlichen Web-Sites zu ermöglichen.“* [26]

Realisiert wird dies durch eine kleine Grafik, die wegen ihrer minimalen Ausmaße auf der Webseite nicht zur Anzeige kommt, aber beim Seitenaufruf auf einem Server des IVW einen Logfile-Eintrag erzeugt, der einen Rückschluss auf die aufgerufene Seite erlaubt sowie eine Klassifizierung des Seiteninhalts beinhaltet. Durch Übergabe von Parametern sollen Proxies umgangen werden, damit jeder Aufruf einer Website, die das Pixel enthält, zu einem Eintrag im Logfile führt.[27]

Eine weitere Technik zur User- und Sessionidentifizierung sind *Session-ID*'s. Sie kommen bei der dynamischen Seitenprogrammierung zum Einsatz und verlängern die URL, die vom Benutzer aufgerufen wird, um eine eindeutige Zeichen- oder Ziffernkombination. Über den Referrer kann so ganz einfach eine Sitzung rekonstruiert werden. Außerdem hat dieses Verfahren den großen Vorteil, dass nie eine schon zuvor von einem anderen Benutzer aufgerufene Seite vom Proxy zur Verfügung gestellt wird, weil jede URL anders ist. Das Wiedererkennen von Usern aus vorangegangen Sitzungen ist allerdings nicht möglich.[28]

Eine weitere Möglichkeit, Benutzer zu erkennen und ihre Bewegungen auf der Website zu verfolgen, ist das als *Subskriptions-Funktion* bezeichnete Anmelden der Benutzer auf der Website. Hierdurch kann in Kombination mit Session-ID's und Cookies eine maximale Transparenz der Useraktionen erreicht werden. Manche Websites knüpfen die Anmeldung an der Website an einen Mehrwert für den Nutzer, indem bestimmte Funktionalitäten des Internetauftritts nur nach Anmeldung verfügbar sind.

---

[24] vgl. Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-a), S. 11 ff
[25] IVW (2003), URL: http://www.ivwonline.de, Zugriff am 11.11.2003
[26] Säuberlich, F. (2002), S. 113
[27] vgl. Säuberlich, F. (2002), S. 113 f
[28] vgl. Säuberlich, F. (2002), S. 111

Tabelle 2: Verfälschende Logfile-Einflüsse und mögliche Gegenmaßnahmen[29]

Einflussfaktor	Auswirkung	Mögliche Abhilfe
Proxy	IP-Verschleierung Verringerung der Logeinträge	Session-ID, Pixelverfahren
Anonymisier-Dienste	Unvollständige oder gefälschte Logeinträge	Subskriptions-Funktion für eine Userfeststellung
Dynamische IP-Adressen	Gleiche IP-Adressen können von unterschiedlichen Computern stammen, Computer wechseln ihre IP-Adressen	Cookies, Subskriptions-Funktion
Browser-Cache	Verringerung der Logeinträge	Session-ID, Pixel-Verfahren
Multi-User-Computer	Verschiedene User an einem Computer können Cookie-Auswertung erschweren	Subskriptions-Funktion
Robots und Spider	Erhöhung der Logeinträge	Entfernen der Einträge aus dem Logfile

Auch bei Kenntnis aller möglichen Einflüsse, die eine exakte Erhebung behindern könnten, sollte der Einsatz von Gegenmaßnahmen immer in einem wirtschaftlich vertretbaren Maß stattfinden. Oft reicht die Kenntnis über potentielle Störfaktoren aus, um am Ende einer Logfileanalyse brauchbare Ergebnisse zu erzielen, ohne kostenintensive Bereinigungen an den Logfiledaten durchgeführt zu haben.

2.4 Logfile-Kennzahlen

Es gibt einige typische Kennzahlen, die in der Praxis eine weite Verbreitung gefunden haben, um die Attraktivität von Websites zu messen.[30] Diese werden im Folgenden in zwei Gruppen eingeteilt: In „einfache Auswertungen“, die allein aus der Betrachtung eines Feldes (zum Beispiel dem Referrer) heraus ein Ergebnis liefern können, und in „fortgeschrittene Auswertungen“, bei denen eine Kombination von Datenfeldern zu einem Ergebnis führt.

Komplexe Sachverhalte und umfangreiche Datenmengen lassen sich durch die Verwendung von Modellen anschaulich abbilden. Dabei werden vorhandene Datenmengen (zum Beispiel die eines Logfiles) in unterschiedlichen Bereichen zusammengefasst und entweder anhand mathematischer Formeln oder über sachliche und logische Zu-

[29] vgl. Säuberlich, F. (2002), S. 114
und vgl. Marschall, N. (2002), S. 51 ff

[30] vgl. Marschall, N. (2002), S. 47 f

sammenhänge miteinander verknüpft.[31] Mathematische Kennzahlensysteme sind hierarchisch aufgebaut, eignen sich aber nach Schwickert / Wendt nicht gut für die Beschreibung der Website-Nutzung:

> *„Zur Beschreibung der Web-Site-Aktivität ist ein hierarchisches System nicht geeignet, da die Messgrößen der Web-Site-Nutzung nur teilweise in einem mathematischen Zusammenhang stehen.“* [32]

Deshalb erfolgt die Einteilung der Logfile-Kennzahlen in sachgegebenen Zusammenhängen. Absolute Zahlen und Verhältnisse, die mit einem zeitlichen Bezug kombiniert werden können, werden ausgewertet. Diejenigen Kennzahlen, die für eine Untersuchung relevante Informationen liefern können, werden betrachtet und gegebenenfalls kombiniert. Der Detaillierungsgrad kann sehr unterschiedlich ausfallen, je nachdem, welchen Hintergrund die Untersuchung hat.[33]

2.4.1 Einfache Auswertungen

Schon mit einfachen Werkzeugen lassen sich aus Logfiles interessante Informationen gewinnen. Tabelle 3 auf Seite 22 zeigt Kennzahlen, die mit Logfile-Analyseprogrammen festgestellt werden können und jeweils auf einem Feld eines Logfiles beruhen. Die einzelnen Felder werden nachfolgend erläutert.

Aus dem Feld „Date / Time“ lässt sich (unter Berücksichtigung der Zeitzone) leicht feststellen, zu welchen Tageszeiten die meisten Zugriffe stattfinden. Nicht selten macht es auch einen Unterschied, welcher Wochentag betrachtet wird. Internetpräsenzen, die private Rezipienten oder Zielgruppen haben, müssen sich zu späten Tageszeiten und an Wochenend- und Feiertagen auf die meisten Zugriffe einstellen. Hingegen sind bei Business-orientierten Internetpräsenzen häufige Zugriffe an Werktagen zu erwarten.

Das Feld „Request“ im Common Logfile Format (CLF) gibt an, welche Ressource angefordert wurde. Ein statistisches Ranking der am *häufigsten angefragten Seiten* lässt Rückschlüsse darauf zu, welche Bereiche einer Website besonders genutzt werden. Es gibt aber auch Seiten, die ungewollt häufig in die Statistik mit einfließen, zum Beispiel Einstiegsseiten. Sie werden häufig als Startseite im Browser definiert und werden bei jedem Browserstart aufgerufen, egal ob der jeweilige Anwender diese dann auch nutzt.

---

[31] vgl. Schwickert, A. C. / Wendt, P. (2000),
URL: http://wi.uni-giessen.de/gi/dl/showfile/Schwickert/1168/Apap_WI_2000_08.pdf –
Zugriff am: 19.11.2003. – S. 3

[32] Schwickert, A. C. / Wendt, P. (2000),
URL: http://wi.uni-giessen.de/gi/dl/showfile/Schwickert/1168/Apap_WI_2000_08.pdf –
Zugriff am: 19.11.2003. – S. 3

[33] vgl. Schwickert, A. C. / Wendt, P. (2000),
URL: http://wi.uni-giessen.de/gi/dl/showfile/Schwickert/1168/Apap_WI_2000_08.pdf –
Zugriff am: 19.11.2003. – S. 4

In der HTTP 1.1 Protokollspezifikation gibt es 18 unterschiedliche Clientfehlermeldungen und 6 weitere Serverfehlermeldungen.[34] Die Häufigkeit der dabei festgestellten *Übertragungsfehler* lässt Rückschlüsse darauf zu, wie gut die Website funktioniert.

Die Gesamtmenge der übertragenen Bytes wird als *Traffic* bezeichnet und ist eine wichtige Information über die Auslastung der Serveranbindung der Internetpräsenz. Vergleicht man den entstandenen Traffic mit der wirklichen Dateigröße, kann Aufschluss darüber erlangt werden, ob Dateiübertragungen (aus unterschiedlichen Gründen) vorzeitig abgebrochen wurden.[35]

Das Feld „Referrer" gibt an, welche Seite im Browser vor dem Aufruf der entsprechenden Ressource angezeigt wurde. Somit kann über dieses Feld in Erfahrung gebracht werden, ob es Portale oder andere Websites gibt, welche einen Link auf die Seite eingerichtet haben und so zum Einstieg genutzt werden. Oft ist durch den Referrer auch eine Eingrenzung der Interessen des Besuchers möglich. Eine Aufstellung der *häufigsten verweisenden Seiten* ist also hilfreich bei der Kategorisierung angesprochener Zielgruppen.

Ebenfalls im Feld „Referrer" sind meist Einträge zu finden, die vom vorherigen Besuch einer Suchmaschine stammen. Durch Extrahieren der in der Suchmaschinen-URL enthaltenen *Suchbegriffe* wird auch hier eine Kategorisierung ermöglicht, welche die Interessen der Websitebesucher abbilden kann.

Ein Ranking der *meistverwendeten Browser* über das Feld „Agent" kann dazu dienen, Maßnahmen zu ergreifen, welche die Funktionalität der Website optimieren. Da jeder Browser unterschiedliche Befehlssätze unterstützt[36], kann neben den protokollierten Statuscodes diese Information für den Webmaster sehr hilfreich sein, um die Gestaltung der Website entsprechend den Vorraussetzungen der Benutzer anzupassen.

Das Zählen der Zugriffe auf den Server (*Hits*) ist eine häufig verwendete, da relativ simple Kennzahl bei Logfile-Analysen und der erste Schritt, das Verhalten der Website-Nutzer zu analysieren. Das Zählen der Hits alleine hat allerdings noch wenig Aussagekraft. Zum einen enthält ein unbereinigtes Logfile viele Zugriffe, die für eine Analyse keine Rolle spielen (Administratorzugriffe, Suchmaschinen-Robots und –Spider, etc.) und andererseits sagt die Zahl der Hits nur dann etwas aus, wenn bekannt ist, wie viele Elemente eine einzelne Webseite hat. Ruft ein Besucher beispielsweise eine Webseite mit drei Grafiken, zwei Frames und einem Hauptframe auf, so entstehen im Logfile sechs Hits.

---

[34] vgl. Anhang C, Seite 58, Tabelle 6: HTTP Status Codes nach HTTP 1.1, / vgl. W3C (1999), http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html#sec10 Zugriff am 12.10.2003

[35] vgl. Marschall, N. (2002), S. 49

[36] vgl. Niederst, J. (2000), S. 6 ff

Tabelle 3: Informationsgehalt einer einfachen Logfileanalyse[37]

Enthaltene Informationen	Betrachtetes Feld
Wochentage und Uhrzeiten mit den meisten Zugriffen	Date / Time (inkl. „Timezone“)
Häufigste Seitenanforderungen	Request
Häufigkeit der verschiedenen Übertragungsfehler	Status
Traffic (Gesamtmenge der übertragenen Bytes)	Bytes (ggf. Soll-Ist-Betrachtung)
Häufigste verweisende Seiten	Referrer
In Suchmaschinen verwendete Suchbegriffe	Referrer
Meistverwendete Browser	Agent
Anzahl der Zugriffe auf den Server (=Hits)	Zeilenanzahl (rohes Logfile)

2.4.2 Fortgeschrittene Auswertungen

Andere Auswertungen und Informationen sind in ihrer Gewinnung oft mit höherem Aufwand verbunden. Gründe hierfür sind, dass einerseits das Logfile zunächst intensiv bereinigt werden muss, andererseits verschiedene Felder kombiniert betrachtet werden müssen und schließlich noch manche Informationen aufeinander aufbauen. So zum Beispiel der Zusammenhang zwischen Hits, Pageviews, Sessions und Usern.

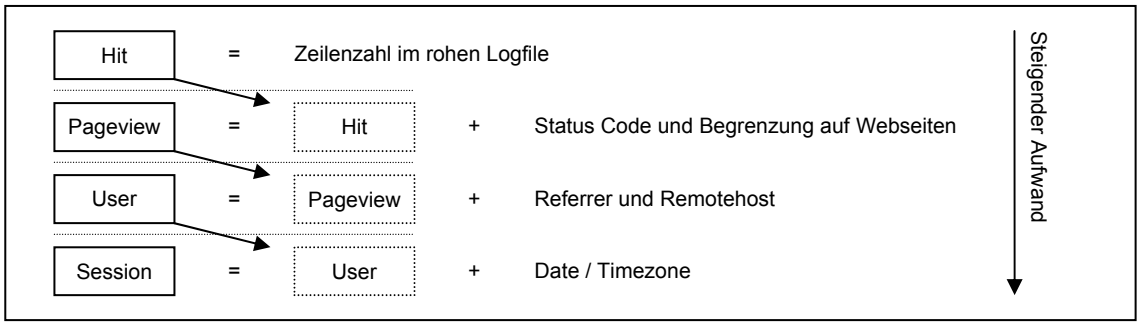

Abbildung 4: Zusammenhang zwischen Hit, Pageview, Session und User[38]

Das Erfassen dieser Kennzahlen bedeutet (in dieser Reihenfolge) einen steigenden Aufwand. Das Zählen der *Pageviews* ist die am meisten verbreitete Methode, um Websites zu vergleichen.[39] Die drei Begriffe „Pageimpression“, „Pageview“ und „View“ werden oft synonym verwendet.[40] Bei der Zählung werden ausschließlich erfolgreich übertragene HTML-Seiten berücksichtigt. Das erfordert die Betrachtung der Felder

[37] vgl. Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-a), S. 19
[38] eigene Darstellung.
[39] vgl. Säuberlich, F. (2002), S. 110
[40] vgl. Marschall, N. (2002), S. 47f

Request und Status Code. In die Zählung gehen Seiten, die durch Proxies, Caches oder Mirrors bereitgestellt wurden nicht ein.

Als *Visit* wird ein zusammenhängender Besuch einer Website bezeichnet.[41] Wichtig ist hierbei, dass der Zugriff, der zur Zählung beiträgt, von außerhalb der Website erfolgt. Über das Feld Referrer kann der Einstieg auf die Website erkannt werden. Da ein Internetnutzer durchaus mehrmals täglich eine Website besuchen kann, versucht man, eine weitere Kennzahl, den *User,* über das Feld „Remotehost" zu erfassen. Das erweist sich in der Praxis aber oft als schwierig, da Internetprovider IP-Adressen dynamisch vergeben. Eindeutige Zuordnungen von Usern können nur über eine Authentifizierung oder bedingt über Cookies erreicht werden, da dort die Benutzer sich per Loginname anmelden, oder ein Cookie einen Benutzer wieder erkennen kann.

Sessions sind in der Praxis sehr aufwändig in ihrer Bestimmung, nicht zuletzt deshalb, weil es unterschiedliche Standards für sie gibt.[42] Als Session wird eine aufeinander folgende Reihe von Page Views durch einen User bezeichnet. Im Vergleich zum Page View und zum Visit ist dabei der zeitliche Faktor von besonderem Interesse. Häufig wird eine Session zur nächsten durch einen Timeout von 30 Minuten abgegrenzt. Es gibt aber auch Vorgehensweisen, die eine tägliche Abgrenzung vornehmen[43] (diese Variante ist auch leichter programmierbar). Je nachdem, welcher Timeout herangezogen wird, kommt eine unterschiedliche Sessionzahl zustande.

Der Begriff *Adimpression* bezeichnet das Erscheinen einer Werbegrafik (auch Banner genannt) auf einer meist externen Website. Setzt sich eine Werbegrafik aus mehreren Dateien zusammen, werden in jedem Fall nur vollständig gezeigte Banner gezählt. Adimpressions werden in ganzen Zahlen gezählt und auf eine Zeiteinheit bezogen (zum Beispiel pro Monat).[44]

Der *Adclick* ist die simple Zählung der Klicks auf einen Werbebanner innerhalb einer Zeiteinheit (zum Beispiel Monat). Es wird auch zwischen den Webseiten unterschieden, auf denen der Banner eingebunden ist. Eine Unterscheidung wird zum Beispiel durch eine dynamische Seitenprogrammierung möglich, bei der, abhängig vom jeweiligen Banner, ein unterschiedlicher Parameter beim Seitenaufruf übergeben werden kann, der durch den Adclick entsteht.

Die *Clickthrough-Rate* setzt die Kennzahlen Adimpression und Adclick zueinander in Beziehung. Hat ein Banner 1.000 Adimpressions pro Monat und 30 Adclicks pro Monat, so weist die Clickthrough-Rate einen Wert von 3% (= 30 ÷ 1.000) auf.

Eine ungefähre geographische Lokalisierung der Websitebesucher ist ebenfalls anhand des Remotehost möglich. Außerdem ist das Feststellen der Nationalitäten der Besucher anhand des Landes, in dem die verwendete IP-Adresse registriert ist, mög-

---

[41] vgl. Marschall, N. (2002), S. 48
[42] vgl. Marschall, N. (2002), S. 48
[43] vgl. Marschall, N. (2002), S. 48
[44] vgl. Bürlimann, M. (1999), S. 207

lich. Es gibt verschiedene Anbieter von Software, die eine solche Analyse unterstützen, zum Beispiel NetGeo Inc., San Jose, U.S.A..[45] Eine beispielhafte Auswertung findet sich in Anhang B auf Seite 57.

