Steffen Leja

Web-Mining und dessen Einsatzmöglichkeiten im modernen Unternehmen

Diplomarbeit

Diplom.de

Bibliografische Information der Deutschen Nationalbibliothek:

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de/ abrufbar.

ISBN: 9783832470685

Steffen Leja

Web-Mining und dessen Einsatzmöglichkeiten im modernen Unternehmen

Diplom.de

Steffen Leja

Web-Mining und dessen Einsatzmöglichkeiten im modernen Unternehmen

Diplomarbeit
an der Fachhochschule Würzburg-Schweinfurt
Fachbereich Betriebswirtschaft
März 2003 Abgabe

Diplom.de

Diplomica GmbH
Hermannstal 119k
22119 Hamburg

Fon: 040 / 655 99 20
Fax: 040 / 655 99 222

agentur@diplom.de
www.diplom.de

ID 7068

ID 7068
Leja, Steffen: Web-Mining und dessen Einsatzmöglichkeiten im modernen Unternehmen
Hamburg: Diplomica GmbH, 2003
Zugl.: Fachhochschule Südwestfalen, Fachhochschule, Diplomarbeit, 2003

Diplomica GmbH
http://www.diplom.de, Hamburg 2003
Printed in Germany

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

Abb.	Abbildung
Bd.	Band
BDSG	Bundesdatenschutzgesetz
bspw.	beispielsweise
bzw.	beziehungsweise
Diss.	Dissertation
et al.	et alii
etc.	et cetera
f.	folgende
ff.	fortfolgende
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IMAP	Internet Message Access Protocol
IuKDG	Informations- und Kommunikationsdienste-Gesetz
LAN	Local Area Network
MDStV	Mediendienste-Staatsvertrag
MIME	Multimedia Internet Mail Extension
OPS	Open Profiling Standard
P3P	Platform for Privacy Preferences Project
POP	Post Office Protocol
rev.	revidiert(e)
S.	Seite
SMTP	Simple Mail Transfer Protocol
TCP/IP	Transfer Control Protocol/Internet Protocol
TKG	Telekommunikationsgesetz
u.a.	und andere(s)
Univ.	Universität
URL	Uniform Resource Locator
Vol.	Volume
W3C	World Wide Web Consortium
WWW	World Wide Web
z.B.	zum Beispiel

1 Einführung

1.1 Problemstellung

Das Internet entpuppte sich in den letzten Jahren als wahre Revolution. Keine andere Technologie, auch nicht Telefon oder Fernsehen, hatte zuvor derart schnell Einzug in die private und berufliche Sphäre gehalten wie das World Wide Web (WWW) [1]. E-Commerce und E-Business waren die dominierenden Managementthemen. Es wurden immense Investitionen in den Aufbau, die Optimierung und die interne Integration des neuen Mediums investiert. Durch die Angst getrieben, am großen, gewinnversprechenden Kuchen des E-Business nicht teilzuhaben, stürzten sich viele Unternehmen nach dem Motto „Dabei sein ist alles“ in die Welt des WWW. Wie sich in der Vergangenheit gezeigt hat, konnten die überzogenen Erwartungen nicht erfüllt werden. Negative Schlagzeilen über erfolglose bzw. gescheiterte Online-Projekte oder ganzer „Dot-com“-Unternehmen haben die E-Commerce-Euphorie relativiert. Exemplarisch sind an dieser Stelle nur die Insolvenzverfahren der Internet- und Mediaagentur Popnet, der Internetdienstleister Exodus und Ision oder des Internetportals Sportgate von Boris Becker zu nennen[2]. Auch in den Führungsetagen herrscht wieder eine größere Vorsicht. Die Einsicht, dass auch Internetaktivitäten eine Strategie und davon abgeleitet auch Instrumente der Steuerung und Kontrolle benötigen, setzt sich nun langsam durch[3]. Gerade hier liegt das große Unterstützungspotential des Web Mining.

