Das Geheimnis des Web Mining

Die Suche nach **verborgenen Schätzen**

Die Analyse der Besucher auf den Internet-Portalen – den virtuellen Filialen - wird für die Unternehmen immer wichtiger. Für das Controlling der Besucherdaten und deren Verhalten etablieren sich verschiedene Produkte auf dem Markt. Viel interessanter ist es, die Besucher der Internetseite zu analysieren, während diese auf der Seite surfen, deren Bedürfnisse zu erkennen und passende Internet-Seiten anzubieten.

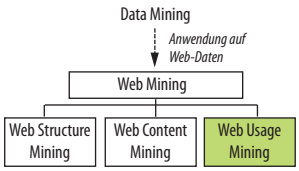
Im folgenden Artikel wird ein Closed-Loop-Ansatz auf Basis von „Web Mining“ vorgestellt, um das Internet als Vertriebskanal zu nutzen.

Oberste Ziele von Web Mining-Projekten sind die Generierung von Kundeninformationen und die Personalisierung der Webseite. Der Artikel soll sich dementsprechend auf das Web Usage Mining fokussieren, wobei die anderen beiden Bereiche nur grob umrissen werden.

Web Mining bezeichnet die Anwendung von Data Mining-Verfahren zur automatischen Entdeckung und Extraktion von Informationen und Mustern auf Datenstrukturen des Internets. Es sollen interessante und interpretierbare Zusammenhänge in den vorhandenen Web-Daten gefunden werden. Im Generellen unterscheidet man nach den Anwendungsgebieten oder den zugrundeliegenden Daten, wie beispielsweise HTML-Dokumenten, Hyperlinks oder Log-Dateien, zwischen den drei Bereichen Web Content Mining, Web Structure Mining und Web Usage Mining.

Anwendungen: Klickpfade, Kaufvohersagen, Warenkorbanalysen, Benutzerprofile, Profitabilitätsanalysen, Cross- & Up-Selling.

Verfahren: Entscheidungsbaumverfahren, Assoziationsanalysen, Sequenzanalysen, Clusterverfahren, Prognosen.



Web Content Mining befasst sich mit der Erkennung und Extraktion von Mustern in Webdokumenten, die aus Texten, Hyperlinks, Grafiken, Audio- und Videostreams bestehen können. Allgemein betrachtet ist der Nutzen für das eigene Unternehmen in der systematischen Beschaffung, dem Vergleich, der Analyse und der Verwertung von Informationen aus dem Internet zu sehen.

Eine besondere Bedeutung kommt der Suche und Gruppierung von Dokumenten nach inhaltlichen Kriterien sowie der Verlinkung zu Nachbarseiten zu. Anders als beim herkömmlichen Data Mining, das Daten in Tabellenform braucht, kommen hier vorzugsweise Algorithmen des Text Minings zum Einsatz, da ein Großteil der Webinhalte nur in schwach- oder unstrukturierter Form vorliegt.

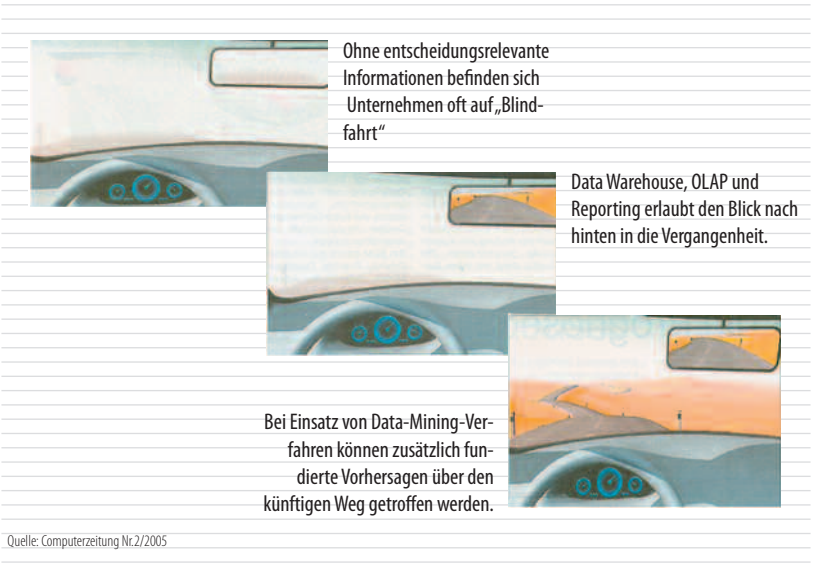


Bild 2: Entscheidungsunterstützung für das Management.