Tabelle 4: Zusammenfassung: Informationsgehalt fortgeschrittener Logfileanalysen[46]

Enthaltene Informationen	Betrachtetes Feld
Pageimpression, Page-view, View,	Request, Status Code
Visit	Request, Status Code, Referrer
User	Request, Status Code, Referrer, Remotehost
Session	Request, Status Code, Referrer, Remotehost, Date/Timezone
Adimpression	Request, (evtl. i. Komb. mit Remotehost und / oder Status Code)
Adclick	Request, Referrer, Status Code, Remotehost
Clickthrough-Rate	Request, Referrer, Status Code, Remotehost
Häufigste Einstiegsseiten	Referrer in Verbindung mit Request
Nationalitäten d. Besucher	Remotehost

Für fortgeschrittene Auswertungen sind – noch mehr als bei einfachen Auswertungen – Logfile-Analyseprogramme oder geeignete Softwaretools nötig. Häufig kommen auch Statistikprogramme oder On-Line Analytical Processing-Tools (OLAP-Tools) zur Bestimmung der angesprochenen Kennzahlen zum Einsatz. Jede Softwarekategorie hat ihre eigenen Vorzüge, wobei OLAP-Tools durch ihre multidimensionalen Fähigkeiten weit flexiblere Analysen erlauben als „einfache“ Analyseprogramme.

Während OLAP- und Logfile-Analysetools hypothesentestend nach dem Top-Down-Ansatz arbeiten, gehen Data Mining-Tools datengesteuert vor.[47]

> *„Ein Data Mining-Tool ersetzt kein Website-Analysetool, aber es eröffnet dem Webadministrator viele zusätzliche Möglichkeiten zur Beantwortung verschiedener Marketing- und Planungsfragen.“* [48]

[45] vgl. NetGeo Inc. (2003) URL: http://www.netgeo.com/technology/technology.html – Zugriff am 01.12.2003.

[46] vgl. Marschall, N. (2002), S. 47 ff

[47] vgl. Mena, J. (2000), S. 90 ff

[48] Mena, J. (2000), S. 92

2.5 Data Mining

Bei der Weiterverarbeitung der Logfile-Daten kommt die eigentliche Mining–Technik zum Einsatz, welche das Web Mining von trivialen Logfile-Analysen differenziert. Es handelt sich hierbei um Data Mining-Methoden, mit deren Hilfe es möglich ist, nach Antworten auf nicht gestellte Fragen zu suchen.[49] Auch wenn diese Aussage sehr abstrakt erscheint, ist genau dies jedoch der entscheidende Vorteil, den Data Mining gegenüber anderen Auswertungstechniken hat. Data Mining ist darauf ausgerichtet, in strukturierten Datenbeständen nach Mustern und Strukturen zu suchen.[50] Dieser strukturierte Datenbestand besteht im Web Log Mining aus den Daten des aufbereiteten Logfiles.

> *„Data Mining bezeichnet den Prozess, der automatisch vorher unbekannte, interessante und interpretierbare Zusammenhänge in großen Datenmengen zu finden vermag.“* [51]

Data Mining wird in einschlägiger Literatur meist als ein Bestandteil der Disziplin „Knowledge Discovery in Databases“ gesehen. Chamoni/Gluchowski sehen KDD als eine dem Data Mining übergeordnete Disziplin, die auf Wissensgewinnung ausgerichtet ist, während Data Mining sich auf Informationsgewinnung aus großen Datenbeständen konzentriert.[52]

Data Mining ist keine genormte Technologie, die sich auf wenige Methoden beschränkt, sondern hat mittlerweile eine Vielzahl an Vorgehensweisen hervorgebracht, die sich unterschiedlich gut im Web Mining einsetzen lassen. Die wichtigsten Data Mining-Methoden im Web Log Mining werden im Folgenden vorgestellt.

2.5.1 Assoziations- und Pfadanalyse

Bei der Assoziations- und Pfadanalyse wird nach häufig auftretenden Mustern im Ausgangsmaterial gesucht. Von besonderem Interesse sind hierbei:[53]

- Mengen von Objekten, die gemeinsam in Nutzungsvorgängen auftreten,
- Abfolgen der zusammen auftretenden Objekte,
- „Wenn-Dann“-Regeln, die aus den Abfolgen abgeleitet werden können.

Es soll also herausgefunden werden, ob beim Aufruf einer Webseite (hier auch als Objekt bezeichnet) bei vielen Nutzungsvorgängen auch eine weitere Webseite (ein weiteres Objekt) aufgerufen wird. Wie häufig Objekte gemeinsam in einem Nutzungsvor-

---

[49] vgl. Walther, R. (2001), S. 16
[50] vgl. Grothe, M. / Gentsch, P. (2000), S. 177
[51] Walther, R. (2001), S. 16
[52] vgl. Chamoni, P. / Gluchowski, P. (1998), S. 20
[53] vgl. Spiliopoulou, M. / Berendt, B. (2002), S. 145

gang auftreten müssen, damit sie assoziiert werden, hängt von der jeweiligen Untersuchung ab, der dafür notwendige Schwellenwert kann prinzipiell frei gewählt werden. Erreichen zwei oder mehrere Objekte den Schwellenwert und wird so zwischen ihnen eine Abhängigkeit festgestellt, dann werden sie in einem so genannten „Itemset“ zusammengefasst. Ein Itemset ist eine Gruppe von Objekten (in diesem Fall Webseiten), zwischen denen eine Abhängigkeit festgestellt wird. Für die Assoziation selbst spielt die Reihenfolge der Objekte im Itemset keine Rolle. Eine in Nutzungsvorgängen gefundene Assoziation könnte also lauten: „Objekt A und Objekt B bilden ein Itemset, weil sie in 15% aller Nutzungsvorgänge nacheinander aufgerufen wurden“.

Bei der Betrachtung von Abfolgen der Objekte ist die Reihenfolge der Aufrufe hingegen entscheidend und es wird eine so genannte „Sequenz“ betrachtet. Wiederum wird ein Schwellenwert festgelegt, der die Häufigkeit definiert, die nötig ist, damit eine Abfolge in die Untersuchung einfließt.[54]

Es können mit dieser Technik auch komplexe Bewegungspfade analysiert werden, und Aussagen über die Häufigkeit ihres Auftretens getroffen werden. Zum Beispiel kann der Besuch der Seiten A und B mit der Symbolik [A, *, B] beschrieben werden, wenn der Pfad zwischen den Seiten A und B irrelevant ist. Sollen zwischen dem Aufruf der Seiten A und B maximal 3 andere Seiten besucht worden sein, kann dies durch die Symbolik [A, [0;3], B] ausgedrückt werden.[55] Diese Technik der Assoziation von Objekten wird im Web Mining gemäß ihrer Anwendung oft auch als Pfadanalyse,.die Analyse von Bewegungspfaden als Clickstreamanalyse bezeichnet. Die Herkunft des Wortes Clickstream basiert darauf, dass der Bewegungspfad eines Websitebesuchers aufgrund von aufeinander folgenden (Stream) Mausaktionen (Klicks) zustande kommt. Gibt ein Internetnutzer über die Tastatur eine andere URL in seinen Browser ein, so verlässt er meist die bis dahin besuchte Website und unterbricht damit den „Clickstream“.

2.5.2 Clusteranalyse

Ziel der Clusteranalyse ist es, den Datenbestand nach bestimmten Kriterien zu gruppieren. Die einzelnen gefundenen Gruppen werden Cluster (oder Segmente) genannt. Die Clusteranalyse ist das Verfahren, welches zur Bildung von Segmenten weite Verbreitung gefunden hat. Wegen des einfachen Aufbaus der Clusterverfahren eignen sie sich sehr gut zur Segmentierung von Datenbeständen.[56]

> *„Unter dem Begriff Clusteranalyse werden multivariante statistische Verfahren subsumiert, die eine umfangreiche und ungeordnete Objektmenge in kleinere, in sich homogene Teilmengen gliedern.“* [57]

[54] vgl. Spiliopoulou, M. / Berendt, B. (2002), S. 145 f
[55] vgl. Spiliopoulou, M. / Berendt, B. (2002), S. 146
[56] vgl. Rapp, R. / Guth, S. (2003), S. 169
[57] Bensberg, F. (2002), S. 176

Die vorgenommene Segmentierung zeichnet sich dadurch aus, dass die nach den Kriterien eingeteilten Segmente in sich möglichst homogen sind, sich aber gegenüber anderen Segmenten möglichst stark unterscheiden. Der Vorgang der Segmentierung erfolgt in zwei Schritten:[58]

1. Dem Bestimmen der Ähnlichkeit von Objekten mithilfe eines „Proximitätsmaßes“.
2. Dem Zusammenfassen ähnlicher Objekte durch einen „Fusionierungsalgorithmus“.

Clusterverfahren gehen entweder hierarchisch oder partitionierend vor, wobei es in der hierarchischen Vorgehensweise die agglomerierende und die divisive Methode gibt.[59] Das partitionierende Verfahren setzt voraus, dass ein Ausgangszustand der Clusterung vorliegt, der optimiert werden soll. Sukzessives Austauschen von Objekten und wiederholtes Überprüfen des Ergebnisses soll dann zu einer Verbesserung der Clusterung führen. Sind die einzelnen Cluster ausreichend heterogen untereinander, was durch einen zuvor festgelegten Wert überprüft werden kann, ist es nicht möglich, eine Verbesserung in der Anordnung zu erreichen und der Austauschprozeß ist abgeschlossen.[60] Das divisive hierarchische Verfahren geht von einem Extrem (gar keine Clusterung) der Clusterung aus und arbeitet sich durch wiederholtes Splitten zum anderen Extrem (maximale Clusterung) vor. Die agglomerierende hierarchische Methode geht den umgekehrten Weg. Um dies zu verdeutlichen, zeigt Abbildung 5 die agglomerative hierarchische Clusterung fiktiver Objekte 1 bis 5.

Als Nachteil der Clusterverfahren muss eingeräumt werden, dass sie für komplexe Kennzahlenstrukturen ungeeignet sind[61]. Die Clusteranalyse ist somit eher bei Objekten mit überschaubarer Kennzahlendichte anzuwenden.

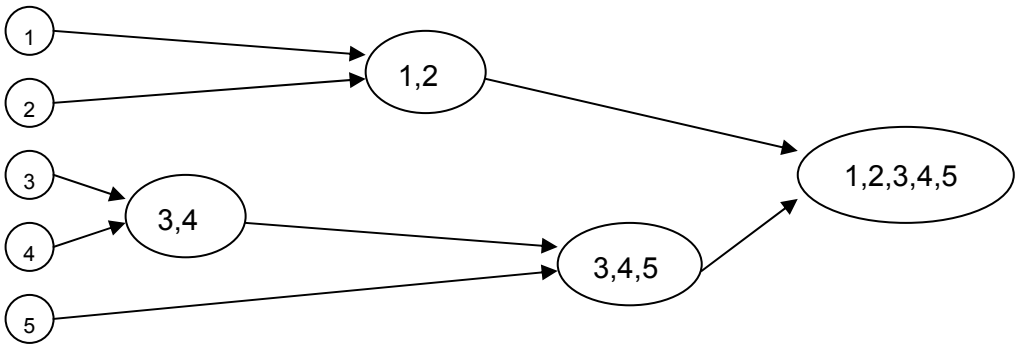

Abbildung 5: Agglomerative hierarchische Clusterbildung[62]

[58] vgl. Bensberg, F. (2002), S. 176ff
[59] vgl. Bensberg, F. (2002), S. 179
[60] vgl. Bensberg, F. (2002), S. 179
[61] vgl. Rapp, R. / Guth, S. (2003), S. 169
[62] vgl. Chamoni, P. / Gluchowski, P. (1998), S. 307

2.5.3 Künstliche Neuronale Netze

Künstliche Neuronale Netze (KNN) eignen sich besonders zur Klassifikation und zur Prognose. Sie wurden als Simulationsmodell für das menschliche Gehirn in den 40er Jahren entwickelt.[63]

> *„Ein neuronales Netz ist ein Verbund von einfachen Informationsverarbeitungseinheiten, die sich über Gewichte gegenseitig Signale zusenden.“* [64]

Ein KNN setzt sich aus Schichten (auch Layer genannt) zusammen. In der Regel sind dies eine Inputschicht, über die Daten eingegeben werden, eine Hiddenschicht, die von äußeren Einflüssen abgeschottet Berechnungen anstellt, und die Outputschicht, welche die vorgenommenen Berechnungen ausgibt. Jede Schicht setzt sich aus Neuronen (den informationsverarbeitenden Einheiten) zusammen, die Anzahl der Neuronen je Schicht bestimmt die Topologie des jeweiligen KNN. Je nach Topologie eignen sich Künstliche Neuronale Netze für unterschiedliche Zwecke. Zwei wichtige Vertreter Neuronaler Netze sind Backpropagation-Netze und Kohonen-Netze.

Bei *Backpropagation-Netzen* (BPN) ist die Topologie dreischichtig. Bei ihnen kommen Trainings- und Validierungsdaten zum Einsatz, die ein neuronales Netz für einen Lernprozess benötigt, um mit einer bestimmten Datenstruktur umgehen zu können. Trainingsdaten werden dazu verwendet, ein Neuronales Netz so zu trainieren, dass es in der Lage ist, bei einem gegebenen Input die richtigen Berechnungen anzustellen, und somit die Ergebnisse möglichst wenig vom gewünschten Wert abweichen. Ist das Neuronale Netz ausreichend trainiert, wird es als „konvergiert“ [65] bezeichnet. Validierungsdaten kommen zum Einsatz, um eine Überanpassung des Netzes an die Trainingsdaten zu verhindern. Überanpassung würde bedeuten, dass das KNN allein mit den Trainingsdaten richtig funktioniert. Da es aber für andere Daten flexibel bleiben soll, werden die Validierungsdaten eingesetzt. In manchen Fällen werden zusätzlich auch Testdaten verwendet, die aber auf den Lernprozess des Neuronalen Netzes keinen Einfluss ausüben.[66]

Kohonen-Netze, oft auch als Self Organizing Maps (SOM) bezeichnet, werden häufig zur Segmentierung eingesetzt. Sie haben eine einfache, zweischichtige Topologie, die sich aus Input- und Outputlayer zusammensetzt. Der Inputlayer besitzt genau die Anzahl an Neuronen, die der Zahl der einzulesenden Informationen entspricht. SOM können komplexe Merkmalsräume zweidimensional abbilden und vorhandene Strukturen in der Datenbasis aufdecken. [67]

---

[63] vgl. Meyer, M. (2002), S. 198 f
[64] Callan, R. (2003), S. 32
[65] vgl. Callan, R. (2003), S. 52
[66] vgl. Meyer, M. (2002), S. 200 f
[67] vgl. Meyer, M. (2002), S. 202 f

Im Gegensatz zu Backpropagation-Netzen lernen Self Organizing Maps unüberwacht. Backpropagation-Netze werden wegen ihrer überwachten Lernphase als strukturprüfendes Verfahren bezeichnet, Self Organizing Maps hingegen als strukturaufdeckendes Verfahren.[68] Dies wird auch durch den Anwendungskontext unterstrichen, da Backpropagation-Netze zur Klassifikation und Prognose eingesetzt werden, Self Organizing Maps hingegen zur Segmentierung.

2.5.4 Entscheidungsbäume

Ziel des Entscheidungsbaumverfahrens ist das Generieren von Klassifikationsregeln.

> *„Zu diesem Zweck versuchen Entscheidungsbäume, aus einer Menge unabhängiger Variablen diejenigen zu identifizieren, die besonders gut zwischen den Klassen der abhängigen Variablen, z.B. zwischen Kauf und Nichtkauf eines Produkts, trennen.“* [69]

Variablen können Ereignisse oder Zustände sein, die durch Kennzahlen beschrieben werden, zum Beispiel das Downloaden einer Datei oder der Abbruch eines Kaufvorgangs.
Ein Entscheidungsbaum besteht aus einem singulären Wurzelelement, auf dem sich der Ereignisbaum aufbauen lässt. Alle Elemente, die wiederum Unterelemente besitzen, werden mit Ausnahme des Wurzelelements als „Innere Knoten“ bezeichnet. Elemente, die kein Unterelement besitzen, werden als „Blattknoten“ bezeichnet.[70] Abbildung 6 soll die Bezeichnungen nochmals verdeutlichen.

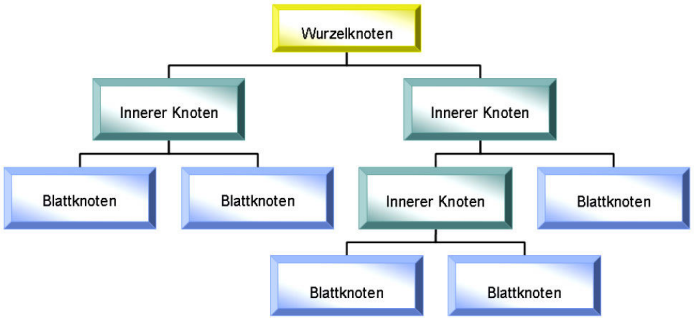

Abbildung 6: Exemplarische Entscheidungsbaumstruktur[71]

Werden die Entscheidungen anhand eines einzigen Merkmals getroffen, wird der Entscheidungsbaum als „univariat“ bezeichnet. Ist mindestens eine Entscheidung im Ereignisbaum auf mehr als einem Merkmal aufgebaut, wird er als „multivariat“ bezeichnet. Für die Unterteilung kommen nominale und quantitative Merkmale in Frage. Wird ein Entscheidungsbaum auf eine Datenbasis angewendet, ergeben sich die Klassifika-

[68] vgl. Meyer, M. (2002), S. 204
[69] Meyer, M. (2002), S. 205
[70] vgl. Meyer, M. (2002), S. 205
[71] vgl. Meyer, M. (2002), S. 205 f

tionsregeln direkt aus dem Aufbau des Entscheidungsbaums.[72] Der Wurzelknoten bildet 100% der Datenbasis ab. Die Daten werden dann in unterschiedlichen Verteilungen an die unteren Ebenen weitergereicht. So lässt sich an den Blattknoten und auch an jeder anderen Stelle im Entscheidungsbaum die Verteilung der Datenbasis anhand der definierten Klassifizierung direkt ablesen.