Noch nie konnten Verantwortungsträger ihre Entscheidungen anhand solch detaillierter und umfangreicher Informationen treffen wie heute. Dies betrifft keineswegs nur Daten auf technischer Ebene. Auch Marketing, Vertrieb und Controlling beginnen inzwischen, das große Potenzial der internetbezogenen Datenquellen für sich zu entdecken[4]. Online-Kunden hinterlassen wissentlich oder unwissentlich eine große Anzahl digitaler Spuren bei ihrem virtuellen Besuch des Unternehmens. Viele dieser Daten liegen zwar in

[1] Das World Wide Web (WWW) stellt als hypertextbasiertes Informationswerkzeug neben Electronic Mail (E-Mail), FTP, Telnet und Newsgroups den wichtigsten Dienst des Internets dar. Eine genaue Definition des Begriffs erfolgt ebenfalls im Punkt 2.1.3 dieser Arbeit.

[2] Vgl. **o. V. (d)**, vgl. **o. V. (e)**, vgl. **o. V. (f)**, vgl. auch **o. V. (l)**

[3] Vgl. **Scheer, A.-W.; Breitling, M.; S.397 ff.**

[4] Vgl. **Bensberg, F. (b), S. 78**

unstrukturierter Form vor, dennoch ist es im Idealfall möglich den gesamten Weg des Kunden, vom Werbemittel bis zur Kaufentscheidung, nachzuzeichnen. Nicht von ungefähr spricht Schida von einem neuen Zeitalter der Erfolgskontrolle[5].

Viele Unternehmen haben es jedoch bis heute verpasst, mittels Web Mining das umfassende Datenmaterial über Kunden und Besucher für die Optimierung des Web-Angebots zu nutzen[6]. Aber das Sammeln der Daten alleine verschafft noch keinen Wettbewerbsvorteil. Entscheidend ist, das zunächst vorhandene neutrale Datenmaterial auszuwerten und in aussagekräftige Informationen über die Besucher und deren Nutzungsverhalten umzuwandeln. Zugegeben stellt dies keine triviale Aufgabe dar! Nur wer es versteht, die entscheidenden Erfolgstreiber zu selektieren und für sich zu nutzen, wer Kunden und Marktsegmente im WWW kennt und weiß, wie die einzelnen Faktoren im Internet zusammenspielen, wird auf Dauer erfolgreich sein und sich gegenüber seinen Konkurrenten durchsetzen können.

1.2 Ziel und Vorgehensweise

Vor dem in der Problemstellung geschilderten Hintergrund und der wachsenden Verlegung von Unternehmensdarstellungen, Kommunikation, Marketing und Vertrieb auf das Internet, einhergehend mit einer zunehmenden Tendenz zur Personalisierung der Kundenansprache, erlangt die Analyse von Online-Daten eine herausragende Bedeutung. Aus diesem Grund soll in der vorliegenden Arbeit ein vollständiger Überblick über das Web Mining, von der Datengewinnung, über die Datenaufbereitung und –auswertung, bis zur Datenverwendung gegeben werden. Ziel ist, die technologischen und die betriebswirtschaftlichen Aspekte des Web Mining möglichst kompakt aber dennoch vollständig zu systematisieren und darzustellen, um Verantwortlichen in den Unternehmen, insbesondere in den Bereichen IT, Marketing, Vertrieb und Controlling einen schnellen Einstieg in das weite Feld des Web Mining zu ermöglichen.

Welche Datenquellen gibt es?
Mit welchen Methoden und Techniken können die Daten erhoben werden?
Wie können die gewonnenen Daten gespeichert werden?

[5] Vgl. **Schida, R.; Busch, V.; Diederichs, M., S. 252**
[6] Vgl. **Mena, J. (c), S.2**

Welche Probleme und Schwierigkeiten existieren bei der Datenerhebung?

Wie kann man die Qualität der erhobenen Daten verbessern?

Wie können die Daten ausgewertet werden?

In welchen Bereichen können die zu Informationen gewordenen Daten genutzt werden?

Diesen und weiteren Fragen wird in den folgenden Kapiteln in ausführlicher Weise nachgegangen.

Das Vorgehen zur Erstellung der Arbeit ist in vier Themenkomplexe abgrenzbar, die sich sowohl in der Gliederung der Arbeit als auch in grafisch veranschaulichter Form widerspiegeln.