Konkrete Beispiele solcher auffindbaren Daten im Web sind Archive mit Zeitungsartikeln, wissenschaftlichen Publikationen, Preislisten, Kundendaten, Qualitätsberichte, Logos, E-Mails oder Meldungen von Nachrichtenagenturen. Weitere Beispiele für den praktischen Einsatz von Web Content Mining sind:

- Automatische Inhaltserschließung: Die wichtigsten Sätze eines Dokumentes können zum Zweck von Zusammenfassungen extrahiert werden. Aus Textpassagen werden Titel generiert.
- Analyse von Onlineinformationsdiensten wie Zeitungen und Newsdiensten zur Identifikation von „Wörtern des Tages“. Einsatz bei: http://www.wortschatz.uni-leipzig.de/wort-des-tages/
- und die Visualisierung der Beziehungen zwischen Objekten (Begriffe, Dokumente) in Form von Graphen.
- Spam-E-Mails von erwünschten Zusendungen unterscheiden
- Klassifizieren von Sprach-, Sound- und Videosegmenten nach inhaltlichen Kriterien für den Einsatz in personalisierten Portalanwendungen
- Analyse von Kundenrezensionen zu Produkten: Produkteigenschaften über die sich Kunden in Kommentaren geäußert haben extrahieren. Aus diesen Meinungsäußerungen Vor- und Nachteile eines Produktes identifizieren oder Produkte gleicher Kategorien anhand ihrer herausgefilterten Bewertungskriterien in einer Vergleichsübersicht den Kunden zur Verfügung stellen.

Text Mining ermöglicht den Aufbau von Wissenskarten (Knowledge-Maps), die dem Nutzer ein begriffliches Leitsystem zur Verfügung stellen oder auch eine interaktives und grafisches Navigieren durch die Webseiten zulassen.

- ein themenspezifischer Internetsuchdienst oder ein Web-Info-Center, bei dem Informationen zu einem bestimmten Themengebiet wie Finanzwesen, juristische Urteile, Sport, etc. automatisch aus dem Internet extrahiert werden

Aufgrund der wachsenden Zahl elektronisch verfügbarer Texte und dem Wunsch nach automatischen Verfahren zur Bewältigung der Informationsflut gehört Text Mining bereits zu einem

durchaus aktiven und interessanten Forschungsgebiet. Ein rasant wachsendes Anwendungsgebiet ist außerdem die Auswertung von Blogs, sog. Online-Tagebüchern, in denen sich die User über aktuelle Trends oder Erfahrungen mit Produkten und Dienstleistungen austauschen.

Mit Web Structure Mining lassen sich Informationen aus der Struktur der Webseite, die sich auf die Anordnung der dort auffindbaren Inhalte bezieht, aufdecken. Zum einen kann sich solch eine Analyse auf die Struktur innerhalb einer Webseite beziehen zum anderen auf die Struktur zwischen den Seiten eines Webauftritts, das heißt, auf die Verlinkung der einzelnen Seiten zueinander. Es geht darum, durch die Extraktion signifikanter Klickprofile Hinweise auf eine verbesserte Gestaltung und Navigation der Webseite zu erhalten, auch in Bezug auf eine möglichst effektive Positionierung von Werbebannern. Ein anderes Anwendungsszenario, das sich jedoch auf das ganze Internet bezieht, ist die automatische Bewertung der Qualität von Internetseiten. Hierbei kommen der von den Google-Gründern Larry Page und Sergey Brin entwickelte Page-Rank-Algorithmus und der HITS-Algorithmus (hypertext induced topic selection) zum Einsatz. Die Grundannahme, je höher die Anzahl der Links, die auf eine Webseite verweisen, desto höher wird deren Autorität bewertet, bildet den Kern dieser Algorithmen. Häufig zitierte Seiten werden dementsprechend als besonders gut und mit einem hohen Rang eingestuft. Die Internetsuchmaschine Google stellt damit ein System für die Bewertung der gelisteten Seiten zur Verfügung.