2.5.5 Zuordnung von Aufgaben im Web Log Mining

Die vorgestellten Data Mining-Methoden Assoziations- und Pfadanalyse, Clusteranalyse, Künstliche Neuronale Netze und Entscheidungsbäume finden im Web Mining an unterschiedlichen Stellen ihre Anwendung. Je nach gestellter Aufgabe sind einzelne für manche Untersuchungen besser geeignet als andere. Abbildung 7 zeigt anhand typischer Web Mining-Fragestellgungen, welche Aufgaben Data Mining-Methoden übernehmen können, um Web Mining-Fragestellungen zu beantworten.

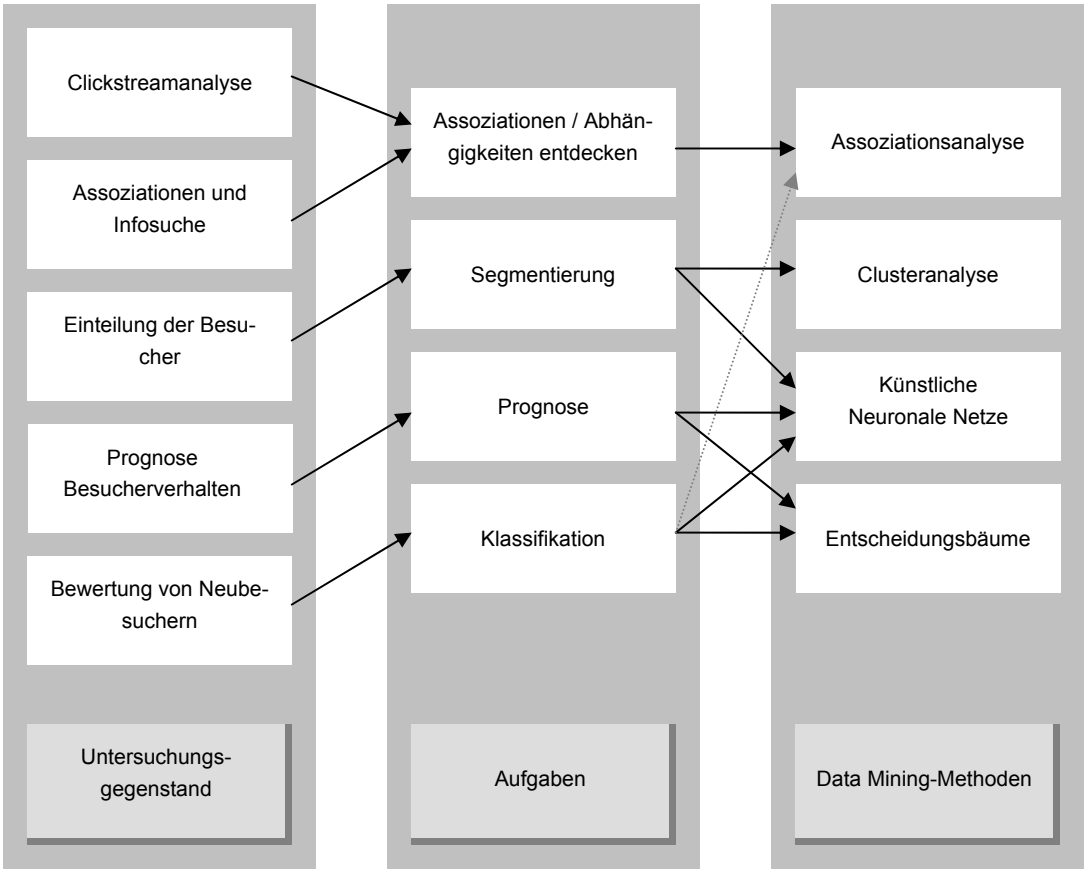

Abbildung 7: Zuordnung von Fragestellungen und Aufgaben im Web Mining zu Data Mining-Methoden[73]

[72] vgl. Meyer, M. (2002), S. 205 ff

[73] vgl. Grothe, M. / Gentsch, P. (2000), S. 179
sowie Klossek, M. (2001), URL: http://www.eworks.de/research/2001/05/WebLogMining/WebLogMining.pdf – Zugriff am 28.10.2003. – S. 11

2.6 Datenschutz

Die Auswertung von Logfile-Daten ist gekennzeichnet von dem Bestreben, möglichst viel über das Verhalten und über die Eigenschaften des Websitebesuchers zu erfahren. Schnell gerät dieses Bestreben aber in Konflikt mit geltendem Recht. Wo die kritischen Stellen bei der Verarbeitung von Logfiles liegen und was datenschutzrechtlich beachtet werden sollte, wird im Folgenden dargestellt.

In Logfiles wird festgehalten, welche Anfragen, ausgelöst durch Aktionen der Webseitennutzer, an den Webserver einer Internetpräsenz gestellt werden. Der Besucher einer Webseite ist sich erfahrungsgemäß nicht bewusst, dass jeder Seitenaufruf und jeder Klick auf einer Internetseite mindestens einen, bis hin zu einer Vielzahl von Einträgen in einem Logfile produziert. Schulzki-Haddouti schreibt in der Online-Zeitschift Telepolis zu diesem Thema:

> *„Fälschlicherweise nehmen einige Nutzer an, dass der Besuch der bei Google im Zwischenspeicher beziehungsweise Cache gespeicherten Websites nicht protokolliert wird. Dabei überträgt der Browser an die Website die Referrer-Meldung samt Cache-Nummer und Suchbegriffen […]“* [74]

Dieser naive Irrglaube kann nur allzu leicht ausgenutzt werden, indem gezielt jene Logfiledaten ausgewertet werden, von denen der Website-Besucher in der Regel nicht erwartet, dass sie aufgezeichnet werden. Vor dem Hintergrund, dass den meisten Internetnutzern nicht bewusst oder gar bekannt ist, welche Daten beim Surfen im Netz gespeichert werden, muss nun klar gestellt werden, welche rechtlichen Grundlagen den Umgang mit Logdaten und deren Verarbeitung regeln.

2.6.1 Rechtliche Grundlagen

Innerhalb der Europäischen Union findet das jeweilige nationale Recht dort Anwendung, wo Daten erhoben, weiterverarbeitet oder genutzt werden. Außen vor bleiben Daten, die sich lediglich im Transit befinden. Internationale Regelungen gibt es derzeit nicht, deshalb kann prinzipiell durch geschickte Standortwahl das EU-Recht umgangen werden. Vorrausetzung dafür wäre, dass die Datenerhebung und Datenverarbeitung in aussereuropäischem Ausland stattfinden müsste und auch die Verantwortlichkeiten bei aussereuropäischen Firmen liegen müssten.[75]

Im Bundesdatenschutzgesetz (BDSG) ist festgelegt, wie mit „personenbezogenen Daten“ umgegangen werden soll, um das Persönlichkeitsrecht des Einzelnen nicht zu verletzen:

---

[74] Schulzki-Haddouti, C. (2003)
URL: http://www.heise.de/tp/deutsch/inhalt/te/14052/1.html - Zugriff am 09.10.2003.

[75] vgl. Arndt, D. / Koch, D. (2002), S. 78

> *„Zweck dieses Gesetzes ist es, den Einzelnen davor zu schützen, dass er durch den Umgang mit seinen personenbezogenen Daten in seinem Persönlichkeitsrecht beeinträchtigt wird.“* [76]

Personenbezogene Daten sind im BDSG definiert als *„[…] Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person (Betroffener).“* [77]

Es werden drei Kategorien personenbezogener Daten unterschieden:[78]

- *Direkter Personenbezug*
  Bei Daten, die einen direkten Personenbezug aufweisen, ist die Identität einer Person erkennbar. Sie liegt offen und muss nicht weiter bestimmt werden.
- *Indirekter Personenbezug*
  Daten mit indirektem Personenbezug erlauben es, die Identität einer Person indirekt herzustellen. Der Personenbezug gilt als bestimmbar.
- *Ohne Personenbezug*
  Daten ohne Personenbezug gelten als anonyme Daten. Ein Personenbezug ist hier gar nicht, oder nur mit unverhältnismäßig hohem Aufwand an Zeit, Kosten und Arbeitskraft herstellbar.

Nur Daten mit direktem oder indirektem Personenbezug unterliegen im Sinne des BDSG dem Datenschutzrecht. Daten ohne Personenbezug sind durch das Datenschutzrecht ebenso wenig geschützt wie so genannte „aggregierte Daten“. Aggregierte Daten weisen keinen Personenbezug auf eine Einzelperson auf, weil sie nur Rückschlüsse auf eine Gruppe von Personen zulassen. Eine Aggregierung im Sinne des BDSG zur Entfernung des Personenbezugs kann auf einer Ebene an Merkmalen von Personen stattfinden, die sie in der Regel voneinander unterscheiden, aber auch bei allen vorhanden sind (Gewicht, Alter, Einkommen, etc.). Das Zusammenfassen einer Menge von Personen zu einer Gruppe über das Merkmal Einkommen in der Form, dass für die Gruppe ein Durchschnittseinkommen angegeben wird, ist dabei eine mögliche Form der Aggregierung. Hingegen wäre ein Zusammenfassen aller Personen, die „Meier“ heißen zu einer Gruppe keine korrekte Aggregierung, weil dann das persönliche Merkmal „Name“ der einzelnen Person nicht verloren geht.[79]

Außerdem unterliegen nur natürliche Personen dem Datenschutzgesetz, juristische Personen haben keinen Schutzanspruch.[80]

---

[76] BDSG, §1, Abs. 1, URL: http://bundesrecht.juris.de/bundesrecht/bdsg_1990/__1.html, Zugriff am 10.11.2003

[77] BDSG, §3, Abs. 1, URL: http://bundesrecht.juris.de/bundesrecht/bdsg_1990/__3.html, Zugriff am 10.11.2003

[78] vgl. Arndt, D. / Koch, D. (2002), S. 81 f

[79] vgl. Arndt, D. / Koch, D. (2002), S. 82

[80] vgl. Arndt, D. / Koch, D. (2002), S. 82

Der Schutz personenbezogener Daten besteht – wie bei jedem deutschen Datenschutzgesetz – neben einem so genannten „Erlaubnisvorbehalt“, der besagt, dass bei einer ausdrücklichen Erlaubnis durch ein Gesetz oder durch die betroffenen Personen das Verbot der Verarbeitung personenbezogener Daten nicht greift.[81]

Eine Erlaubnis betroffener Personen (Einwilligung) ist genau geregelt. Obwohl eine Einwilligung generell schriftlich erfolgen sollte, ermöglicht doch das Teledienstedatenschutzgesetz (TDDSG) eine elektronisch übermittelte Einwilligung. Diese ist an folgende Bedingungen gebunden: Der Nutzer muss seine Einwilligung eindeutig und bewusst abgeben, die Einwilligung darf nicht manipuliert werden können, und die Einwilligung muss nachgewiesen werden können und deshalb protokolliert werden. [82]

Kommt es zu einer (legalen) Erhebung, Verarbeitung oder Nutzung personenbezogener Daten, hat der betroffene Nutzer keineswegs seine Rechte abgegeben. Er hat jederzeit das Recht auf Widerruf seiner Einwilligung, das Recht auf Kenntnis der Speicherung seiner Daten und ein „Auskunftsrecht“. Das Auskunftsrecht besagt, dass der Nutzer stets das Recht auf Einsicht der über ihn oder zu seinem Pseudonym gespeicherten Daten hat.[83] Diese gesetzliche Vorschrift impliziert, dass pseudonymisierte Daten einen indirekten Personenbezug aufweisen müssen, um den Nutzern ihr Auskunftsrecht einräumen zu können. Eine Entfernung des Personenbezugs zum Zwecke des uneingeschränkten Umgangs ist somit nicht praktikabel.

2.6.2 Ethische Aspekte der Logfile-Auswertung

Logfile-Auswertungen werden als innerbetrieblicher Vorgang in der Regel nicht an die Öffentlichkeit kommuniziert. Wenn ein Website-Besucher erfährt, dass seine Bewegungen im Internet nicht nur aufgezeichnet, sondern auch ausgewertet werden, kann das einen Vertrauensbruch zur Folge haben. Es ist zwar allgemein bekannt, dass Provider und Website-Betreiber alle Bewegungen der Internetnutzer nachvollziehen können, doch fühlen sich Internetnutzer vermeintlich „sicher“ beziehungsweise anonym, wenn sie Suchmaschinen nutzen oder sich per Call-by-Call-Provider ins Internet einwählen. Die Tatsache, dass die Speicherung von Weblog-Protokollen ohne das explizite Wissen beziehungsweise ohne das Bewusstsein der Betroffenen darüber geschieht, ist nicht vertrauensbildend und wirkt somit dem Ziel von Public Relations entgegen, Vertrauen aufzubauen und zu erhalten. Gerade deshalb muss der Umgang mit Logfiles in Zusammenhang mit Public Relations sorgfältig geplant und durch selbst gesteckte Grenzen auf ein Mindestmaß beschränkt bleiben.

Überwachung im und durch das Internet ist ein Umstand, der die Anonymität des Einzelnen beeinträchtigen kann. Rötzer, Redakteur der Online-Zeitschrift Telepolis, schreibt hierzu:

---

[81] vgl Arndt, D. / Koch, D. (2002), S. 83
[82] vgl. Arndt, D. / Koch, D. (2002), S. 87
[83] vgl. Arndt, D. / Koch, D. (2002), S. 88

> *„Eine der Formen des Privatseins ist Anonymität, also ganz einfach unbekannt zu bleiben, […]. Das vielgepriesene Internet hat mit seiner zunehmenden Kommerzialisierung jedoch gezeigt, daß die Gefahr der permanenten Identifizierung und Verfolgung jeder Handlung die man im Datennetz unternimmt, nicht nur vom jeweiligen Großen Bruder, sondern auch von den vielen kleinen Brüdern ausgeht.“* [84]

Zudem werden durch die Möglichkeiten der Überwachung, die das Internet bietet, Schutzmaßnahmen und Mechanismen, die über Generationen gewachsen sind und dem Einzelnen ein Recht auf Privatheit und Anonymität bieten, aufgeweicht. Es wird als Selbstverständlichkeit angesehen, dass es ein Grundrecht jeder Person ist, sich im Internet genauso anonym zu bewegen, wie in der Öffentlichkeit (z. B. auf der Straße).[85]

Ein verantwortungsvoller Umgang mit Logfiles sollte aber oberstes Prinzip sein, vor allem wenn es um Public Relations geht. Die angesprochenen Bedenken zum Urteil der Website-Besucher über eine Logfile-Auswertung, als auch die rechtlichen Grundlagen, führen zwingend zu einem verantwortungsvollen Umgang mit den Daten, die Logfiles bereitstellen. Gerade der Einsatz von Techniken des Data Mining bietet hier eine ideale Grundlage, mit möglichst wenig personenbezogenen Daten aufschlussreiche und nützliche Ergebnisse zu erhalten. Denn anders als beim „Integrated Web Usage Mining“[86] verzichtet Web Log Mining auf das Zusammenführen des Logfiles mit anderen Datenquellen wie Kundenstammdaten oder anderen personenbezogenen Quellen. Broder hebt in seinem im Jahr 2000 veröffentlichten Artikel „Data Mining, the Internet, and Privacy“ hervor, dass Experten den Konflikt, über Internetnutzer Informationen zu sammeln ohne deren Anonymität verletzen zu wollen, durch das Data Mining ideal berücksichtigt sehen.[87]

---

[84] Rötzer, F. (1999)
URL: http://www.heise.de/tp/deutsch/inhalt/te/5053/1.html. - Zugriff am 10.11.2003

[85] vgl. Bäumler, H. / Mutius, A. von (Hrsg) (2003), S. 1 ff

[86] vgl. Abbildung 2: Einordnung des Web Log Mining, S. 12, Kapitel 2.1

[87] vgl. Broder, A. J. (2000), S. 56

3 Web Log Mining im Rahmen der Online-PR

3.1 Online-PR

Online-PR ist in den letzten Jahren zu einem integrierten Bestandteil der Public Relations im Allgemeinen avanciert. Das Medium Internet hat sich für die PR neben klassischen Kommunikationsmedien (Zeitung, Fernsehen, Hörfunk, etc.) in vielen Fällen als unverzichtbar herausgestellt und sollte dabei als fester Bestandteil des „Instrumentenmix“ der Public Relations betrachtet werden.[88] Online-PR (respektive das Internet) ersetzt keine klassischen Kommunikationsinstrumente, sondern ergänzt diese.[89] Dies hat zur Folge, dass ein Abstimmungsprozess zwischen der PR, die über das Internet kommuniziert wird, und der PR über andere Medien vorhanden sein muss. Es sollte klar definiert sein, welche Zielgruppen online angesprochen werden sollen, und welche Informationen online angeboten werden – wie schon erwähnt – in Abstimmung mit den anderen verwendeten Kommunikationsinstrumenten.[90]

Die speziellen Vorteile der Online-PR sollten berücksichtigt und bewusst eingesetzt werden.