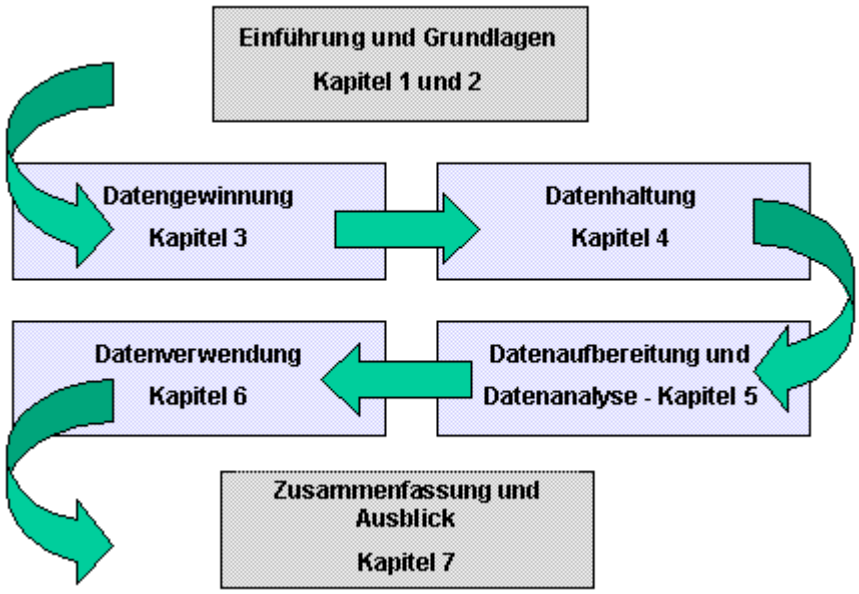

Abb. 1: Vorgehensweise

Was ist Web Mining? Was ist das WWW? Wie funktioniert die Kommunikation im WWW? Was ist das HTTP? Die Klärung dieser grundlegenden Fragestellungen findet im Kapitel 2 statt. Die Zielsetzung ist demnach, ein Basiswissen aufzubauen, mit dem die folgenden Abschnitte leichter nachvollziehbar werden.

Im Kapitel 3 werden die unterschiedliche Quellen und Techniken der Datengewinnung aufgezeigt. Natürlich liegt der Schwerpunkt auf der Gewinnung von Web-Server-Daten, da diese die zentrale Quelle darstellen. Im Rahmen dieser Arbeit werden auch weitere bedeutende Quellen im und außerhalb des Mediums WWW vorgestellt. Zu benennen sind hier die Sammlung von Benutzerdaten mittels Web-Formularen und, als Quelle außerhalb des Mediums WWW, insbesondere die unternehmensinternen Daten aus der Finanzbuchhaltung oder dem Vertrieb.

Aufgrund der Fülle von Daten wird im Kapitel 4 die dateibasierte der datenbankbasierten Datenhaltung gegenübergestellt und die jeweiligen Vor- und

Nachteile herausgearbeitet. In diesem Kontext wird zudem das Data Warehouse-Konzept erläutert.
Leider gibt es bis heute in diesem Kontext noch nicht die „eierlegende Wollmilchsau“! Aus diesem Grund werden in Kapitel 5 zunächst die derzeit bedeutendsten Probleme und Beschränkungen im Web Mining beschrieben. Zum Glück existieren inzwischen Methoden und Techniken, die diese Probleme teilweise beheben oder zumindest reduzieren können. Anzuführen sind hier beispielsweise Schlagworte wie Session-ID, Cookies, Web Bugs, Data Cleaning, usw..
Im Rahmen dieses Kapitels wird zudem die Datenauswertung behandelt. Grundsätzlich wird zwischen Verfahren der hypothesengestützten und der hypothesenfreien Entdeckung von Mustern unterschieden. Es werden zum einen hypothesengestützte Verfahren der statistischen Analyse sowie OLAP und zum anderen Methoden, die aus dem Bereich des Data Mining stammen, aufgeführt.
Wie kann man nun die gewonnenen Daten nutzen? Dies ist Hauptgegenstand des Kapitels 6. Darin sollen Möglichkeiten aufgezeigt werden, in welchen Bereichen Online-Daten, insbesondere Web-Server-Daten Verwendung finden können. Hier soll, aufgrund der eingehend geschilderten Situation in der Wirtschaft, besonderes Augenmerk auf das Web-Controlling gelegt werden.
In „Zusammenfassung und Ausblick“ (Kapitel 7) werden die wichtigsten Ergebnisse der Arbeit noch einmal dargestellt und es wird nach der Rolle des Web Mining für die Zukunft gefragt.