Das Web Content Mining und das Web Structure Mining haben ähnliche Beziehungen, da es sehr wahrscheinlich ist, dass die Webdokumente Links enthalten und beide die realen oder Primärdaten im Web verwenden. Demzufolge kommen diese beiden Techniken in einer Anwendung oft gemeinsam zum Einsatz. Diese beiden Web Mining Gebiete tragen nicht dazu bei Informationen über meine Online-Nutzer aufzuspüren daher liegt der Fokus auf dem im Folgenden vorgestellten Web Usage Mining.

WEB-TIPP:

http://www.w3.org/Daemon/User/Config/Logging.html

http://www.symposion.de/ecrm-hb/e-crm-04.htm

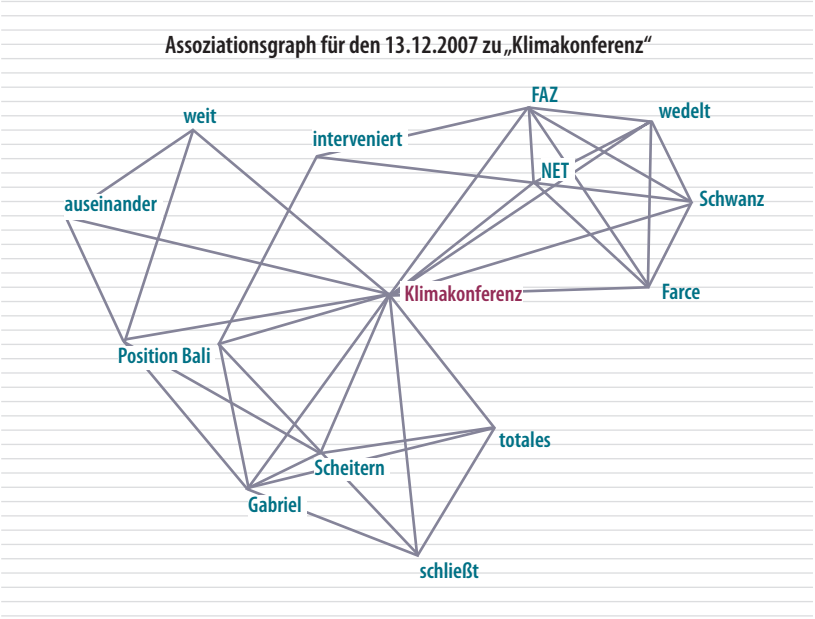


Bild 2: Tagesaktueller Assoziationsgraph.

Das Web Usage Mining beschäftigt sich mit der Aufgabe, das Benutzerverhalten mit Hilfe von Web-Server-Logdateien auszuwerten. Im Vordergrund stehen hierbei Personalisierungsmaßnahmen, mit denen sich die e-Commerce Unternehmen auf die Bedürfnisse der Kunden konzentrieren wollen. Gegenstand der Personalisierung können neben den Informationen auf der Webseite ebenso die Produkte und die Dienstleistungsgestaltung, als auch die Interaktion mit den Online-Kunden sein. Bild 4 stellt in Anlehnung an das Leistungssystem nach Belz, die drei Ebenen der elektronischen Interaktion zwischen Kunde und Anbieter dar. Die innere Ebene bezieht sich auf die Personalisierung der Produkte und Leistungen des Unternehmens wobei die äußeren Ebenen (Layout, Interface, Kommunikation) die Personalisierungsmaßnahmen der Interaktion mit den Online-Kunden beschreiben. Der Aufbau und das Layout eines Online-Portals kann an die Bedürfnisse des Kunden angepasst werden, um ihm ein möglichst gutes Zurechtfinden auf den Seiten und einen bestmöglichen Überblick über die von ihm bevorzugten Informationen zu garantieren.

Verborgene Informationen, die durch den Einsatz von Web Mining gewonnen werden, sind von zentraler Bedeutung.

Personalisierungsmaßnahmen

Die Personalisierung der Kommunikation kann zum einen auf die Kundenansprache und zum anderen auf die Kommunikationsart angewendet werden. Dem Kunden sollte die Möglichkeit gegeben werden, den von ihm bevorzugten Kommunikationskanal (E-Mail, Webseite, Telefon, Chat) zu wählen, vor allem dann wenn der Kunde online auf der Seite beraten werden soll.