3.1.1 Spezielle Merkmale der Online-PR

Das Internet als Medium der Online-PR ist vorrangig Pull-Medium[91], hat am Rande aber auch einen Push-Charakter, wenn beispielsweise Emails gezielt zur Informationsverbreitung eingesetzt werden.[92] Websites haben einen durchgängigen Pull-Charakter: der Nutzer ruft von sich aus Informationsinhalte ab. Informationen müssen also auf einer Website bereitgestellt werden. Die Informationen und Angebote einer Website werden aber nur dann vom Nutzer abgerufen, wenn sie für ihn interessant und neu sind, oder der Nutzer durch sie einen so genannten „Mehrwert“ erhält, da der Besucher einer Website nicht nur selbst aktiv werden muss, sondern er muss für seinen Besuch sowohl Zeit als auch Geld aufwenden, und hat jederzeit die Möglichkeit, seinen Besuch abzubrechen.[93]

---

[88] vgl. Fuchs, P. / Möhrle, H. / Schmidt-Marwede, U. (1999), S. 8

[89] vgl. Fuchs, P. / Möhrle, H. / Schmidt-Marwede, U. (1999), S. 13

[90] vgl. Herbst, D. (2001), S. 26 f

[91] Bei einem Pull-Medium muss der potenzielle Rezipient selbst aktiv werden, um Informationen zu erhalten. Das Gegenteil, ein Push-Medium, trägt Informationen aktiv an den Rezipienten heran.

[92] vgl. Sauvant, N. (2002), S. 171 f

[93] vgl. Sauvant, N. (2002), S. 171

Online-PR ist schnell und flexibel – schneller und flexibler als andere Kommunikationskanäle, wie zum Beispiel die Print-Medien.[94] Diese Eigenschaft muss als echter Vorteil von Online-PR erkannt werden. Gleichzeitig bedeutet dies aber, dass auch schnell Fehlentscheidungen publik werden können. Wird z. B. eine Unternehmens-Website zur Kommunikation von Nachrichten oder Informationen eingesetzt, wird von den Besuchern dieser Website automatisch vorausgesetzt, dass die Inhalte hoch aktuell und wahr sind. Sollte sich herausstellen, dass ein anderes Medium, z. B. eine Nachrichtensendung in Hörfunk oder Fernsehen, aktuellere Informationen bietet, wird nicht nur das Ansehen der Internetpräsenz, sondern auch das Image des Unternehmens in Mitleidenschaft gezogen. Das würde den Zielen der PR entgegenwirken und unterstreicht wiederum die Notwendigkeit, dass Online-PR mit anderen Kommunikationsmedien der PR abgestimmt werden muss.

Angebotene Informationen bei Online-Angeboten müssen so zeitnah wie möglich bereitgestellt werden, dies wird von Fuchs sehr treffend beschrieben: *„If its not just in time, its too late.“* [95] Diese Aussage bezieht Fuchs nicht nur auf die Aktualität der Website selbst, sondern gleichermaßen auf Antwortzeiten von Email-Anfragen der Websitebesucher. Diese Anforderung entsteht daraus, dass Informationen im Internet sofort nach ihrer Veröffentlichung unabhängig von Ort und Zeit abgerufen werden können und werden.

Da sich Online-PR nahezu vollständig auf Pull-Technologie stützt, muss den Zielgruppen ein Mehrwert angeboten werden, der sie dazu animiert, in regelmäßigen Abständen die Website zu besuchen. Je nach Zielgruppe kann dieser unterschiedlich beschaffen sein. So kann der Zugriff auf einen Presseserver des Unternehmens über die Website einen enormen Mehrwert für Journalisten darstellen[96], für Privatpersonen hingegen sind Produktinformationen oder ein Online-Spiel Mehrwert-Faktoren einer Website. Herbst schreibt hierzu: *„Ein fehlender einzigartiger Nutzen macht den Webauftritt unattraktiv.“* [97]

Im Laufe der Zeit haben sich drei Gruppen von Netzen herausgebildet, auf denen Webseiten eingesetzt werden. Sie grenzen sich durch ihre unterschiedliche Zugänglichkeit voneinander ab: Das Internet selbst, Intranets und Extranets. Werden im Internet nahezu alle denkbaren Zielgruppen angesprochen, schränken Intranets und Extranets ihre Zielgruppen ein. Auf Intranets können meist nur diejenigen Personen zugreifen, die auch Zugang zum Firmennetzwerk haben, z. B. Mitarbeiter. Intranets sind also „nach innen“ ausgerichtet. Extranets erlauben den Zugriff auf ausgewählte Informationen „von außen“. So können beispielsweise wichtige Benutzergruppen wie Handelspartner, Lieferanten oder Journalisten Zugriff auf firmeninterne Daten erhalten, die für

[94] vgl. Fuchs, P. / Möhrle, H. / Schmidt-Marwede, U. (1999), S. 13
[95] Fuchs, P. / Möhrle, H. / Schmidt-Marwede, U. (1999), S. 17
[96] vgl. Grudowski, S. (2001), S. 83 ff
[97] Herbst, D. (2001), S. 12

andere Personen nicht zugänglich sind. Durch Benutzeraccounts wird sichergestellt, dass nur auf freigegebene Ressourcen zugegriffen werden kann.[98]

Wichtige Vorteile der Online-PR sind:

- schnelle Ansprache der Zielgruppen, schnelle Aktualisierung möglich
- Unabhängigkeit von Ort und Zeit
- gute Möglichkeit für das Einbinden von Mehrwert-Inhalten
- Überprüfbarkeit der Verbreitung durch Web Log Mining
- weitere Kommunikationsschnittstellen wie Email, SMS, Chat, Web-TV oder Web-Radio können in ein Online-PR-Konzept eingebunden werden.

Allerdings erreicht Online-PR nur diejenigen Internetnutzer, welche die Website besuchen, dies muss als Nachteil erwähnt werden.

3.1.2 Zielgruppen der Online-PR

Unabhängig von Web Log Mining und Controlling werden in der PR Zielgruppen definiert, um eine klare Vorstellung zu haben, wer durch PR-Aktivitäten angesprochen werden soll. Die Aufteilung der Zielgruppen kann ganz unterschiedlich gestaltet sein. Sie kann aus Sicht des Unternehmens in externe und interne Zielgruppen, aus Sicht der Rolle der Rezipienten in Kunden, Geschäftspartner, Journalisten, Mitarbeiter, Geldgeber, u. s. w., oder aber auch aus einer Sicht über Merkmale der Rezipienten wie Alter, Einkommen, Geschlecht, Interessen, Bildung, u. s. w., erfolgen.

Cornelsen unterscheidet vier PR-Zielgruppen: Institutionen und Unternehmen, Informelle Gruppen, Soziale Gruppen und die Medien.[99] Bei dieser Einteilung werden die externen und die internen Zielgruppen miteinander vermengt, denn alle vier Zielgruppen können sowohl extern als auch intern bei einem Unternehmen vorkommen. *Unternehmen und Institutionen* bieten sowohl durch ihre nach innen als auch durch ihre nach außen gerichtete Infrastruktur sehr gute Kommunikationswege. Schwarze Bretter, Email (-Verteiler), Intranets, Extranets u. s. w. machen es leicht, die Zielpersonen zu erreichen. *Soziale Gruppen* bestehen aus Mitgliedern mit ähnlichen oder gleichen Interessen, die aber nicht in Kontakt miteinander stehen (z. B. die Brillenträger). Das macht es schwierig, sie gemeinsam anzusprechen. *Informelle Gruppen* hingegen sind durch eine lose Verbindung untereinander organisiert und damit besser ansprechbar. In Abgrenzung zu Organisationen haben Informelle Gruppen laut Cornelsen keine explizite Rechtsform. Die PR-Arbeit wird wesentlich erleichtert, wenn Soziale Gruppen zu Informellen Gruppen überführt werden können. Das Internet bietet dafür eine ideale Plattform, beispielsweise durch das Gründen einer virtuellen Community. Auch die Praxis vieler Produkthersteller, die ihre Kunden dazu anhalten, nach dem Kauf ihre Produkte

[98] vgl. Fuchs, P. / Möhrle, H. / Schmidt-Marwede, U. (1999), S. 35 ff
[99] vgl. Cornelsen, C. (2002), S. 31 f

zu „registrieren“, ist oft nichts anderes, als der Versuch, aus den Sozialen Gruppen der „Produktanwender“ Informelle Gruppen von (bekannten und ansprechbaren) Produktbesitzern zu machen. Die *Medien*, als vierte Zielgruppe nach Cornelsen, übernehmen die Funktion eines „wichtigen Bindeglieds“ zur Kommunikation mit sozialen Gruppen. Sie werden mit Informationen versorgt, mit dem Ziel, dass diese wiederum die eigentliche Zielgruppe ansprechen.[100]

Eine besonders wichtige Zielgruppe für die Online-PR sind die von Trendforscher Horx als Premium-Zielgruppe bezeichneten „Online-Worker“.[101] Vor allem Journalisten - mittlerweile ohne Ausnahme mit Internetanschluss am Arbeitsplatz versorgt und ständig auf Informationssuche – gelten als typische Online-Worker und bedürfen somit bei der Online-PR besonderer Aufmerksamkeit. Je gezielter den Journalisten Informationen bereitgestellt werden, umso höher ist der Multiplikationseffekt der Medien an der eigenen PR-Strategie und unterstützt diese. Online-Worker erwarten Informationen im Internet. Wenn dieser Informationsbedarf vom Unternehmen gedeckt wird, kann die Erwartungshaltung befriedigt werden und darüber hinaus können die Informationen dem Unternehmensbild entsprechend gestaltet werden. Nichts wäre schlimmer für das Unternehmensimage, als dass der Informationsbedarf über unabhängige Dritte gedeckt würde, der sich dann nicht direkt kontrollieren ließe und sich auch nicht am strategischen Unternehmensbild orientiert.

Egal, welche Einteilung bevorzugt wird, wichtig bei der Einteilung der anzusprechenden Zielgruppen ist, dass alle relevanten, potentiellen Rezipienten erfasst und entsprechend bedient werden. Außerdem bietet es sich an, entsprechend der zu verwendenden Kommunikationsmedien die Abgrenzung abzustimmen. Es können sowohl Überschneidungen verschiedener Zielgruppen auftreten, als auch unerfasste Lücken, die durch keine definierte Zielgruppe abgedeckt werden. Je weniger Lücken entstehen, umso besser können die Zielgruppen erreicht werden, und je weniger Überschneidungen festzustellen sind, umso gezielter können die jeweiligen Zielgruppen direkt angesprochen werden.

3.1.3 Inhalte der Online-PR

Woran soll sich Online-PR orientieren – welches sollen die Inhalte sein? Ein Mehrwert für den Rezipienten ist dabei von entscheidender Bedeutung.

> *„Denn worauf es letztlich ankommt, das ist der Mehrwert, den Ihr Internetauftritt Ihren Zielgruppen und Ihrem Unternehmen bringt.“* [102]

Mehrwert kann sich auf unterschiedliche Art und Weise herausbilden lassen, und Websites können auf unterschiedlichste Weise „nützlich“ sein, und somit durch den Nutzen für den Besucher das Image eines Unternehmens anheben. Eine zusätzliche Steige-

---

[100] vgl. Cornelsen, C. (2002), S. 31 f

[101] vgl. Fuchs, P. / Möhrle, H. / Schmidt-Marwede, U. (1999), S. 16

[102] Sauvant, N. (2002), S. 209

rung des Imagegewinns kann durch die Qualität des Mehrwert-Angebots erzielt werden. Ein Beispiel dafür ist das Anbieten von Mehrwert für Journalisten. Journalisten (bzw. die Medien) sind Multiplikatoren für die Weiterverbreitung von Informationen und somit eine besonders interessante Zielgruppe. Ein Mehrwert für Journalisten ist zum Beispiel möglich durch:

- Virtuelle Pressekonferenzen. Ein solches Angebot spart Zeit, und Zusatzinformationen können online präsentiert werden, zur direkten Weiterverarbeitung für die redaktionelle Arbeit.[103]
- Das Bereitstellen digitaler Pressemappen zu Pressemitteilungen entweder zum Download über die Website oder über den Versand als Email-Anhang. Bei einer Umfrage der Kommunikationsagentur Schrader unter Online-Journalisten im Herbst 2002 gaben 23,8 Prozent der Befragten an, sie bevorzugten Pressemappen virtuell zum Download, 31,7 Prozent gaben an, sie bevorzugten Pressemappen im Email-Anhang, 36,5 Prozent bevorzugten Pressemappen klassisch per Post, und 7,9 Prozent erklärten, sie würden keine Pressemappen nutzen.[104]
- Den Zugang zu einem Presseserver, der alle veröffentlichten Pressemitteilungen und anderes Informationsmaterial in für die Pressearbeit aufbereiteter Form bereitstellt.[105] Ein eventuell verwendetes Subskriptionsmodell würde einen Überblick darüber erlauben, welche Medien dieses Angebot nutzen können. Bei Zustimmung durch die Nutzer des Angebots könnte sogar ausgewertet werden, welche Informationen abgerufen wurden. Da aber auch andere Gruppen, zum Beispiel Stakeholder oder auch die allgemeine Öffentlichkeit, Interesse an diesen Presseinformationen zeigen könnten[106], wäre eine frei zugängliche oder eine in den Zugangsrechten abgestufte Lösung besser.
- Das Anbieten von Bild-, Film- und Tonmaterial über den Presseserver.[107]
- Den Versand von Pressemitteilungen per Email. Ebenfalls bei der schon erwähnten Umfrage der Kommunikationsagentur Schrader wurde erhoben, wie Journalisten Pressemitteilungen in Emails übermittelt bekommen möchten. Hierbei gaben insgesamt 41,1 Prozent an, sie bevorzugen Pressemitteilungen als Attachement (Email-Anhang) in individuellen Formaten (.doc, .rtf, .pdf und ASCII wurden genannt). 12,7 Prozent bevorzugten die Pressemitteilung direkt

[103] vgl. Sauvant, N. (2002), S. 209 ff

[104] vgl. ECIN (2003). – URL: http://www.ecin.de/marketing/onlinejournalisten/ Zugriff am 16.11.2003

[105] vgl. Grudowski, S. (2001), S. 83 ff

[106] vgl. Grudowski, S. (2001), S. 83

[107] vgl. Grudowski, S. (2001), S. 84

in der Email mit Umbrüchen nach 70 Zeichen, 46 Prozent direkt in der Email ohne harte Umbrüche.[108]

- Weitere, hier nicht näher aufgeführte Serviceleistungen, die Journalisten die Arbeit erleichtern.

Der Mehrwert für die Nutzer, und damit die Online-PR an sich, muss und sollte sich aber nicht darauf beschränken, Informationen bereitzustellen oder in Newslettern zu versenden. Vor allem für die breite Öffentlichkeit gibt es eigentlich keine Grenzen das Medium Internet für PR-Aktionen zu nutzen. Volkswagen bietet zum Beispiel auf seiner Homepage (www.volkswagen.de) einen Menüpunkt „Mobile Services“ an, der viele nützliche Zusatzinformationen und Download-Möglichkeiten bietet – unter anderem Logos und Klingeltöne für Mobiltelefone. Für diesen Bereich muss sich der Nutzer anmelden und erhält Leistungen, die durchweg in Beziehung zu Volkswagen stehen.[109]

Weiterhin können, je nach Branche, Online-Inhalte wie Betriebsanleitungen, Softwareupdates, Stadtpläne und Routenplaner, Kochrezepte oder auch der Service der Sendungsverfolgung bei Paketdiensten einen Mehrwert darstellen, um nur einige zu nennen.

3.2 Online-PR-Controlling

3.2.1 PR-Controlling

Controlling kann an unterschiedlichen Stellen des PR-Prozesses ansetzen und somit PR-Aktivitäten planen, koordinieren, steuern und überwachen. Diese Ansatzpunkte sind, wie es Ebert und Steinhübel darstellen, die PR-Ziele, die PR-Maßnahmen, die PR-Instrumente und das PR-Ergebnis.[110] Diese lassen sich den Controlling-Bereichen Prämissencontrolling, Durchführungscontrolling und Realisationscontrolling zuordnen. Angesichts der Tatsache, dass Web Log Mining das Verhalten der Rezipienten und damit den Erfolg des PR-Prozesses anhand von bestimmten Kennzahlen evaluiert, rückt das „Realisationscontrolling“ im Folgenden in den Mittelpunkt der Betrachtungen. Abbildung 8 zeigt eine Darstellung von Ebert, welche um die Domäne des Web Log Mining erweitert wurde. Web Log Mining untersucht das *Verhalten* der *Rezipienten*, die in Kontakt mit konkreten *PR-Aktionen* gekommen sind. Über das Verhalten können so auch Rückschlüsse auf das *Ergebnis* selbst, die *Rezeption*, gezogen werden. Dies sind die Stellen, an denen Web Log Mining im Sinne eines Controllinginstrumentes ansetzen kann.