2 Grundlagen

2.1 Web Mining

Gegenstand des Web Mining ist die allgemeine Anwendung moderner Verfahren des Data Mining auf Datenstrukturen des Internets[7]. In der Literatur finden sich unterschiedliche Definitionen zum Thema Web Mining. Nach Cooley et al. kann Web Mining definiert werden als:

> *Die Entdeckung und Analyse nützlicher Informationen des WWW. Dies umfasst die automatisierte Suche in online verfügbaren Informationsquellen (Web Content Mining) sowie die automatische Generierung von Navigationsmustern bzgl. der Besucher einer Website (Web Usage Mining).*

Demnach umfasst Web Mining die Analyse aller Daten des WWW (Web-Daten), incl. Nutzungs- und Kundendaten, sowie inhaltliche Daten des WWW. Dies sind Inhalte von HTML-Dokumenten als auch Daten über die Struktur einer Website. Aus diesem Grund werden die Daten in Nutzungs-, Inhalts- und Strukturdaten differenziert.

- *Nutzungsdaten* beschreiben die Nutzungsmuster der Besucher einer Website. Dies schließt den Host des Besuchers, verweisende Websiten, sowie das Datum und die Zeit des Zugriffs ein. In diese Kategorie fallen außerdem E-Commerce-Daten wie Transaktionsdaten und spezielle Kundendaten.
- *Inhaltliche Daten* sind Informationen, die in den Web-Seiten enthalten sind. Im Allgemeinen beinhalten diese Daten Texte und Graphiken.
- Die Strukturinformationen innerhalb einer Web-Seite bzw. Website, wie die Anordnung der Web-Seiten, Einstiegsseiten usw. bilden die *Strukturdaten*.

Im Kontext des Web Mining spielt die Website als Messobjekt eine zentrale Rolle. So verkörpert eine Website die Präsenz eines Unternehmens im elektronischen Wirtschaftsgefüge[8]. Die technologische Basis hierfür bietet der Internet-Dienst WWW, der es erlaubt, verknüpfbare Dokumente mit multimedialem Inhalt anzuzeigen. Dabei

[7] Vgl. **Mobasher, B.; Jain, N.; Han,E.—H., Srivastava, J., S. 1, vgl. auch . Cooley, R.; Mobasher, B.;Sirvastava, J. (b), S. 568**

[8] Vgl. **Schwickert, A.C. (c), S. 7**

beschränkt sich eine Website nicht alleine auf den öffentlich zugänglichen Bereich, die so genannte Homepage, sondern beinhaltet auch die abgesicherten Bereiche zur Kooperation mit anderen Unternehmen (Extranet) und für die unternehmenseigene Kommunikation (Intranet) [9]. Aus technischer Sicht setzt sich eine Website aus mehreren, durch sogenannte Hyperlinks verbundene Seiten zusammen, die auf einem Web-Server[10] vorliegen. Inhaltlich lassen sich die Seiten nach ihrer Funktion in Navigationsseiten, Informationsseiten und in interaktive Anwendungen unterteilen [11].

Durch Web Mining können leistungsorientierte Größen wie Verfügbarkeit und übertragenes Datenvolumen erfasst werden. Zusätzlich lassen sich aber auch die Nutzungsvorgänge auf einer Website beobachten, wodurch der Umgang des Nutzers mit den dargebotenen Inhalten ebenfalls zum Beobachtungsobjekt wird. Insbesondere besteht durch Web Mining die Möglichkeit, selbständig Muster in den Nutzungsdaten aufzufinden. Gerade derartige Muster im Verhalten der Online-Kunden können im zunächst anonymen Medium Internet jedoch von hoher Bedeutung für die Informationsgewinnung sein. Daher bietet es sich zur automatischen Mustererkennung an, klassische Data Mining-Verfahren auf Internetdaten anzuwenden, um tiefer gehende Informationen über die Nutzer einer Website aufzupüren[12].

Zu den Zielen des Web Mining zählen hier einerseits die qualitative Verbesserung der Website und die Beseitigung von Fehlern, aber auch die Gewinnung von Informationen über die Nutzer und deren Verhalten.