Der Einsatz spezieller Personalisierungsmaßnahmen ist vor allem dann sinnvoll, wenn die Informationsmenge und die Nutzeranzahl sehr hoch sind. Vorreiter für den Einsatz von Personalisierungen von Produkten und Leistungen bei Online-Shops ist das Social-Commerce Versandhaus Amazon.de. Dort wird analysiert, welche Produkte oft zusammen gekauft werden und es

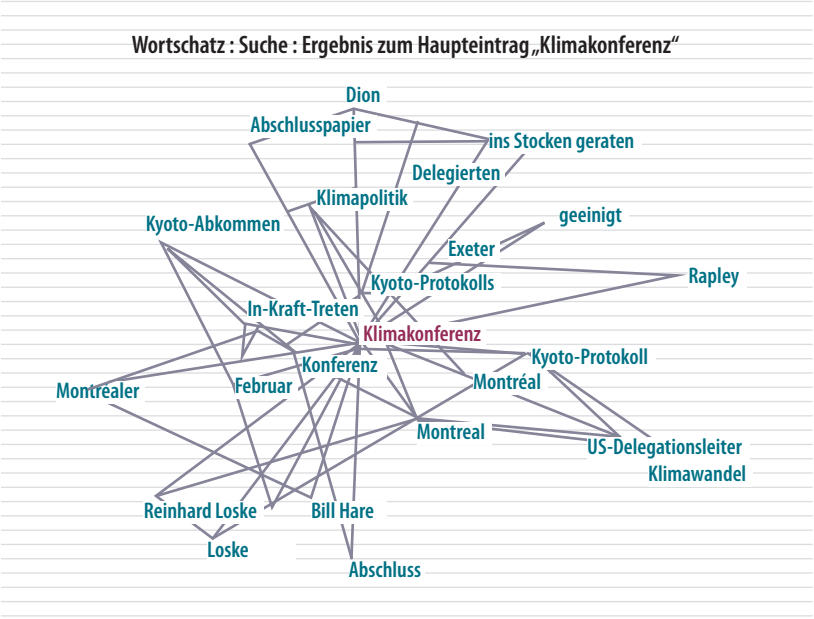


Bild 3: Assoziationsgraph aus dem Referenzcorpus. Der aktuelle Assoziationsgraph vom 13.12.2007 in Bild 2 hebt für das derzeit häufige Wort „Klimakonferenz“ aktuelle Beziehungen (etwa Bali, Gabriel) hervor, wobei im Referenzgraphen (Bild 3) eher grundsätzliche Beziehungen (Klimapolitik, Abschlusspapier) verdeutlicht werden.

werden Vorschläge in der Form „Kunden, die diesen Artikel gekauft haben, kauften auch:“ oder „Kunden, die Artikel gekauft haben, welche Sie sich kürzlich angesehen haben, kauften auch:“ generiert. Damit kann der Kunde bei der Produktauswahl unterstützt werden und es kann eine Absatzsteigerung komplementärer Produkte erzielt werden.

Das hierbei eingesetzte Verfahren nennt sich Warenkorbanalyse, bei der mittels Assoziationsregeln untersucht wird welche Produkte oft zusammen gekauft werden, um Marketing-Aktivitäten wie Cross-/Up-Selling zu planen. Cross-Selling („Quer-Verkaufen“) meint dabei, dem Kunden weitere, komplementäre Produkte oder Dienstleistungen anzubieten. Hingegen wird beim Up-Selling („Aufwärts-Verkaufen“) dem Kunden die nächst höhere und preisintensivere Produkt- oder Dienstleistungskategorie angepriesen und so schmackhaft gemacht, dass sich der Kunde zum Kauf entschließt. Die Warenkorbanalyse bedarf aber der Integration von Transaktionsdaten, da diese in den Logfiles nicht gespeichert werden.

Assoziationsanalysen

Assoziationsanalysen ermöglichen weiterhin die Beantwortung der Fragestellung: „Welche Informationsangebote/Seiten werden typischerweise zusammen aufgerufen?“. Als Sonderform der Assoziationsanalyse können anhand von Sequenzanalysen zusätzlich Aussagen über die Reihenfolge der aufgerufenen Seiten beziehungsweise der Online-Aktivitäten getroffen werden. Eine typische Fragestellung diesbezüglich ist, welche Informationen/Seiten die Besucher nach der Startseite aufrufen. Somit ist feststellbar welche Informationen und Produkte den größten Interessantheitsgrad für die Nutzer darstellen und welche Seiten am engsten mit dem Verkauf eines Produktes verbunden sind.