---

[108] vgl. ECIN (2003). – URL: http://www.ecin.de/marketing/onlinejournalisten/ Zugriff am 16.11.2003

[109] vgl. Volkswagen AG (2003). – URL: http://mobileservices.volkswagen.de/0_mobilizer/0_index/?serviceid=handy Zugriff am 16.11.2003

[110] vgl. Ebert, G. / Steinhübel, V. (1995), S. 10 ff

Aufgrund der nicht einfach erfassbaren wechselseitigen Beziehungen zwischen der Öffentlichkeit und Unternehmen wird häufig auch von „Evaluierung“ der Public Relations gesprochen. Für eine praxisnahe Evaluierung und ein erfolgreiches Controlling ist es außerdem von Vorteil, wenn schon bei PR-Konzepten die Messbarkeit und Bewertbarkeit herausgearbeitet wird. Operationale Ziele sind hierbei der Schlüssel zu einer besseren Umsetzung des PR-Controlling.[111]

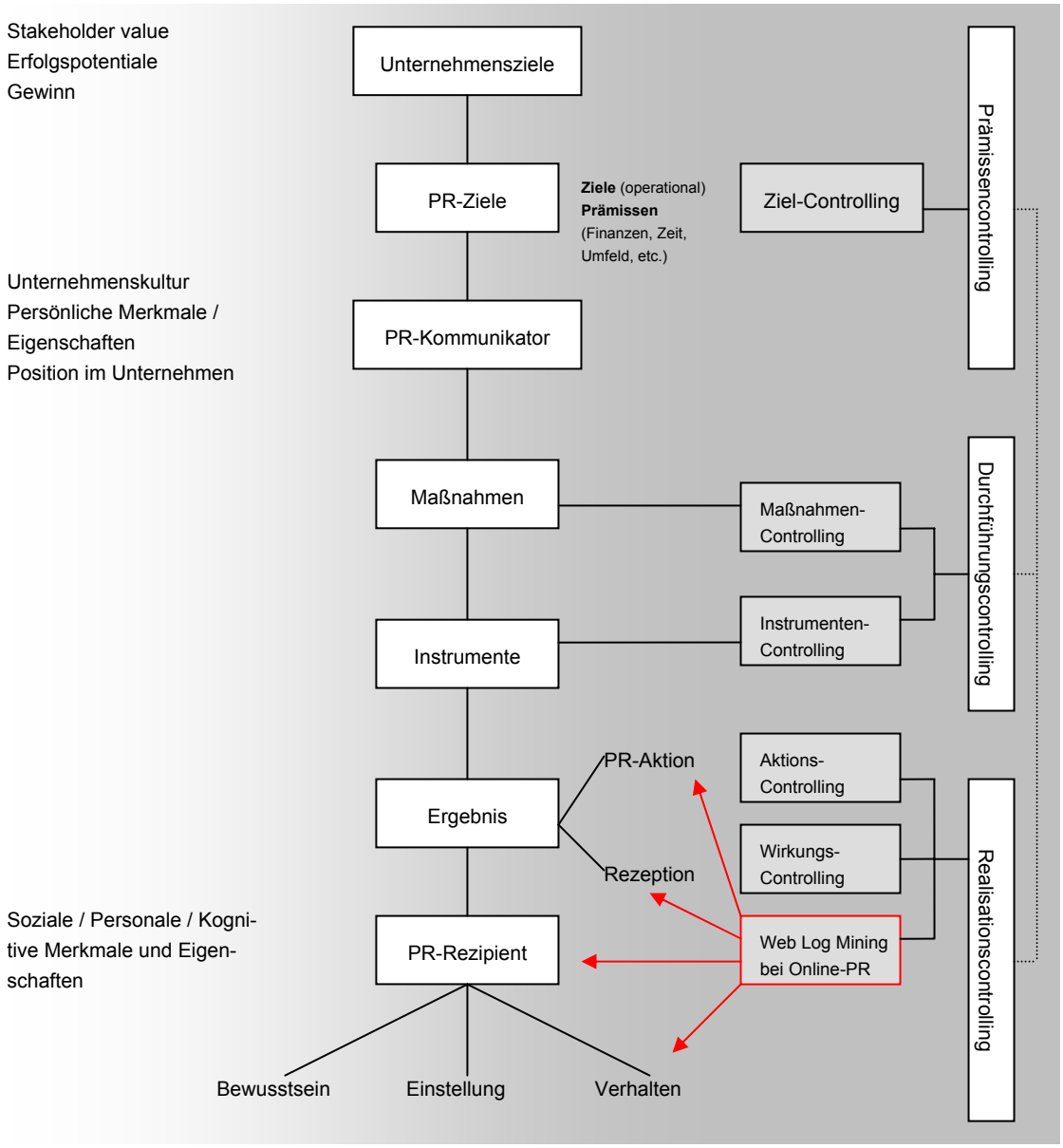

Abbildung 8: PR-Controlling[112]

Die einzelnen Controllingstufen in obiger Grafik sind nicht isoliert voneinander zu betrachten. Es finden Rückkopplungen statt, da sich die Controllingprozesse über Pla-

[111] vgl. Perry, T. (1995), S. 32 f

[112] Aus: Ebert, G. / Steinhübel, V. (1995), S. 14, erweitert durch Web Log Mining

nung und Information aufeinander einwirken.[113] Die Darstellung zeigt demnach keinen vollständig abgebildeten Controllingprozess, sondern Ansatzpunkte einzelner Controllinginstrumente.

3.2.2 Kennzahlen

Kennzahlen sind zentrale Größen, die im Controlling dazu verwendet werden, Messungen und Planungen durchzuführen. In länger etablierten Controlling-Domänen existieren einheitliche Kennzahlensysteme. Kennzahlensysteme bestehen aus einer Reihe von Kennzahlen, die in ihrer Gesamtheit umfassend und konzentriert über einen Sachverhalt informieren.[114] Ein einheitliches Kennzahlensystem für die Auswertung von Website-Aktivitäten existiert derzeit nicht.[115]

Es gilt nicht nur für die Analyse der Website-Aktivität durch Web Log Mining geeignete Kennzahlen festzulegen, sie müssen auch mit Kennzahlen der Online-PR kompatibel sein, beziehungsweise mit ihnen in Verbindung gebracht werden können. Das gemeinsame Medium Internet (bzw. die Website) ist hierbei der gemeinsame Nenner, um diese Kennzahlen zu definieren.

Um geeignete Kennzahlen zu finden, können zum Beispiel folgende Fragestellungen aus PR-Sicht hilfreich sein. Aus Web Log Mining-Sicht müssen dann entsprechend Antworten gefunden werden, um aus einer Fragestellung eine Kennzahl festzulegen.

- Werden die Soll-Zielgruppen erreicht?
  Je eher die im PR-Konzept definierten Zielgruppen auf der Website Inhalte abrufen und je mehr sie mit den tatsächlichen Ist-Zielgruppen übereinstimmen, umso besser werden die Soll-Zielgruppen erreicht.
- Was rufen die Zielgruppen ab?
  Hier kann bestimmt werden, ob alle Informationsangebote in gleichem Maße genutzt werden. Auch eine Aufstellung des nicht berücksichtigten Informationsbedarfs ist denkbar.
- Wann werden Informationen abgerufen?
  Vor allem von Informationen, bei denen es in besonderem Maße auf Aktualität ankommt, ist es wichtig zu wissen, wie zeitnah sie den Empfänger erreichen.
- Von wo aus werden Informationen abgerufen?
  Wenn bekannt ist, welche Nationalität die Website-Besucher haben, kann es

---

[113] vgl. Ebert, G. / Steinhübel, V. (1995), S. 11

[114] vgl. LEGAmedia (2001),
URL: http://www.legamedia.net/lx/result/match/22710757561423d5ad818dd198/index.php,
Zugriff am 19.11.2003

[115] vgl. Schwickert, A. C. / Wendt, P. (2000),
URL: http://wi.uni-giessen.de/gi/dl/showfile/Schwickert/1168/Apap_WI_2000_08.pdf –
Zugriff am 19.11.2003. – S. 1

gegebenenfalls sinnvoll sein, das Informationsangebot um eine Sprache zu erweitern.

- Wie sieht die technische Ausrüstung der Rezipienten aus?
  Hieraus kann ersehen werden, ob die Informationen im richtigen Format angeboten werden, oder ob die technische Ausstattung vieler Rezipienten geringer ist als sie für die Anforderungen der Website nötig wäre.

Diese und weitere Fragestellungen lassen sich recht gut mit Methoden des Web Log Mining beantworten. Dabei kommen sowohl die hypothesengetriebenen Analyseprogramme als auch die datengetriebenen Data Mining-Techniken zum Zuge. Erstere beispielsweise bei der Feststellung, woher die Rezipienten kommen, letztere, um zum Beispiel Ist-Zielgruppen festzustellen.

3.2.3 Erfolgsmessung von Online-PR

Ausgangspunkt jeder Erfolgsmessung sind zuvor definierte und avisierte Ziele. Das sind in diesem Fall Ziele eines PR-Konzepts bzw. einer PR-Strategie, die zumeist eng mit allgemeinen Unternehmenszielen verknüpft sind (vgl. Abbildung 8, S.41). Letztendlich kann die Erreichung einzelner Ziele anhand von Aussagen überprüft werden, die aufgrund ermittelter Kennzahlen getroffen werden können.

Einzelne Kennzahlen sagen etwas darüber aus, ob eine PR-Maßnahme erfolgreich war. Zum Beispiel sind hierfür ein Anstieg der Clickthrough-Rate oder der Visits aussagekräftige, quantitative Indikatoren.[116] Diese lassen allerdings nur begrenzt auf den qualitativen, unternehmerischen Erfolg schließen. Es könnte sein, dass ein Anstieg der Visits dadurch zustande kommt, dass eine Protestwelle gegen konkrete Vorhaben des Unternehmens zu einem Anstieg der Besucherzahlen auf der Website führte.

Qualitativer Erfolg von Public Relations bedeutet, dass die generellen Ziele der PR, das Unternehmen in einem positiven Licht erscheinen zu lassen und Beziehungen zu den verschiedenen Partnern des Unternehmens aufrecht zu erhalten und zu pflegen, erreicht werden. Dies könnte durch Maßnahmen untersucht werden, die konkrete Rückmeldungen oder Bewertungen via Emails oder Umfragen auswerten.[117] Allerdings zählt das dann nicht mehr zum Web Log Mining. Web Log Mining ermöglicht also keine direkte qualitative Erfolgskontrolle der PR, sondern allenfalls eine indirekte qualitative Einschätzung durch die Entwicklung quantitativer Kennzahlen.

Ein weiterer Diskussionspunkt ist die Genauigkeit, die bei der Erfolgsmessung angewendet werden soll. Bürlimann schreibt zum Thema „Genauigkeit am falschen Ort“:

> *„Die Realität zeigt, dass die Verantwortlichen häufig nicht die Nerven haben, es bei einer pragmatischen Untersuchungstiefe zu belassen. Sie*

[116] vgl. Fuchs, P. / Möhrle, H. / Schmidt-Marwede, U. (1999), S. 132
[117] vgl. Fuchs, P. / Möhrle, H. / Schmidt-Marwede, U. (1999), S. 132 f

versuchen oft krampfhaft, Erfolg und Misserfolg bis ins Letzte auszuleuchten.“ [118]

Das darf aber nicht darüber hinwegtäuschen, dass ohne eine absolut einwandfreie Datenbasis jedes Web Log Mining zum Scheitern verurteilt ist, wenn Data Mining betrieben werden soll. Einfache Logfile-Auswertungen (vgl. Tabelle 3 auf Seite 22) sind mit Einschränkungen auch mit geringerer Datenqualität durchführbar.

3.3 PR-spezifisches Web Log Mining

3.3.1 Vergleich von Kosten und Nutzen

Da der Aufwand im Web Log Mining recht hoch werden kann, vor allem gut durchgeführte ETL-Maßnahmen und Data Mining-Methoden erfordern ein hohes Maß an Sorgfalt und Aufwand, muss im Einzelfall abgewogen werden, ob sich intensive Web Log Mining-Analysen lohnen. In der Regel werden nicht allein aus PR-Beweggründen Web Log Mining-Untersuchungen angestellt. Neben den Webmastern und Administratoren einer Website sind beispielsweise auch oft der Vertrieb und die Unternehmensführung selbst an Web Usage Mining-Ergebnissen interessiert.

Bezogen auf den PR-Bereich wird in der Praxis Web Log Mining nur dann durchgeführt werden, wenn die Kosten hierfür auch im Rahmen des bisherigen PR-Budgets stehen. Web Log Mining kann den Informationsbedarf einer (Online-)PR-Abteilung keinesfalls decken, sondern ist ein Instrument, um Informationen über die Website-Nutzung zu gewinnen. Ganz einfache Analysen können annähernd zum Nulltarif erstellt werden („hits“ zu ermitteln erfordert kaum Aufwand), die Aussagekraft dieser einfachen Kennzahlen ist aber auch sehr gering. Je höher der Aufwand für die Ermittlung einer Kennzahl ist, umso höhere Kosten entstehen.

Umfassendes Web Log Mining aus PR-Sicht ist somit nur großen Unternehmen zu empfehlen, mittlere und kleinere Unternehmen werden wahrscheinlich nur dann darauf zurückgreifen, wenn ihre PR-Arbeit hauptsächlich auf deren Website stattfindet, wie dies zum Beispiel bei Firmen der Fall ist, die in den e-Business-Markt eingestiegen sind.

Der größte (und schnellste) Nutzen kann erzielt werden, indem die dringlichsten Fragen und Probleme, die Web Log Mining aufdecken kann, behoben und angegangen werden (beispielsweise ein toter Link, der über das Auftreten eines Fehlercodes entdeckt wird, oder eine Zielgruppenabweichung, die durch eine Clusteranalyse festgestellt wird). Was schon unter der Überschrift „Erfolgsmessung von Online-PR“ erwähnt wurde, gilt auch bei der Abwägung von Kosten und Nutzen: Bei einfacheren Logfile-Analysen genügen Genauigkeiten von 80 bis 90 Prozent, um brauchbare Ergebnisse

[118] Bürlimann, M. (1999), S. 217

zu erhalten.[119] Kommt Data Mining mit ins Spiel, ist Genauigkeit bei der Datenbasis unabdingbar, und der Aufwand und die Kosten steigen erheblich.[120]

3.3.2 Zielgruppenidentifikation

Es ist wichtig zu wissen, ob die Zielgruppen, welche im PR-Konzept als Rezipienten der PR-Strategie vorgesehen sind, auch wirklich erreicht werden. Deshalb müssen diejenigen Personen, die durch eine Online-PR-Kampagne angesprochen werden, identifiziert werden. Dies kann auf verschiedene Weise erreicht werden, zum Beispiel, indem der Website-Besucher befragt wird. Eine häufig verwendete Methode ist es, den Website-Besucher nach seinen Interessen in einem Fragebogen zu befragen, wenn er sich bei einer Webseite ein Benutzerkonto einrichtet. Abgekoppelt von den eigentlichen Identifikationsmerkmalen (wie Name, Adresse) der Person selbst lässt sich so problemlos eine Datenbank aufbauen, welche Angaben über Website-Nutzer enthält, die ein Benutzerkonto besitzen.

Auch eine Online-Meinungsumfrage zum Online-Angebot kann Aufschluss über die Zusammensetzung der Website-Nutzer geben. In Kombination mit einem Gewinnspiel kann so auch direkt ein „Interesse am Kunden“ vermittelt werden.

Eine gute Möglichkeit, Zielgruppen zu identifizieren und gleichzeitig mit den Zielgruppen in Kontakt zu treten, sind Newsletter. Dem Website-Besucher wird angeboten, einen Newsletter zu abbonieren. Dabei gibt der Nutzer Themen an, die ihn interessieren. Im Idealfall orientieren sich diese Themen natürlich an den PR-Inhalten, die das Unternehmen vermitteln will. Aus Datenschutzgründen könnten die „Bestelldaten“ des Newsletters aggregiert werden und für eine Data Mining-Auswertung herangezogen werden – zum Beispiel um eine Klassifizierung der Benutzer anhand ihrer Interessen vorzunehmen.

Einen interessanten Ansatzpunkt könnte die Link- bzw. Menüstruktur innerhalb der Website bieten, um Zielgruppen zu identifizieren. Anhand der Logfiledaten kann festgestellt werden, wie häufig und in welcher Kombination (in einzelnen Benutzerpfaden) Verlinkungen innerhalb der Website genutzt werden. Ein Abgleich der ermittelten tatsächlich Angesprochenen (Ist-Zielgruppen) mit den beabsichtigt Angesprochenen (Soll-Zielgruppen) leistet dann einen wesentlichen Beitrag zum Controlling der PR.