Bisher wird Web Mining vor allem im E-Commerce-Bereich[13], also in der Beziehung zwischen Unternehmen und der Masse, von in der Regel anonymen Nutzern betrieben. Eine Verwendung im Intranet bzw. Extranet - besonders in großen Konzernintranets und –extranets - ist jedoch ebenfalls denkbar.

---

[9] Vgl. **Schwickert, A.C. (c); S. 6 ff.**. Ein Intranet ist ein unternehmensinternes, informationsverteilendes, IP-basiertes Netzwerk, das allerdings vom öffentlich zugänglichen Internet durch eine Firewall abgekoppelt ist. Ein Extranet ist ebenfalls vom Internet abgekoppel, jedoch ist ein unternehmensübergreifendes Netzwerk. Vgl. **Meyer, M.; Weingärtner, S.; Döring, F.; S. 5 ff.**. Eine Abgrenzung der Begriffe Internet, Intranet und Extranet erfolgt im Punkt 2.1.3 dieser Arbeit.

[10] Unter Web-Server ist ein Rechner oder eine Gruppe von Rechnern, die zusammen eine Web-Anwendung realisieren. Ist der Oberbegriff für HTTP-, Applications- und Daten-Server, wird aber auch als Synonym für HTTP-Server verwendet. Vgl **Rahm, E.; Stöhr, T.; S.477**

[11] Vgl. **Schwickert, A.C. (c); S.16**

[12] Vgl. **Bensberg, F.; Weiß, T.; S. 426**

[13] E-Commerce umfasst im wesentlichen den Kontakt mit Kunden über das Internet als Vertriebskanal. Vgl. **Meyer, M.; Weingärtner, S.; Döring, F.; S. 9**

2.1.1 Richtungen des Web Mining

Web Mining wird, je nach Auffassung des Autors in zwei oder drei Teilbereiche unterteilt. Es beinhaltet grundsätzlich die Analyse von Seiteninhalten (Web Content Mining) als auch die Untersuchung des Nutzerverhaltens (Web Usage Mining). Zaiane bildet eine weitere Kategorie, die die Seitenstrukturen als Grundlage für die Wissensentdeckung heranzieht und wird daher als „Web Structure Mining" bezeichnet[14]. Dieser Aufteilung folgt auch Sirvastava[15]. Bensberg und Spiliopoulou hingegen grenzen lediglich die Teilgebiete „Web Content Mining" und „Web Usage Mining" voneinander ab[16]. Diese spalten jedoch das Web Usage Mining nochmals in zwei Teilbereiche auf. Bei Spiliopoulou wird in Abhängigkeit, ob die bei der Analyse zur Verfügung stehenden Daten personenbezogen sind oder nicht, zwischen „Web Usage Mining - Impersonalized" und „Web Usage Mining – Personalized" unterschieden. Letzterer Bereich konzentriert sich auf die Erstellung von Nutzerprofilen und die Anwendung dieser Profile für die Einrichtung personalisierter Dienste[17]. Bensberg schlägt eine andere Klassifizierung vor: Er unterscheidet zwischen dem „Web Log Mining"(WLM) und dem „Integrated Web Usage Mining"(IWUM). Bei der Ausprägungsform WLM beschränkt sich die Analyse ausschließlich auf die Protokolldateien des Web-Serves. Fließen neben den Protokolldateien noch weitere Datenbestände in den Analyseprozess mit ein, so spricht man vom IWUM [18]. In der folgenden Abbildung ist ein Überblick möglicher Unterteilungen des Web Mining aufgezeigt:

[14] Vgl. **Zaiane, Osmar R. (b); S. 17 u. S. 20**

[15] Vgl. **Sirvastava, J.; Cooley, R.; Deshpande M.; Tan, P.-N.; S. 12**

[16] Vgl. **Bensberg, F. (**b**), S. 131**; **Bensberg, F.; Weiß, T.; S. 426**, vgl. auch **Spiliopoulou, M.; S 490 f**. Bensberg folgt der dreigeteilten Kategorisierung nicht, da seiner Ansicht nach Strukturdaten als inhaltsbezogene Daten erfasst werden, deren Analyse durch den Aufgabenbereich des Web Content Mining abgedeckt wird. Vgl. **Bensberg, F. (b); S. 131**

[17] Vgl. **Spiliopoulou, M.; S 490**

[18] Vgl. **Bensberg, F. (b); S. 131**; **Bensberg, F.; Weiß, T.; S. 426**