Beispiel: 100 % der Besucher betreten bei der „Flight“-Seite den Webauftritt und 48 % schauen sich danach die „Hotel“-Seite an (siehe Bild 5). Des Weiteren kann der Online-Shop hinsichtlich seiner kritischen Geschäftsprozesse untersucht werden, um Aufschluss über Prozessdurchlaufzeiten und Konversionsraten zu geben.

Klassifikationsverfahren

Als weiteres Anwendungsfeld ermöglicht das Web Usage Mining mit Hilfe von Klassifikationsverfahren ein Objekt einer von mehreren vordefinierten Klassen oder Kategorien, wie etwa „Käufer“ oder „Nicht-Käufer“ zuzuordnen. Dabei werden diejenigen Variablenkombinationen gesucht, die eine möglichst gute Zuordnung zu den Klassen Käufer oder Nicht-Käufer gewähren. Hierfür ist besonders der Einsatz von Entscheidungsbäumen geeignet. Als Ergebnis entstehen Regeln der Form „WENN-DANN“, die zumeist in einer Baumstruktur abgebildet werden. Bild 6 stellt einen solchen Entscheidungsbaum für eine Webseite mit Online-Shop dar. Wenn ein Besucher die Seite „Sales.html“ besucht, auf dieser länger als zehn Minuten verweilt und wenn der Besuch zwischen Mittwoch und Freitag erfolgt, dann liegt die Kaufwahrscheinlichkeit für ein Produkt bei 16,52 %. Dieser Erkenntnis zum Anlass könnte sich ein Unternehmen für die Versendung eines Newsletters jeweils an einem Mittwoch entscheiden, der auf die Seite „Sales.html“ verlinkt.

Ist keine Information bezüglich einer Klassenzugehörigkeit in den Daten enthalten, so können diese anhand von Clusterverfahren automatisch geschätzt werden. Ein (optimales) Cluster ist eine Gruppe von Objekten (etwa Besucher), deren Eigenschaften (Kriterien) möglichst ähnlich, von anderen Clustern aber deutlich getrennt sind. Die einzelnen Merkmale werden dabei über ein Proximitätsmaß (zum Beispiel „euklidische Distanz“) miteinander verglichen. Die Clusteranalyse dient vorrangig der Segmentierung. Die Internetuser können anhand ihrer Eigenschaften in ver-

Warenkorbanalysen helfen **Benutzerpräferenzen** und **personalisierte Empfehlungssysteme** anzubieten.

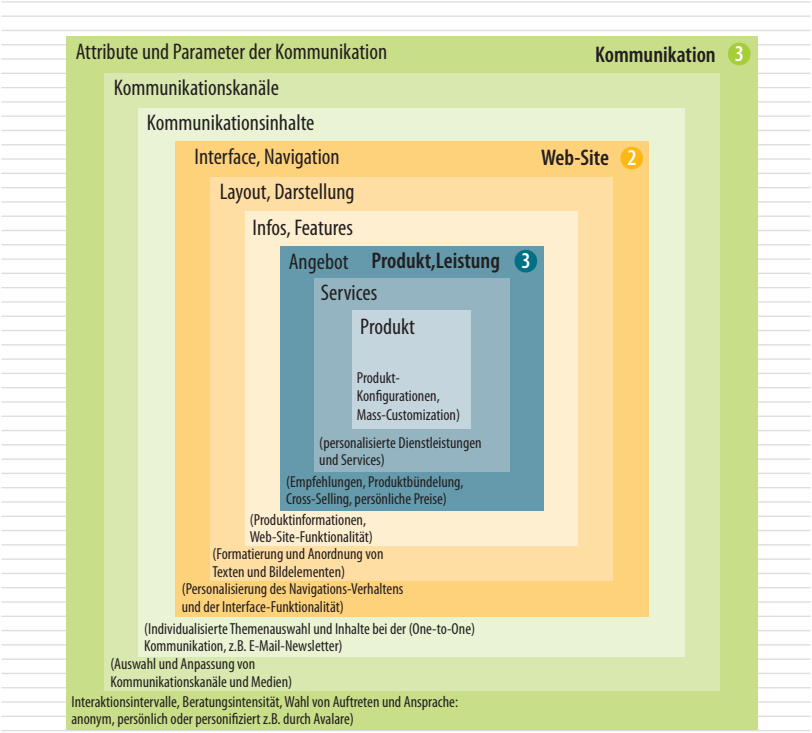


Bild 4: Leistungssystem der Personalisierung nach Belz [Riemer/Klein, 2002 S. 53].

schiedene Bereiche eingeteilt werden. Mögliche Segmentierungskriterien sind eingegebene Suchbegriffe, die Herkunft oder aufgerufene Seiten. Anhand dieser Merkmale kann versucht werden auf die Informationsbedürfnisse der Besucher-Gruppen zu schließen.