Einen Ansatz zur Bestimmung von Merkmalen der Website-User, ohne diese zu befragen, beschreiben Murray und Durrell in ihrem Artikel „Inferring Demographic Attributes of Anonymous Internet Users“. Sie erstellten ein System, das mit der Kombination aus der Information Retrieval-Methode LSA[121] (Latent Semantic Analysis) und einem Neu-

[119] vgl. Bürlimann, M. (1999), S. 217
[120] vgl. Rapp, R. / Guth, S. (2003), S. 175
[121] Näheres zu LSA siehe Ferber, R. (2003), S. 239

ronalen Netz des SCG-Algorithmus[122] (Scaled Conjugate Gradient-Algorithmus) fähig ist, aus einfachen Nutzungsdaten demographische Aussagen bezüglich der Nutzer zu treffen. Die Idee, ein solches System zu entwerfen, entstand dadurch, dass es für das Schalten von Werbebannern keinen praktikablen Weg gab, zu erkennen, welche demographischen Merkmale ein Nutzer aufweist. Da jede Bannerschaltung (jeder Adview) mit Kosten verbunden ist, galt es, die Bannerschaltungen auf die jeweilige Zielgruppe einzuschränken, um eine höhere Rentabilität bei Bannerschaltungen zu erzielen. Zu Beginn wurden Nutzungsdaten mit Vektoren abgebildet, wobei jeder Nutzungsvorgang als eigenes „Dokument" und somit als eigener Vektor in eine Matrix aufgenommen wurde. Anschließend wurde ein dreischichtiges Neuronales Netz mit Logfile-Beispieldatensätzen trainiert. Zu den Beispieldatensätzen waren die demographischen Angaben der jeweiligen User bekannt. Es wurde auf eine ausgeglichene Datenbasis hinsichtlich demographischer Merkmalsausprägungen geachtet. Erfasst wurden sechs demographische Merkmale: Geschlecht, Altersgruppe, Einkommen, Familienstand, Bildung und Nachkommen in jeweils binärer Form. Das System konnte viele der Merkmale mit 60- bis 80-prozentiger Bestimmtheit bei neuen Nutzungsdaten ermitteln (bei denen die demographischen Angaben nicht von vornherein bekannt waren). Dieses Ergebnis bedeutete einen enormen Fortschritt, weil dadurch die Zielgruppe weitaus häufiger gezielt angesprochen werden konnte (wenn auch nicht in 100% der Fälle), und somit die Verschwendung von Werbemaßnahmen an die falschen Empfänger reduziert werden konnte. Eine Auswertung und Verarbeitung der Nutzungsdaten in Echtzeit war zum Zeitpunkt der Fertigstellung des Artikels noch nicht realisiert, jedoch war dies ein geplanter nächster Entwicklungsschritt dieses Systems. [123]

Dieses Projekt ist ein praktisches Beispiel dafür, wie Data Mining-Methoden im Web Log Mining Informationen aus einer Datenbasis generieren können. Der Hintergrund, dass Werbemaßnahmen in Form von Bannern optimiert werden sollten, ist gut auf den PR-Kontext übertragbar, so dass hier PR-Controlling mittels Web Log Mining hervorragend umgesetzt werden könnte. Bezogen auf Public Relations wäre eine Promotion-Kampagne auf anderen Webseiten realisierbar, Banner, die ein Unternehmen oder die Website eines Unternehmens bewerben, könnten so besser auf die Zielgruppen abgestimmt werden. Eine andere, ebenfalls denkbare Anwendung der eben beschriebenen Technologie von Murray und Durrell wäre, auf komplexeren Webseiten, wie die eines TV-Senders, dem Website-Besucher an unterschiedlichen Stellen Links und Hinweise zu geben, die auf seine Zielgruppe abgestimmt sind. Zum Beispiel einen News-Bereich am Rand der Website, der dynamisch zielgruppengerechte Schlagzeilen bereitstellt.

---

[122] Näheres zu SCG siehe Felzer, T. / Heidger, A. / Wiesiollek, M. (2002) – URL: http://www.st.informatik.tu-darmstadt.de:8080/felzer/nn_sem.pdf – Zugriff am 13.11.2003. – S. 45 ff

[123] Murray, D. / Durrell, K. (2000), S. 7 ff

4 Praktische Möglichkeiten der Umsetzung und Vorteile für die PR

Der prozessuale Charakter, durch den Data Mining in den Kontext des Knowledge Discovery in Databases eingebunden ist, kommt auch beim Web Log Mining zum Tragen. Der Ablauf des Web Log Mining ist in einzelne Schritte einteilbar (wie dies in Kapitel 2 auf Seite 14 dargestellt ist). Zu Beginn steht eine Zieldefinition, die aussagt, was untersucht werden soll, und welche Untersuchungstiefe angestrebt wird. Damit wird auch zu Beginn festgelegt, welche Methoden zum Einsatz kommen werden. Die evaluierten und interpretierten Ergebnisse fließen auch beim Web Log Mining als Informationen dem Domänenwissen (dem Fachwissen der Public Relations und des Controllings) zu.

Diese enge Bindung an die Fachabteilungen legt eine Einbindung des Web Log Mining in andere Strukturen und Prozesse nahe. Im Idealfall kann Web Log Mining in ein Controllingkonzept integriert werden, wie dies in Abbildung 8 auf Seite 41 dargestellt wurde. Was diese Abbildung aber auch deutlich zeigt: Web Log Mining kann nicht *das*, sondern nur *ein* Instrument für ein PR-Controlling sein, da es nicht alle wichtigen Felder der PR abdecken kann (es ist auf Online-PR zugeschnitten und untersucht dort die Nutzung des Kommunikationskanals „Website“).

Die technischen Vorraussetzungen für ein effektives und effizientes Web Log Mining können im Einzelfall unterschiedlich sein. Dies liegt daran, dass auf geeignete vorhandene Website-Strukturen nur in manchen Fällen zurückgegriffen werden kann. Beispielsweise bietet eine Webserverarchitektur eines großen Online-Shops mit Anbindung an ein ERP-System (Enterprise Resource Planning-System) bessere Vorraussetzungen komplexe Web Mining-Schnittstellen einzurichten, als ein einfacher Webauftritt ohne Anbindung an andere Informationssysteme. Je integrierter in andere beteiligte Systeme Web Log Mining betrieben wird, umso eher sind auch Realtime-Auswertungen möglich. Realtime-Auswertungen ermöglichen das Analysieren und Betrachten aktueller Bewegungen auf der Website und stellen eine Vorraussetzung dafür dar, direkt auf Aktionen der Website-Besucher reagieren zu können. Direktes Reagieren auf Benutzerbewegungen auf der Website ist zum Beispiel bei Werbestrategien von Vorteil, da nach den Wünschen des Users entsprechende Werbebotschaften gesendet werden könnten, ohne den User selbst nach seinen Präferenzen zu fragen. Allein sein Navigieren auf der Website könnte ausreichend Informationen darüber liefern, für welche Dinge er sich interessiert.

Ein weiterer wichtiger Punkt ist, dass für eine kompetente Analyse von Web-Logfiles die Struktur der Website bekannt sein muss (am Besten in Form von Metadaten), damit eine Interpretation der Web Log-Einträge möglich ist. Pfadbeziehungen, Berechtigungsebenen, Stylesheets und das Wissen über Änderungen an der Website sind nur einige Gründe dafür, dass Web Log Miner (diejenigen, welche das Web Log Mining

durchführen) wissen sollten, wie der Web-Auftritt strukturiert ist, aufgebaut ist und wie er funktioniert.

Je nach Abteilung oder Sichtweise werden aber die Vorteile des Web Log Mining meist anders gewichtet.

- Web-Admins und Content-Manager legen Wert auf Auswertungen über die technischen Vorraussetzungen der Besucher: Die Browserversion und das Betriebssystem. Weiter ist für sie besonders interessant, Fehlercodes zu analysieren, da diese Aufschluss über fehlgeschlagene Aufrufe liefern.
- Die PR-Abteilung und die PR-Entscheider haben ein Interesse an der Zielgruppenidentifikation. Sie wollen wissen, wie viele Kunden die Website besucht haben und ob PR-Inhalte die Empfänger erreicht haben.
- Die Unternehmensführung ist an wirtschaftlichen Fakten interessiert. Hierbei ist von Interesse, ob der Web-Auftritt sich wirtschaftlich lohnt und ob beispielsweise der Einsatz des Mediums Website für eine PR-Kampagne profitabel war („Führte die PR-Kampagne zu einer Imageverbesserung oder zu einem Kundenzuwachs?").

Im Folgenden wird nun erörtert, für welche PR-spezifischen Inhalte sich die Methoden des Web Log Mining besonders eignen. Hierbei wird, wie in Kapitel 2, zwischen Logfile-Kennzahlen und Data Mining-Verfahren unterschieden, allerdings werden hier nun die Belange der PR näher erklärt, weniger der technische Hintergrund der Logfile-Kennzahlen.

4.1 Exemplarische Untersuchungen

In Anlehnung an die Zieldefinition einer Web Log Mining-Untersuchung gilt es festzustellen, welche PR-relevanten Sachverhalte analysiert werden sollen. Hier werden sechs Fragen herausgegriffen, denen Logfile-Kennzahlen und Data Mining-Techniken zugeordnet werden.

- *Wer ist an der Website an sich interessiert?*
  Wird dieser Frage nachgegangen, können auch ähnliche Fragestellungen bezüglich der erreichten Zielgruppen mit beantwortet werden (z. B.: *„Wie setzen sich die Zielgruppen zusammen?", „Sind hauptsächlich neue Besucher oder bereits bekannte Besucher auf der Website?"*). Sollen möglichst viele Details über die Besucher bekannt werden, (der Begriff „wer" kann relativ weit gefasst werden), so können gleich einige Datenfelder des Logfiles interessant sein. Sehr wichtig ist das Feld „Remotehost". Mit einer Auswertung der zugreifenden IP-Adressen lassen sich vor allem größere Firmen und damit deren Interesse identifizieren, da sie zusammenhängende IP-Adressbereiche nutzen. Das Logfile-Feld „Agent" gibt Auskunft über die technische Ausstattung der Website-Besucher. Dies lässt in gewissem Umfang auch Rückschlüsse auf andere demographische Merkmale der Besucher zu. Auch Data Mining-Verfahren können

Antworten auf die gestellte Frage geben. Eine Zielgruppenidentifikation, wie sie in Kapitel 3.3.2 auf Seite 45 beschrieben wurde, ist ein eindrucksvolles Beispiel dafür, wie mit Methoden des Information Retrieval und des Data Mining eine Einschätzung über demographische Attribute der Besucher vorgenommen werden kann. Weiterhin sind die Nationalitäten der Besucher je nach Fragestellung sehr interessant. Auch sie können über „Remotehost“ und entsprechende Zuordnungen der IP-Adressen zu Providern und Firmen ermittelt werden. Wird dabei festgestellt, dass häufig Zugriffe aus einem Land stattfinden, das noch nicht in ausreichendem Maße auf der Website sprachlich berücksichtigt wird, kann dies ein Anstoß zur Überarbeitung der Website sein. Ist ein Subskriptionsmodell in der Website integriert, können zusätzlich Analysen der Besuche registrierter Nutzer (unter Beachtung des Datenschutzes) angestellt werden.

- *Wird die Website regelmäßig oder eher unregelmäßig besucht?*
  Neben registrierten Nutzern, die sich recht leicht ermitteln lassen, können durch intensivere Auswertungen unter anderem Userprofile ermittelt werden. Bei der Hinzunahme der Felder „Date“ und „Timezone“ ist es dann möglich, bei wiederkehrenden, identifizierten Usern eine eventuelle Regelmäßigkeit der Besuche dieser Website festzustellen. Die Identifizierung der User würde durch den Einsatz von Cookies oder Session-ID’s auf der Website erleichtert, machen die Auswertung aber nicht weniger kompliziert, weil sie nicht immer erfolgreich eingesetzt werden können.

- *Werden PR-spezifische Informationen (wie z. B. Pressemitteilungen) auch gelesen?*
  Eine Auswertung des Feldes „Request“ bringt an den Tag, welche Seiten der Internetpräsenz besonders oft aufgerufen werden. Eine Analyse der „Sessions“ kann außerdem die Dauer der Rezeption in die Analyse mit einbeziehen. Für eine Analyse ist es aber erforderlich, die Struktur der Website zu kennen, und zu wissen, welche Dateien und Inhalte PR-relevant oder PR-spezifisch sind. Das erfordert die Zusammenarbeit der PR-Abteilung mit den Administratoren der Website.

- *Nutzen Journalisten in größerem Umfang das PR-Angebot?*
  Um diese Frage zu beantworten, ist es von Vorteil, wenn (allgemein oder für den Pressebereich) ein Subskriptionsverfahren auf der Website verwendet wird. Anhand der registrierten Benutzer kann so festgestellt werden, wie viel Prozent der registrierten Nutzer regelmäßig wiederkehren. Aber auch die Analyse des Felds „Remotehost“ kann eventuell Aufschluss darüber geben, ob häufig Zugriffe von größeren Verlagen (und somit von Journalisten) erfolgen.

- *Unter welchen Suchbegriffen wird die Website in Suchmaschinen gefunden?*
  Im Feld „Referrer“ finden sich die in einer Suchmaschine eingesetzten Suchbegriffe, wenn zuvor eine Suchmaschine besucht wurde und dort ein über Suchbegriffe erzieltes Ergebnis den Link zu der Seite bot und genutzt wurde. Neben der Information, dass die Website über die jeweilige Suchmaschine ge-

funden werden kann, wird so auch bekannt, welche Suchbegriffe Interessierte mit dem Web-Angebot assoziieren.

- *Wie schnell erreichen neue Inhalte die Rezipienten?*
  Anhand der Felder „Request", „Date" und „Timezone", sowie einem Überblick über das Erscheinungsdatum neuer Seitenbereiche, einzelner Downloadangebote oder Artikel auf der Website, kann festgestellt werden, wie zeitnah und in welchem Umfang neue Inhalte den Benutzer erreichen. Hierfür bietet es sich an, Zeitabstände festzulegen, welche die Neuheit eines Inhalts beschreiben. Zum Beispiel könnte ein Zeitraum bis 24 Stunden nach Veröffentlichung die Bezeichnung „hochaktuell" erhalten, der daran anschließende Zeitraum bis zu drei Tagen die Bezeichnung „aktuell", und so weiter. Eine Aufstellung, wie viele Aufrufe Texte in ihrer jeweiligen Phase erhalten, lässt eine Einschätzung zu, wie zeitnah neue Inhalte rezipiert werden. Diese Untersuchung kann auch sehr wertvoll für Kampagnen der Krisen-PR sein, bei der es stets darauf ankommt, schnell und punktgenau zu informieren.[124] Ein Wissen darum, wie ein typischer Rezeptionszyklus von Nachrichten ist, gibt in einem solchen Fall Klarheit darüber, wann die Nachricht ankommt. Mit dieser Methode kann auch gemessen werden, wann und in welchem Umfang (bezüglich der Zielgruppe) die Krisen-Nachricht selbst angekommen ist.

Die theoretische Beantwortung der obigen exemplarischen Fragen verdeutlicht, dass Logfile-Kennzahlen und die Möglichkeiten des Data Mining einen engen Bezug zur Analyse PR-relevanter Themen bieten. Sicherlich gibt es eine Reihe weiterer Punkte im PR-Interesse, die sich ähnlich gut untersuchen lassen.

4.2 Grenzen von Web Log Mining bei Online-PR

Trotz der zahlreichen Auswertungsmöglichkeiten, von denen eben einige beschrieben wurden, kann Web Log Mining nicht alle Fragestellungen eines umfassenden Online-PR-Controllings abdecken. Dies ist ein weiterer Grund, weshalb Web Log Mining Teil eines umfassenden Controlling-Konzeptes sein sollte, und nicht „isoliert" angewendet werden sollte.

Im Gegensatz zur Akzeptanz der Website durch die angesprochenen Zielgruppen, die durch Kennzahlen wie der Clickthrough-Rate oder der Sessiondauer ermittelt werden kann, ist eine Beurteilung der Bewertung durch die Benutzer nicht allein durch Web Log Mining zu bewerkstelligen. Eine solche Beurteilung kann mit Hilfe von Umfragen ermittelt werden. Diese sollten sich nicht auf Online-Umfragen beschränken, da unter Umständen gerade diejenigen, die große Vorbehalte gegen den Web-Auftritt in seiner jetzigen Form haben, online nur schwer befragt werden können, weil sie den Kontakt mit der Website meiden.

---

[124] vgl. Sauvant, N. (2002), S. 180 f

Zudem gestaltet sich der Vergleich verschiedener durchgeführter Web Log-Untersuchungen einer Website als schwierig. Da sich – in aller Regel – das Web-Angebot verändert, und sich externe sowie interne Rahmenbedingungen wie wirtschaftliche Lage oder das Produktspektrum des Unternehmens geändert haben können, sollte bei einem Vergleich, genau so wie bei der Bewertung einzelner Untersuchungen, darauf geachtet werden, welche externen und internen Variablen zur Veränderung beigetragen haben könnten. Sinn und Zweck eines Vergleichs verschiedener Web Log-Untersuchungen ist es, zu beurteilen, ob die Maßnahmen, die aufgrund von Erkenntnissen zurückliegender Web Log-Untersuchungen getroffen wurden, zu einer Verbesserung der Effizienz und Effektivität beigetragen haben. Um dies richtig beurteilen zu können, ist das Berücksichtigen von externen und internen Einflüssen, die nicht mit den getroffenen Verbesserungsmaßnahmen in Verbindung stehen, nötig.

4.3 PR-Nutzen

Der konkrete Nutzen, den Web Log Mining der PR bieten kann, schlägt sich in verschiedenen Punkten nieder, die im Folgenden aufgezählt werden.

Die Zielgruppenansprache über das Internet kann besser abgestimmt werden. Web Log Mining bietet Möglichkeiten zur Analyse der Zielgruppen, welche die Website besuchen. Der Besucher der Website kann von der direkteren Ansprache ebenfalls profitieren. Die Steuerung dynamischer Webseiten, welche die präsentierten Inhalte vom Verhalten des Besuchers abhängig macht, kann ideal durch Web Log Mining ergänzt werden.

Gezieltere Ansprache und die Auswertungsmöglichkeiten über genutzte und nicht genutzte Bereiche einer Website bieten ein hohes Potenzial an Einsparungsmöglichkeiten. Beispielsweise kann der Einsatz von Werbebannern für Promotionaktionen im Rahmen der Onilne-PR besser gesteuert werden, wenn mehr über demographische Merkmale der Rezipienten bekannt ist. Web Log Mining kann dazu eingesetzt werden, die demographischen Merkmale über das Nutzerverhalten zu ermitteln.[125] Wird festgestellt, dass der Bekanntheits- oder Nutzungsgrad der Website bei einer oder mehreren Zielgruppen nicht den Erwartungen entspricht, kann durch eine entsprechende Kampagne die Website gezielt promotet werden.

Die Kenntnis über den Grad der Nutzung der Website durch die Besucher hebt das Medium Internet als PR-Kommunikationsmedium auf eine höhere Akzeptanzebene bei Kritikern, da die Nutzung des Angebotes transparent wird.

Durch Web Log Mining kann auf Veränderungen im Umfeld der Online-PR schnell reagiert werden. Seitenbesuche, die über Suchmaschinentreffer zustande kommen, Zielgruppenveränderungen, Suchbegriffe auf der Website-internen Suchfunktion und Abru-

[125] vgl. Murray, D. / Durrell, K. (2000), S. 7 ff

fe von Informationsmaterial lassen einen schnellen Überblick über vorhandene und veränderte Interessen zu.

Mit einer Analyse der Logdateien kann auch die Website in ihrer Struktur verbessert werden. Der Verwertung der Ergebnisse (Interpretation und Umsetzung) kann zu deutlichen Verbesserungen auf der Website führen. Folgende Verbesserungen an einer Website lassen sich mithilfe von Web Log Mining beispielsweise anstreben:

- Navigationspfade können optimiert werden. Wird festgestellt, dass viele Website-Besucher sich über mehrere Seiten zum den gewünschten Inhalten „durchklicken“, kann Ihnen durch eine Optimierung der Navigationspfade der Inhalt direkt angeboten werden.
- Werden große Dateien oft abgefragt, können sie im Sinne einer Verbesserung für den User mehrfach oder in kleinerer Version angeboten werden.
- Sucht der Besucher nach Informationen, die ihm noch nicht angeboten werden (dies kann über eine Website-interne Suchfunktion festgestellt werden), kann dies als Ansatzpunkt für eine Erweiterung des Online-Angebots genutzt werden.

Bei Zielgruppen-Veränderungen kann schneller reagiert werden, das Angebot der Website an die Zielgruppen anzupassen. Sowohl Veränderungen bei Soll-Zielgruppen (wie sie in den PR-Zielen vorkommen) als auch Veränderungen bei Ist-Zielgruppen, die bei Web Mining-Untersuchungen bestimmt werden können, sind hiervon betroffen.

5 Fazit

Es wurde gezeigt, dass Logfile-Analysen zahlreiche Kennzahlen zur Beurteilung der Nutzung von Internetpräsenzen bereitstellen. Diese können von verschiedenen Nutzergruppen (Web-Administratoren, Marketing-Abteilung, Unternehmensführung, etc.) zur Kontrolle und für Verbesserungen beim Einsatz des Online-Mediums Internet eingesetzt werden. Die Domäne der (Online-)Public Relations kann sowohl aus Logfile-Analysen als auch mit dem Einsatz von Data Mining-Methoden wichtige und wertvolle Informationen gewinnen, um die Zielgruppenansprache zu verbessern, Inhalte schneller und zielgerichteter zu vermitteln, und weitere Aspekte der PR-Arbeit zu optimieren.

Web Log Mining ist als Controllinginstrument für die (Online-)Public Relation aus mehreren Gründen gut geeignet. Der prozessuale Ansatz, abgeleitet aus dem Vorgehen bei Knowledge Discovery in Databases, ermöglicht ein Einbinden des Web Log Mining in vorhandene Controllingprozesse und –systeme mit einer minimierten Anzahl an Medienbrüchen (im Idealfall gar keinen).

Je besser Web Log Mining an andere Systeme angebunden werden kann, umso höher ist auch der Nutzen für das gesamte Unternehmen. Die Ergebnisse könnten in gegebenenfalls vorhandenen Reporting-Systemen bereitgestellt werden. Web Log Mining liefert Kennzahlen, welche strategische Entscheidungen der PR unterstützen können, beispielsweise, ob eine Website als Teil des Instrumentenmix weiter ausgebaut werden soll, oder nicht. Der hypothesenfreie Ansatz des Data Mining eignet sich auch für Aufgaben der Planung und Steuerung, wenn es zum Beispiel darum geht, weitgehend unbekannten Website-Besucher allein anhand ihres Online-Verhaltens bestimmte Merkmale zuzuordnen.[126]

Bei allen Möglichkeiten, durch Logfiles Informationen über Website-Besucher, Zielgruppen und Kunden zu erfahren, darf der Datenschutz nicht vernachlässigt werden. Personenbezogene Daten dürfen ohne das Einverständnis der Betroffenen weder erhoben, noch weiterverarbeitet werden. Entweder muss der Personenbezug entfernt werden, zum Beispiel durch Pseudonymisierung, oder das Einverständnis der Betroffenen muss vorliegen, beispielsweise durch eine Zustimmung bei der Registrierung eines Benutzeraccounts für die Website.

Als zukünftige Entwicklung wird angenommen, dass Data Mining-Technologie künftig auch in Webserverpakete integriert werden wird. Dadurch könnten eBusiness-Aktivitäten sinnvoll ergänzt und direkt unterstützt werden, da direkt aus der Webserverumgebung heraus Data Mining-Untersuchungen betrieben werden könnten.[127]

---

[126] vgl. Murray, D. / Durrell, K. (2000), S. 7 ff und Kapitel 3.3.2: Zielgruppenidentifikation, S. 45 f

[127] vgl. Mena, J. (2000), S. 443

Ein weiteres Ziel künftiger Systeme ist die Fähigkeit, Daten in Echtzeit auszuwerten. Bei Logfile-Analysen ist dies heute schon möglich, bei Data Mining-Techniken noch längst nicht selbstverständlich. Ein Grund hierfür ist, dass Data Mining-Produkte häufig noch nicht ausgereift sind und selten einen so großen Bedienungskomfort bieten wie andere Softwareprodukte.[128] Echtzeit-Auswertungen könnten die Interaktivität von Websites noch weiter steigern, indem Website-Besuchern gezielt PR-Inhalte (Unternehmensinformationen, Pressemitteilungen, Produktinformationen und ähnliches) angeboten werden.

Die verzahnte Verarbeitung von Logfiles einer Website und PR-Material erfordert eine enge Zusammenarbeit aller Beteiligten. Weder PR-Verantwortliche noch die Administratoren der Website können ohne das Wissen der anderen Disziplinen effektive Analysen durchführen. Dies gilt besonders für Data Mining-Aufgaben: *„Das Data Mining von Websites ist zudem ein interdisziplinäres Fach, das fächerübergreifendes Know-How benötigt. […]“.* [129]

Damit ist Web Log Mining eine Technik, die sich als Controllinginstrument dann sinnvoll einsetzen lässt, wenn sie in vorhandene Controllingprozesse eingebunden wird und von allen beteiligten Abteilungen unterstützt wird.

[128] vgl. Mena, J. (2000), S. 443 f
[129] Mena, J. (2000), S. 444

Anhang A: Auszug aus einem Logfile

Tabelle 5: Exemplarische Logfile-Einträge[130]

Remotehost	Ident	A.User	Date, Timezone		
Request				**Code**	**Bytes**
p50918db7.dip0.t-ipconnect.de	-	-	[14/Sep/2003:05:06:57 +0200]		
"GET /%7Exx32/tid al/gifs/k1_read.gif HTTP/1.1"				304	0
p50918db7.dip0.t-ipconnect.de	-	-	[14/Sep/2003:05:06:57 +0200]		
"GET /%7Exx32/tid al/gifs/k2_mp.gif HTTP/1.1"				304	0
p50918db7.dip0.t-ipconnect.de	-	-	[14/Sep/2003:05:07:32 +0200]		
"GET /%7xx32/tidal/sound/tidal%20-%20warten%20auf%20godot.mp3 HTTP/1.1"				206	5378823
pd9eb2f1a.dip.t-dialin.net	-	-	[14/Sep/2003:05:09:27 +0200]		
"GET /~xx13/images/descent_extra.gif HTTP/1.1"				200	15223
200_215_86.uio.satnet.net	-	-	[14/Sep/2003:05:09:28 +0200]		
"GET /style.css HTTP/1.0"				200	2148
6191124168.mad.wi.charter.com	-	-	[14/Sep/2003:05:09:42 +0200]		
"GET /~xx19/files/fire.jpg HTTP/1.1"				200	6577
80.146.212.98	-	-	[14/Sep/2003:05:10:01 +0200]		
"GET /mw HTTP/1.0"				301	422
200_215_86.uio.satnet.net	-	-	[14/Sep/2003:05:10:06 +0200]		
"GET /view_news?ident=news20030827143215 HTTP/1.0"				200	34279
200_215_86.uio.satnet.net	-	-	[14/Sep/2003:05:10:12 +0200]		
"GET /bilder_navigation/zurueck.gif HTTP/1.0"				200	900
200_215_86.uio.satnet.net	-	-	[14/Sep/2003:05:10:12 +0200]		
"GET /bilder_navigation/mailsenden HTTP/1.0"				200	923
208104.n1.vanderbilt.edu	-	-	[14/Sep/2003:05:59:13 +0200]		
"GET /~xx05/images/rueckblick.jpg HTTP/1.1"				200	1925
208104.n1.vanderbilt.edu	-	-	[14/Sep/2003:05:59:13 +0200]		
"GET /~xx05/images/home.gif HTTP/1.1"				200	716
208104.n1.vanderbilt.edu	-	-	[14/Sep/2003:06:00:51 +0200]		
"GET /~xx05/Navigation_links/nav_studienzeit.htm HTTP/1.1"				200	3323
20.103.132.92	-	-	[14/Sep/2003:06:10:40 +0200]		
"GET /~xx89/pdf/Guia_001-038.pdf HTTP/1.1"				200	1158601

[130] Das Logfile stammt von der Website www.hdm-stuttgart.de aus dem Zeitraum 14.09.2003 bis 19.10.2003 und wurde aus Datenschutzgründen von personenbezogenen Daten bereinigt. Diese 14 Einträge in Tabelle 5 sind einige der ersten Einträge von 4.688.248 Hits (Logfileeinträgen) im Auswertungszeitraum.

Anhang B: Grafische Darstellungen

Die folgenden Abbildungen sind Auswertungen eines Logfiles der Hochschul-Website www.hdm-stuttgart.de vom Zeitraum 14.09.2003 bis 19.10.2003. Die Auswertung wurden mit der Software AccessLog 5.01[131] erstellt.

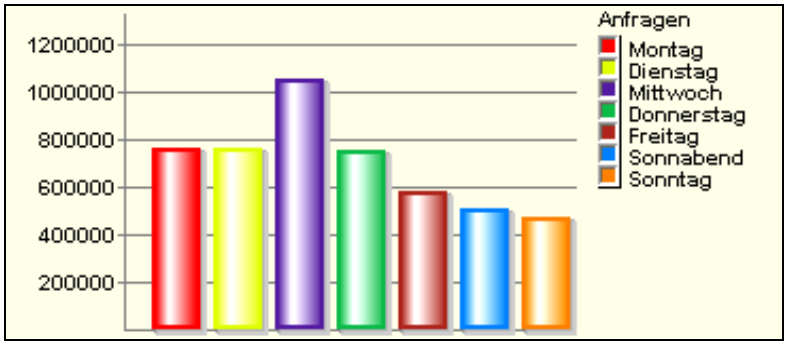

Abbildung 9: Zugriffszahlen auf Wochentage kumuliert

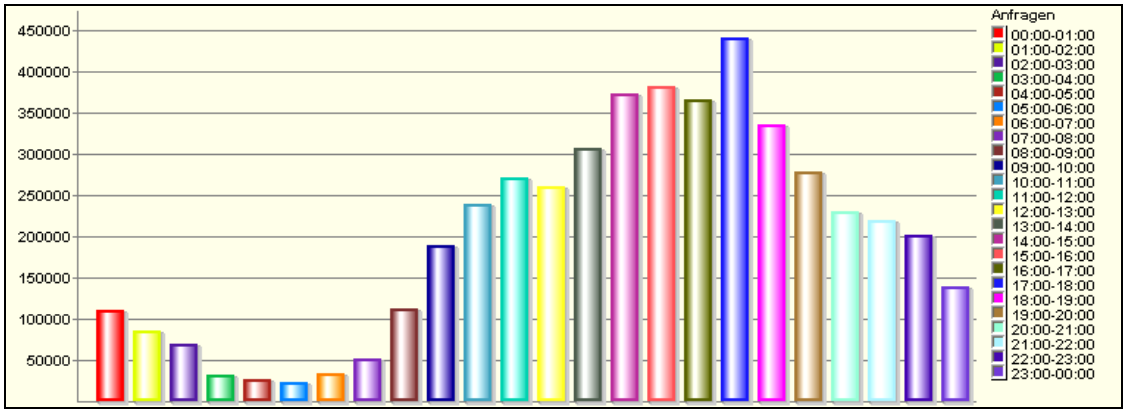

Abbildung 10: Anzahl Zugriffe auf Tageszeiten kumuliert

[131] Siehe: Frenz, Volker (2002). URL: http://www.accesslog.de/ – Zugriff am 05.12.2003.

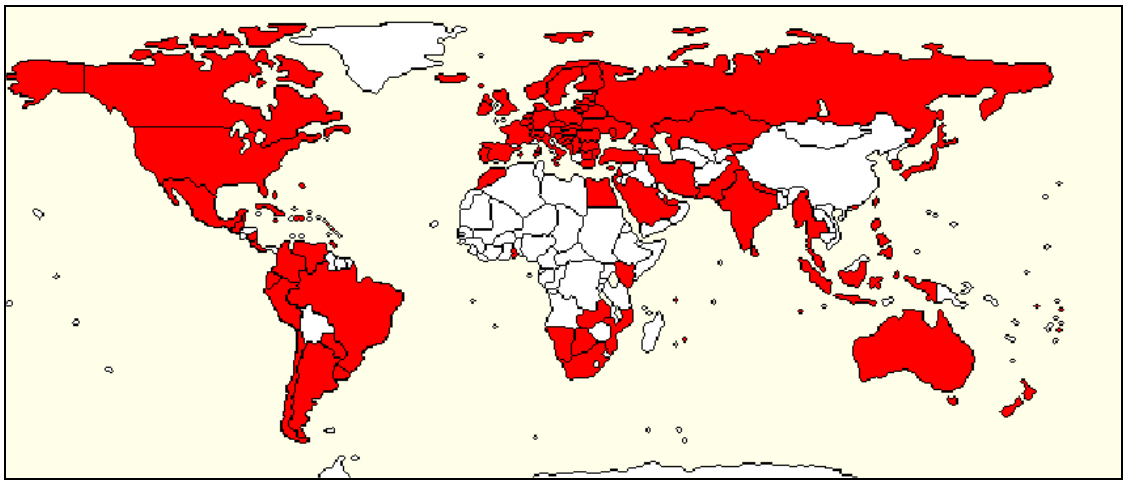

Abbildung 11: Geografische Herkunft der Website-Zugriffe auf der Weltkarte dargestellt

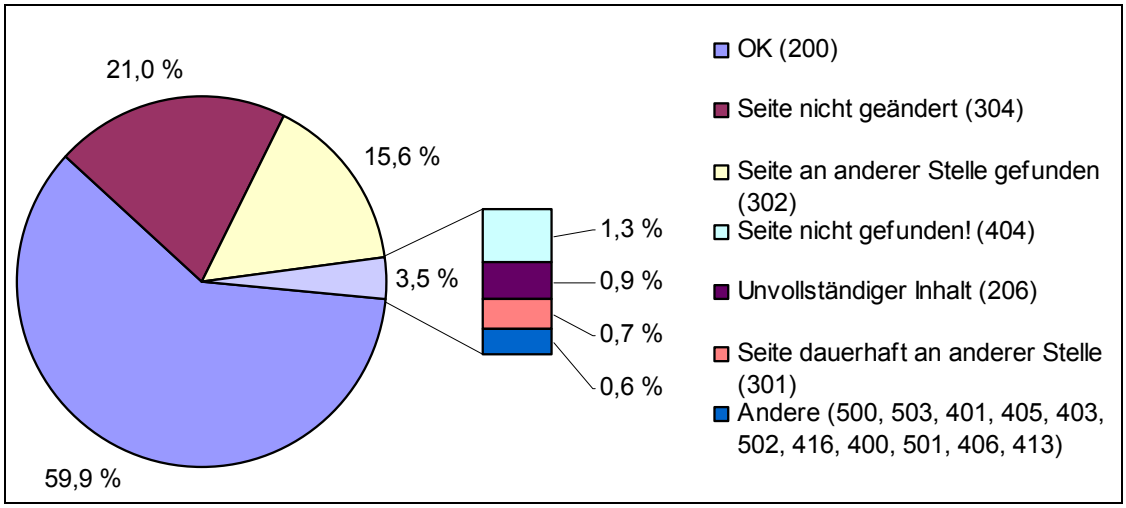

Abbildung 12: Häufigste Status Code-Meldungen im Auswertungszeitraum

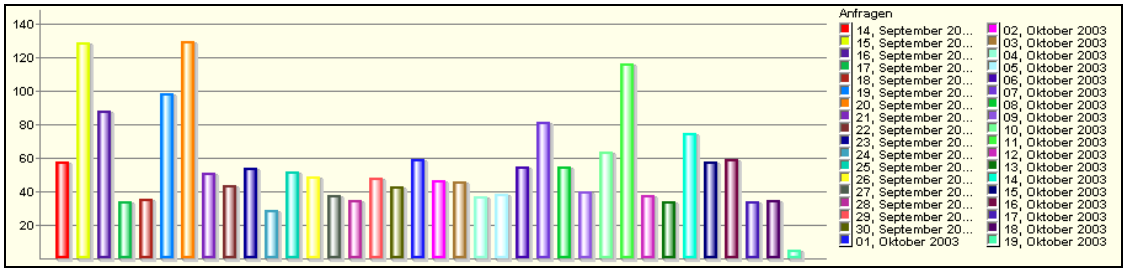

Abbildung 13: Anzahl Downloads nach Tagen geordnet

Anhang C: HTTP Status Codes[132]