Der Prozess

Der Prozess des Web Mining durchläuft die im Bild 8 aufgezeigten Schritte der Datenselektion, Datenbereinigung, Datenvorbereitung, Datenanalyse (Data Mining) und Evaluation. Am Anfang eines Web-Mining-Projektes steht eine exakte Beschreibung der betriebswirtschaftlichen Problemstellung. Erst danach kann mit der Auswahl des Datenbestandes begonnen werden. Jeder Besucher hinterlässt mit jedem Mausklick auf der Webseite eine elektronische Spur in Form von Logfiles auf dem Web-Server. Als Standard gilt das Common Logfile Format. Andere Formate wie beispielsweise das Expanded Common Logfile Format leiten sich von diesem ab und enthalten zusätzliche Felder, wie beispielsweise Referrer und User-Agent. Ein Beispiel für ein Expanded Common Logfile Format ist aus Bild 7 ersichtlich. Es wird eine Abfrage eines Nutzers, der mit der IP-Adresse 123.456.78.1 am 06.12.2007 auf die Seite info.html zugriff, demonstriert. Wenn es sich um eine geschützte Seite handelt werden an der zweiten Position der Benutzername und an dritter Position das zugehörige Passwort aufgeführt.

Der HTTP-Status-Code 200 steht für eine erfolgreiche Seitenübertragung, zudem wurden 2198 Bytes gesendet. Der Nutzer kam über einen Link der Seite „Q.html“ und benutzte den Firefox Browser in der Version 2.0.0.11 in Kombination mit dem Betriebssystem Windows Server 2003. Es wird kein Referrer mit gesendet, wenn die Adresse per Hand im Browser eingegeben wurde oder Bookmarks verwendet wurden.

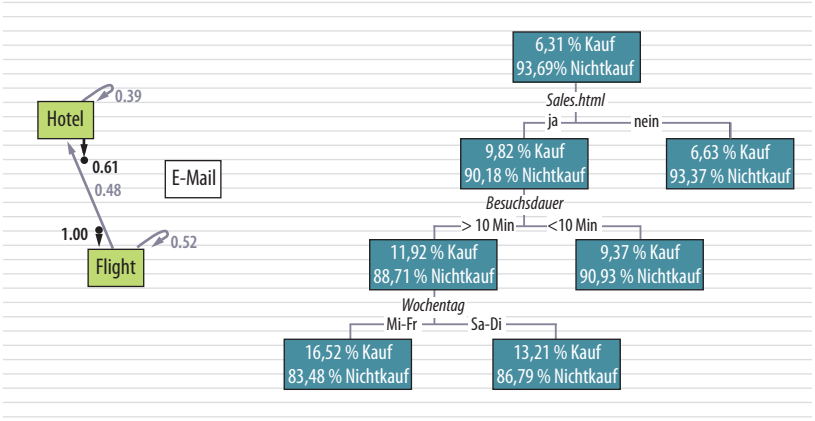


Bild 5: Sequenzanalyse [Quelle: Microsoft]. **Bild 6:** Exemplarische Entscheidungsbaumstruktur.

Schrittweises Vorgehen

Bevor die Data Mining Techniken sinnvoll auf den Datenbestand angewandt werden können, ist es erforderlich diese, aufgrund von technischen Begebenheiten und unsauberen Daten entsprechend aufzubereiten. Dabei können folgende Teilschritte angewendet werden:

- Entfernen von Ausreißern: Bereinigung atypischer Beobachtungen
- Behandlung fehlender Werte: Löschen der Datensätze anderenfalls Ersetzen mittels statistischer Schätzungen oder empirisch erhobener Werte

Die Qualität der Web-Daten ist ohne Frage ein entscheidender Faktor für valide Ergebnisse.

- Reduktion der Variablenanzahl/Datenvolumen: unwichtige Merkmale entfernen, Aggregation der Informationsobjekte (etwa Zusammenfassung nach zeitlichen Kriterien Tag → Woche → Monat → Quartal → Jahr), Summierung, Mittelwertbildung.
- Umkodierung von Variablen: Skalentransformation, Zusammenfassung in Klassen/Gruppen.
- Entfernen von Suchroboter-Einträgen (Spider): Es werden Einträge im Logfile generiert, die keine Hinweise auf das Verhalten der Besucher bringen. Diese Einträge sind vor der Analyse zu entfernen.
- Ziehung von Stichproben, um die Anzahl der zu analysierenden Daten per hinreichend großer Stichprobe zu reduzieren und die rechentechnische Effizienz dementsprechend zu erhöhen.
- Weitere grundlegende Prozessschritte der Datenaufbereitung sind die Identifikation von Besuchern und Sessions sowie von Seitenaufrufen. Die Herausforderung besteht in der Identifikation der Besucher, da es hier zu Unschärfen kommt, die selbst durch ein Analyse-Werkzeug kaum exakt ausgeglichen werden können.

Der Datenauswahl und Datenaufbereitung ist eine besonders hohe Aufmerksamkeit zu schenken, da diese mit bis zu 80 Prozent die zeitaufwendigste Phase im Data-Mining Prozess darstellen und zudem die Datenqualität einen entscheidenden Faktor für valide Ergebnisse darstellt.

Zur Rettung aus einem möglichen Datenchaos bietet sich die Datenhaltung in einem Data Warehouse an. Hier werden alle operativen Datenquellen, welche zumeist heterogen und unterschiedlicher Qualität sind, zusammengeführt, geordnet und strukturiert. Bei der Datenintegration können beispielsweise Registrierungsdaten wie Name, Adresse, Alter usw. mit den Logfile-Daten zusammengeführt werden. Auf Basis eines solch integrierten Datenbestandes sind übergreifende Auswertungen und Analysen sowie eine schnelle Suche nach Daten möglich.

Der nächste Schritt im Web Mining-Prozess ist die Anwendung automatischer Mustererkennungs-Verfahren auf die vorverarbeiteten Daten. Einige besonders geeignete Verfahren wurden bereits im Rahmen des Web Usage Mining vorgestellt.

Die Nachbereitung der Datenanalyse umfasst die Interpretation, Dokumentation und Evaluation der Analyseergebnisse, um darauf aufbauend Verbesserungsmaßnahmen zu ergreifen.

Bild 8 stellt die Prozess-Schritte und Methoden noch einmal veranschaulicht dar.

Kompetenz und Kontinuität

Business Intelligence-Projekte im Allgemeinen und Data Mining-Projekte im speziellen sollten nicht nur als Aufgabe der IT-Abteilung angesehen werden. Wichtig ist es sämtliche Kompetenzträger und Fachabteilungen (etwa Marketing, HR, Technik) in solche Projekte einzubeziehen, da diese individuelle Parameter setzen müssen.

Web Mining ist weiterhin keinesfalls als einmaliges Projekt anzusehen sondern es lebt langfristig nur als fester Bestandteil des Customer Relationship Managements. Letztendlich ist es durch ein Zusammenspiel von Akteuren und Technik gekennzeichnet, das nur so gut und stark sein kann, wie das schwächste Glied in dieser Kette- und das schwächste Glied ist of der Mensch.

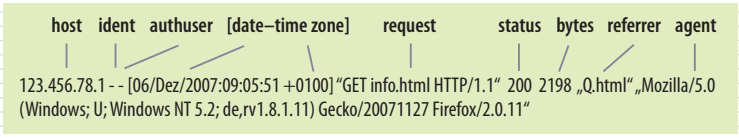


Bild 7: Logfile-Format.