Tabelle 6: HTTP Status Codes nach HTTP 1.1

Code	Bezeichnung	Kurz-Beschreibung[133]
1xx	***Informativ (informational)***	
100	Continue	Der Client sollte seine Anforderung fortsetzen, der Server muss die Übertragung bestätigen.
101	Switching Protocols	Der Server stimmt einer Protokollwechselanfrage des Clients zu.
2xx	***Erfolgsmeldungen (successful)***	
200	OK	Die Anfrage führte zum Erfolg. Die übertragene Information hängt von der Ü.-Methode ab.
201	Created	Begründet durch eine Anfrage wurde eine neue Ressource erfolgreich erstellt.
202	Accepted	Eine Anfrage befindet sich in der Verarbeitung, ist aber noch nicht abgeschlossen.
203	Non-Authoritative Information	Metainformationen sind nicht vom Originalserver.
204	No Content	Der Server kann neue oder aktualisierte Metainformationen senden, keinen Content.
205	Reset Content	Der Server hat die Anfrage ausgeführt und der Client sollte die Anzeige aktualisieren.
206	Partial Content	Der Server hat einen partiellen Request über die Methode GET beantwortet.
3xx	***Weiterleitungen (redirection)***	
300	Multiple Choices	Die angeforderte Ressource ist an mehreren Stellen verfügbar.
301	Moved Permanently	Die angefragte Ressource ist ab sofort unter anderer (mit übermittelter) URL zu erreichen.
302	Found	Die angefragte Ressource ist temporär unter anderer (mit übermittelter) URL zu erreichen.
303	See Other	Der Client wird angewiesen, an einer anderen Stelle anzufragen.
304	Not Modified	Die Ressource wurde nicht verändert seit dem letzten Aufruf.
305	Use Proxy	Die angefragte Ressource muss über den angegebenen Proxy bezogen werden.
306	(Unused)	-
307	Temporary Redirect	Die angefragte Ressource ist temporär an anderer Stelle gespeichert.
4xx	***Client-Fehler (client error)***	
400	Bad Request	Die Anfrage konnte wegen eines Fehlers in der Anfrage nicht bearbeitet werden.
401	Unauthorized	Es fehlt die Berechtigung, auf eine Ressource zuzugreifen.
402	Payment Required	Ein kostenpflichtiges Angebot wurde angefordert.
403	Forbidden	Der Zugriff wurde unterbunden (selbst bei vorliegender Identifikation).
404	Not Found	Die angefragte Ressource konnte nicht gefunden werden.
405	Method Not Allowed	Die verwendete Methode ist nicht erlaubt.
406	Not Acceptable	Die Anfrage kann nicht akzeptiert werden.
407	Proxy Authentication Required	Für die Verwendung des Proxy-Servers ist eine Authentifizierung erforderlich.
408	Request Timeout	Der Client hat zu lange keine Anfrage gestellt und der Server hat das Warten eingestellt.
409	Conflict	Die Anforderung kann nicht erfüllt werden aufgrund des aktuellen Zustands der Ressource.
410	Gone	Die angefragte Ressource ist nicht mehr verfügbar.
411	Length Required	Der Server verweigert die Bearbeitung ohne die Angabe der Länge des Inhalts.
412	Precondition Failed	Vorbedingung ist nicht erfüllt.
413	Request Entity Too Large	Die angefragte Ressource ist zu groß um vom Server verarbeitet werden zu können.
414	Request-URL Too Long	Die Angefragte URL (Adresse) ist zu lang, um vom Server verarbeitet werden zu können.
415	Unsupported Media Type	Das Format der angefragten Ressource wird nicht unterstützt.
416	Requested Range Not Satisfiable	Dem angeforderten Bereich kann nicht entsprochen werden.
417	Expectation Failed	Ein erwarteter Zustand ist nicht eingetroffen.
5xx	***Server-Fehler (server error)***	
500	Internal Server Error	Der Server konnte die Anfrage nicht bearbeiten, weil ein interner Fehler aufgetreten ist.
501	Not Implemented	Der Server unterstützt die angeforderte Funktion nicht.
502	Bad Gateway	Ein anderer Server gab eine ungültige Antwort.
503	Service Unavailable	Der Server kann derzeit keine Anfragen bearbeiten.
504	Gateway Timeout	Ein Gateway oder Proxy musste zu lange auf Daten von einem anderen Server warten.
505	HTTP Version Not Supported	Die HTTP-Version wird vom Server nicht unterstützt.

[132] vgl. W3C (1999), http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html#sec10 Zugriff am 12.10.2003

[133] Die Kurz-Beschreibungen entsprechen nicht der kompletten Spezifikation des W3C, sondern sollen einen Eindruck von der praktischen Verwendung der Status Codes vermitteln.

Literaturverzeichnis

Monographien und Zeitschriftenartikel

Arbeitskreis Evaluation der GPRA (1997): Evaluation von Public Relations : Dokumentation einer Fachtagung. Frankfurt am Main, Institut für Medienentwicklung und Kommunikation GmbH in der Verlagsgruppe Frankfurter Allgemeine Zeitung GmbH (IMK), 1997.

Arndt, D. / Koch, D. (2002): Datenschutz im Web Mining – Rechtliche Aspekte im Umgang mit den Nutzerdaten, S. 76-103, Erschienen in: Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-b)

Bäumler, H. / Mutius, A. von (Hrsg) (2003): Anonymität im Internet : Grundlagen, Methoden und Tools zur Realisierung eines Grundrechts. 1. Aufl., Braunschweig, Wiesbaden: Vieweg, 2003.

Bensberg, F. (2001): Web log mining als Instrument der Marketingforschung: ein systemgestaltender Ansatz für internetbasierte Märkte. 1. Aufl., Wiesbaden: Dt. Univ.-Verl., 2001.

Bensberg, F. (2002): Segmentierung im Online-Marketing. S. 162-190. Erschienen in: Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-b).

Broder, A. J. (2000): Data Mining, the Internet, and Privacy. S. 56-73. Erschienen in: Spiliopoulou, M. / Masand, B. (2000).

Brosius, G. (2001): Data Warehouse und OLAP mit Microsoft. Bonn: Galileo Press GmbH, 2001.

Bürlimann, M. (1999): Web Promotion – Professionelle Werbung im Internet. St. Gallen/Zürich: Midas Management Verlag, 1999.

Callan, R. (2003): Neuronale Netze im Klartext [Übers.: Javier Botana]. München: Pearson Studium, 2003.

Chamoni, P. / Gluchowski, P. (1998): Analytische Informationssysteme: Data warehouse on-line analytical processing, data mining. Berlin et al.: Springer, 1998.

Cornelsen, C. (2002): Das 1x1 der PR: Öffentlichkeitsarbeit leicht gemacht [unter Mitarb. v. Stephanie Schwinn]. 4. überarb. Aufl., Freiburg i. Br., Berlin: Haufe, 2002.

Deutscher Kommunikationsverband BDW e.V. (1995): PR-Controlling – Dokumentation zum Fachtag ’95. 1995.

Ebert, G. / Steinhübel, V. (1995): Grundlagen des PR-Controlling. S. 5-14. Erschienen in: Deutscher Kommunikationsverband BDW e.V. (1995).

Englbrecht, A. (2002): Deskriptive Logfile-Analysen – Durchführung und Einsatzpotenziale. S. 124-139. Erschienen in: Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-b).

Fayyad, U. M. et al. (1996): Advances in Knowledge Discovery and Data Mining. Menlo Park, Calif.: AAAI Press, 1996.

Ferber, R. (2003): Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web. 1. Aufl., Heidelberg: dpunkt-Verlag, 2003.

Fuchs, P. / Möhrle, H. / Schmidt-Marwede, U. (1999): PR im Netz: Online-Relations für Kommunikations-Profis. Ein Handbuch für die Praxis. 2. Aufl., Frankfurt am Main: Institut für Medienentwicklung und Kommunikation GmbH in der Verlagsgruppe Frankfurter Allgemeine Zeitung GmbH (IMK), 1999.

Grothe, M. / Gentsch, P. (2000): Business Intelligence : aus Informationen Wettbewerbsvorteile gewinnen – München ; Boston [u.a.] : Addison-Wesley, 2000.

Grudowski, S. (2001): Web Public Relations. S. 79-98. Erschienen in: Riekert, W.-F. / Michelson, M. (2001).

Heindl, E. (2003): Logfiles richtig nutzen. 1. Aufl., Bonn: Galileo Press, 2003.

Herbst, D. (2001): Internet-PR [Besonderheiten der PR im Netz; Pressearbeit im Netz; Kommunikation mit wichtigen Bezugsgruppen]. 1. Aufl., Berlin: Cornelsen, 2001.

Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-a): Grundlagen des Web Mining – Prozess, Methoden und praktischer Einsatz. S. 2-31. Erschienen in: Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-b).

Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-b): Handbuch Web Mining im Marketing: Konzepte, Systeme, Fallstudien. 1. Aufl., Braunschweig, Wiesbaden: Vieweg, 2002.

Kotler, P. et al. (2003): Grundlagen des Marketing. 3., überarb. Aufl., München: Addison Wesley in Pearson Education Deutschland GmbH, 2002.

Lindner, W. (2003): PR@www: Öffentlichkeitsarbeit in Zeiten vom SMS und Internet. 1. Aufl., Essen: Stamm, 2003.

Marschall, N. (2002): Logfile-Analysen zur absatzpolitischen Auswertung von Internetpräsenzen [Elektronische Ressource: 1 CD-ROM]. Marburg: Tectum-Verl., 2002.

Martin, W. (1998): Data Warehousing: Data Mining – OLAP. 1. Aufl., Bonn: ITP GmbH, 1998.

Mena, J. (2000): Data Mining und E-Commerce – Wie Sie Ihre Online-Kunden besser kennen lernen und gezielter ansprechen [Übers. und dt. Bearb. Beate Meister]. Düsseldorf: Symposion Publ., 2000.

Meyer, M. (2002): Einsatz von Klassifikation und Prognose im Web Mining. S. 192-216. Erschienen in: Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-b).

Müller, A. (2002): Controlling-Konzepte : Kompetenz zur Bewältigung komplexer Problemstellungen. Stuttgart, Berlin, Köln: Kohlhammer, 2002.

Murray, D. / Durrell, K. (2000): Inferring Demographic Attributes of Anonymous Internet Users. S. 7-20. Erschienen in: Spiliopoulou, M. / Masand, B. (2000).

Niederst, J. (2000): HTML : kurz & gut [Dt. Übers. von Eva Wolfram]. 1. Aufl., Köln et al.: O'Reilly, 2000.

Payne, A. / Rapp, R. (2003): Handbuch Relationship-Marketing: Konzeption und erfolgreiche Umsetzung. 2., völlig überarb. und erw. Aufl., München: Vahlen, 2003.

Perry, T. (1995): PR-Erfolg und Kontrolle von Erfolg. S. 32-34. Erschienen in: Deutscher Kommunikationsverband BDW e.V. (1995).

Porter, M. E. (1999): Wettbewerbsstrategie – Methoden zur Analyse von Branchen und Konkurrenten (= Competitive strategy) [Dt. Übers. von Volker Brandt]. 10. durchges. und erw. Aufl, Frankfurt am Main, New York: Campus Verlag, 1999.[134]

Rapp, R. / Guth, S. (2003): Data Mining Anwendungen im Relationship Marketing. S. 165-179. Erschienen in: Payne, A. / Rapp, R. (2003).

Riekert, W.-F. / Michelson, M. (2001): Informationswirtschaft: Innovation für die neue Ökonomie. 1. Aufl., Wiesbaden: Dt. Univ.-Verl., 2001.

Säuberlich, F. (2002): Vorverarbeitung von Web-Daten – Pre-Processing. S. 106-123. Erschienen in: Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-b).

Sauvant, N. (2002): Professionelle Online-PR: Die besten Strategien für Pressearbeit, Investor relations, interne Kommunikation, Krisen-PR. Frankfurt am Main, New York: Campus-Verl., 2002.

Schaar, P. (2002): Datenschutz im Internet: Die Grundlagen. München: Verlag C. H. Beck, 2002.

Spiliopoulou, M. / Berendt, B. (2002): Assoziations- und Pfadanalyse – Entdeckung von Abhängigkeiten. S. 142-161. Erschienen in: Hippner, H. / Merzenich, M. / Wilde, K. D. (2002-b).

Spiliopoulou, M. / Masand, B. (2000): Web usage analysis and user profiling: revised papers / International WEBKDD '99 Workshop, San Diego, CA, USA, August 15, 1999. Berlin et al.: Springer, 2000.

Vollmuth, H. J. (2000): Controlling-Instrumente von A-Z. 5., erw. Aufl., Planegg / München: WRS Verl., 2000.

Vollmuth, H. J. (2003): Controllinginstrumente. 2., durchges. Aufl., Planegg / München: Haufe, 2003.

Walther, R. (2001): Web Mining S. 16-18. Erschienen in: Informatik Spektrum, Nr. 24, Ausgabe 02/2001.

[134] e-Book: Seitenzahlen orientieren sich an der e-Book-Ausgabe.

Internetquellen

Bolz, C. (2001): Web Mining Software und Dienstleistungen im Vergleich. URL: http://www.bolz.org/Vergleich_Web_Mining_Software.PDF – Zugriff am 06.12.2003.

Bundesdatenschutzgesetz – In: juris (2003). URL: http://bundesrecht.juris.de/bundesrecht/bdsg_1990/ – Zugriff am 10.11.2003.

ECIN (2003): Was Sie schon immer über Online-PR wissen wollten… Ergebnisse einer Umfrage von prdienst.de. 13.03.2003. URL: http://www.ecin.de/marketing/onlinejournalisten/ – Zugriff am 16.11.2003.

explido (2003): Online-PR – Instrumente und Möglichkeiten. URL: http://www.promotionwelt.de/marketingmix_online_pr.htm – Zugriff am 15.10.2003.

Fayyad, U. M. / Piatetsky-Shapiro, G. / Smyth P. (1996): From Data Mining to Knowledge Discovery in Databases. Erschienen in: AI Magazine 17(3): Fall 1996, S. 37-54 URL: http://www.aaai.org/Library/Magazine/Vol17/17-03/Papers/AIMag17-03-002.pdf – Zugriff am 29.11.2003.

Felzer, T. / Heidger, A. / Wiesiollek, M. (2002): Beaufsichtigtes Lernen - Vortragsausarbeitung im Seminar Neuronale Netze im SS 1993, TH Darmstadt. URL: http://www.st.informatik.tu-darmstadt.de:8080/felzer/nn_sem.pdf – Zugriff am 13.11.2003.

Frenz, Volker (2002): AccessLog Homepage (Programm zur Auswertung von Logfiles). URL: http://www.accesslog.de/ – Zugriff am 05.12.2003.

IVW (2003): URL: http://www.ivwonline.de . – Zugriff am 11.11.2003.

KDnuggets (2003): Web Mining and Web Usage Mining Software. URL: http://www.kdnuggets.com/software/web.html – Zugriff am 06.12.2003.

Klossek, M. (2001): Web Log Mining. URL: http://www.eworks.de/research/2001/05/WebLogMining/WebLogMining.pdf – Zugriff am 28.10.2003.

LEGAmedia (2001): Kennzahlensystem, Lexikon Controlling und Kostenrechnung. URL: http://www.legamedia.net/lx/result/match/22710757561423d5ad818dd198/index.php – Zugriff am 19.11.2003.

NetGeo Inc. (2003): Technology. URL: http://www.netgeo.com/technology/technology.html – Zugriff am 01.12.2003.

Rötzer, F. (1999): Anonymität im Internet. 07.07.1999. URL: http://www.heise.de/tp/deutsch/inhalt/te/5053/1.html – Zugriff am 10.11.2003.

Schulzki-Haddouti, C. (2003): Digitale Spuren. 31.01.2003. URL: http://www.heise.de/tp/deutsch/inhalt/te/14052/1.html – Zugriff am 09.10.2003.

Schwickert, A. C. / Beiser, A. (1999): Web Site Controlling. (Arbeitspapiere WI Nr. 7/1999). URL: http://wi.uni-giessen.de/gi/dl/showfile/Schwickert/1155/Apap_WI_1999_07.pdf – Zugriff am 19.11.2003.

Schwickert, A. C. / Wendt, P. (2000): Controlling-Kennzahlen für Web Sites. (Arbeitspapiere WI Nr. 08/2000). URL: http://wi.uni-giessen.de/gi/dl/showfile/Schwickert/1168/Apap_WI_2000_08.pdf – Zugriff am:19.11.2003.

SevenOne Media (2003): Aktuelle Ergebnisse der Langzeitstudie "Time Budget". 24.11.2003. URL: http://www.sevenonemedia.de/unternehmen/presse/pm/index.php?pnr=83478 – Zugriff am 24.11.2003.

Volkswagen AG (2003): Mobile Services URL: http://mobileservices.volkswagen.de/0_mobilizer/0_index/?serviceid=handy – Zugriff am 16.11.2003.

W3C (1999): Status Code Definitions. URL: http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html#sec10 – Zugriff am 12.10.2003.