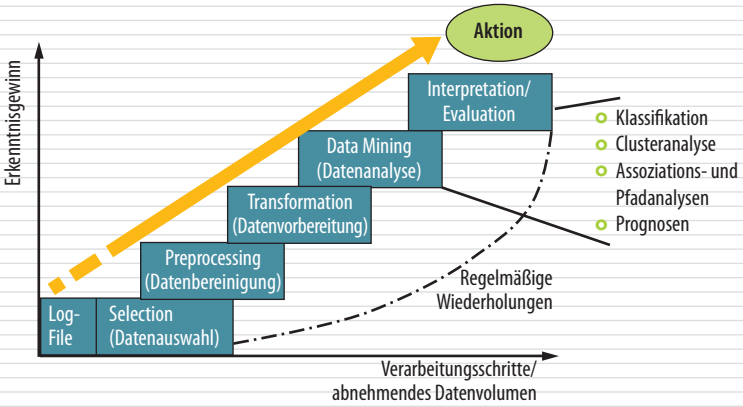


Bild 8: Prozess und Methoden des Web Usage Mining [Quelle: in Anlehnung an Symposion Publishing GmbH2].

Ein Werkzeug, das beispielsweise komplexe Data-Mining- und Web-Mining-Projekte über alle Phasen (von der Datenintegration und Datenaufbereitung bis zur Visualisierung der Ergebnisse) unterstützt ist der Microsoft SQL Server 2005. Im Besonderen mit den Analysis Services des MS SQL Server 2005 wird dem Anwender eine umfangreiche Palette an Data Mining Algorithmen zur Verfügung gestellt:

- Microsoft Decision Trees-Algorithmus
- Microsoft Clustering-Algorithmus
- Microsoft Naive Bayes-Algorithmus
- Microsoft Association-Algorithmus
- Microsoft Sequence Clustering-Algorithmus
- Microsoft Time Series-Algorithmus
- Microsoft Neural Network-Algorithmus (SSAS)
- Microsoft Logistic Regression-Algorithmus
- Microsoft Linear Regression-Algorithmus

Fazit

Web Mining als Anwendungsfeld von Data Mining-Techniken wie Clustering, Sequenz & Assoziationsanalysen und Neuronale Netze wird zukünftig noch weit aus intensivere Aktivitäten im Unternehmensumfeld versprechen. Diejenigen Unternehmen die es erkannt haben, die Informationen und die Masse an Benutzerdaten am effektivsten zu nutzen werden von einem strategischen Wettbewerbsvorteil profitieren.

Folgende Fragestellungen und Aufgaben können mittels Web Mining, also mit der Verwendung verschiedener Data Mining-Algorithmen bewältigt werden.

- Wie kann ich mein Service-Angebot verbessern?
- Wer sind meine Kunden und wo kommen sie her?
- Welche Profile haben meine wichtigsten Kunden?
- Wie unterscheiden sich Besucher von Käufern?
- Welche Produkte werden oft zusammen gekauft? (Cross Selling-Potenziale)
- Welcher Pfad führt Kunden häufig zu einem Kauf/einer Bestellung?
- Welche Seiten werden nach dem Besuch der Startseite aufgerufen?
- Nach welchen Keywords suchen interessante Zielgruppen?
- Welche Partner (Suchmaschinen, Werbung) generieren die meisten/umsatzstärksten Besucher?
- Welche Werbemaßnahmen sollte ich für welche Kunden einsetzen?

Im Rahmen des Web Usage Mining werden personenbezogene Daten verarbeitet. Es sind die entsprechenden Gesetze des Datenschutzes (BDSG, TDDSG) zu berücksichtigen.

Im nächsten Artikel stellen wir einen Ansatz vor, wie Sie die Besucher Ihrer Internetseite während ihres Besuches beraten und kontaktieren können.

Mit hohen Budgets für Internetmarketing versuchen die Unternehmen zwar viele Besucher auf ihre Website zu bringen, eine Kontaktaufnahme zum Zeitpunkt des Besuches unterbleibt jedoch mangels geeigneter Möglichkeiten und fehlender Informationen über den Interessenten.

Gerade der Besuchszeitpunkt ist aber für die Kundengewinnung und die Auftragsakquise von hoher Bedeutung. Der potentielle Kunde beschäftigt sich aktuell mit der Thematik und ist in der Entscheidungsphase. Findet der Internetbesucher nicht das Gewünschte oder gibt es eine offene Frage, verlässt er das virtuelle Unternehmen ohne eine Bestellung oder eine Kontaktaufnahme vorgenommen zu haben. Der nächste Anbieter ist nur einen Mausklick entfernt.

Michael Deinhard
Janine Oswald

Microsoft SQL Server Analysis Services bieten verschiedene implementierte Algorithmen für ein erfolgreiches Web Mining.