# *Web Mining*

Ralf Walther

**Web Mining ist die Übertragung von Data-Mining-Techniken auf Datenmaterial, das im Zusammenhang mit Prozessen im Internet gesammelt wird.**

Schon vor vielen Jahren ist eine der unangenehmen Auswirkungen des Informationszeitalters in den EDV-Abteilungen und Rechenzentren sichtbar geworden. Euphorisch über die neuen technischen Möglichkeiten der Datenverarbeitung wurden und werden immer mehr und immer detailliertere Informationen gespeichert. Die Folge ist unübersichtliches, inkonsistentes und hochgradig redundantes Datenmaterial, dem schon bei einfachen Fragestellungen kaum eine Antwort zu entlocken ist.

Ein Datawarehouse [1] kann in diesem Fall helfen, Ordnung in die Daten zu bringen. In einem Datawarehouse werden alle operativen Datenquellen, die meist heterogen und von sehr unterschiedlicher Qualität sind, zusammengeführt, um sie unternehmensweit und für zukünftige Ansprüche skalierbar zur Verfügung zu stellen. Auch im Falle des Web Minings muss die Datenhaltung, die auch als Data Webhouse bezeichnet werden kann, an die neuen Anforderungen angepasst sein [2].

Auf diesen kumulierten Daten kann anschließend mit Data Mining [3] nach Antworten auf nicht gestellte Fragen gesucht werden. Data Mining bezeichnet den Prozess, der automatisch vorher unbekannte, interessante und interpretierbare Zusammenhänge in großen Datenmengen zu finden vermag. Data Mining ist stets im Kontext von Wissensentdeckung zu sehen (Knowledge Discovery in Databases, KDD), welche den gesamten Prozessablauf bezeichnet, der im Wesentlichen aus Datenaufbereitung, Data Mining und Interpretation der Ergebnisse besteht. Übertragen auf Web Mining wird der Prozess mit „Web KDD" [4, 5] bezeichnet.

## Spuren im Datensand

Durch die rasante Entwicklung des Internets, insbesondere im kommerziellen Sektor, fallen zunehmend auch umfangreiche Daten aus der Benutzung von Internetdiensten an. Dazu zählen primär die Protokolldateien der Web-Server, die Informationen über die angeforderten Inhalte und auch über denjenigen speichern, der sie angefordert hat. Auf diese Weise hinterlässt jeder Internet-Anwender seine Spur in Form von Seitenanfragen (Hits) in den entsprechenden Protokolldateien (Hitlog). Eine weitere Datenquelle ergibt sich, wenn sich der Internetbenutzer aus seiner Anonymität heraus bewegt und sich innerhalb eines Internetdienstes als Person zu erkennen gibt (Name, Anschrift usw.). Navigiert er zudem noch in einem zusammenhängendem Web-Auftritt (z. B. ein e-Commerce-System), können alle Benutzeraktionen, die von dem System angeboten werden, registriert werden. Im einfachen Fall sind dies lediglich die Interaktionen des Benutzers mit der Maus (click stream) bzw. die Navigationsanfragen; sie können jedoch beliebig detailliert sein (z. B. Informationen über Suchanfragen, Bestellvorgänge, Warenkorbanwendungen, Gewinnspiele usw.).

Eine weitere wesentliche Informationsquelle, die in besonderem Maße die Verarbeitung der gesammelten Daten beeinflusst, sind Metainformationen über die zu untersuchenden Seiten. Das ist insbesondere die Topologie (Aufbau und Verbindungen) und das Wissen über die Funktionalität. Jeder Internetseite wohnt eine der

**Ralf Walther**
mindUp Web + Intelligence GmbH, Blarerstraße 56,
D-78462 Konstanz
E-Mail: ralf.walther@mindup.de

* Vorschläge an: Prof. Dr. Frank Puppe, Institut für Informatik, Universität Würzburg, Am Hubland, 97074 Würzburg und Dieter Steinbauer, GEZ, Freimersdorfer Weg 6, 50829 Köln. Eine Liste aller „Aktuellen Schlagwörter" seit 1988 gibt es unter http://ki.informatik.uni-wuerzburg.de/~puppe/as.htm

Anwendung sehr spezifische Funktionalität inne und die Interaktion des Benutzers wird entscheidend von den möglichen Prozessen beeinflusst. Die in einem Auftritt möglichen Prozessabläufe sind zwar sehr individuell, jedoch liefern sie sehr detaillierte und damit sehr wertvolle Informationen.

## Daten sind nicht gleich Daten

Die Datenaufbereitung spielt beim Web Mining dieselbe wichtige Rolle wie beim Data Mining. Aufgrund technischer Begebenheiten (z. B. Caching) sind die Daten zumeist unsauber. Es finden an dieser Stelle die vom Data Mining bekannten Techniken zur Verbesserung der Datenqualität Anwendung. Extremwerte und Randwertartefakte können gefiltert (Outlier/Noise-Elimination) und fehlende Werte können aufgrund von statistischer Schätzung bzw. empirisch erhobenen Werten ersetzt werden. Da die Datenaufbereitungsschritte stets sehr stark auf die Datenquellen ausgerichtet sind, müssen beim Web Mining die internetspezifischen Umstände berücksichtigt werden, mit denen die Daten gesammelt werden. Das Internet besteht nicht nur aus international heterogenen Datenquellen, sondern auch die Benutzer können jeder Herkunft und fast jeden Alters sein. Die Überlegungen für das Data Mining von Kundendaten einer Filialbank sind auf solche Daten nicht ohne weiteres übertragbar.

Die Ziele des Web Minings können in zwei Gruppen aufgeteilt werden: Zum einen geht es darum, die einzelnen Internetseiten auf ihre Inhalte zu untersuchen (Web Content Mining), und zum anderen, das Benutzerverhalten zu erkunden (Web Usage Mining). Häufig wird noch ein zusätzlicher Bereich genannt, der sich explizit mit der Analyse der Struktur im Web befasst (Web Structure Mining). Hier wird dieses Gebiet im Zusammenhang mit Web Content Mining behandelt.

## Web Content Mining

Das Internet bietet zu jedem Thema unzählige Informationen und vielfältige Dateninhalte. Web Content Mining befasst sich mit der Analyse von den im Netz befindlichen Daten. Dazu gehören textuelle und multimediale Informationen jeglichen Formats und auch die Verbindungen (Links) zu den Nachbarseiten.

Ein Ziel kann die Einteilung (Segmentierung) der Daten in thematisch zusammenhängende Bereiche sein. Hierfür können Algorithmen der automatischen Klassifikation (Clusteranalyse) herangezogen werden, wobei die Wahl der Metriken die entscheidende Rolle spielt. Die Ähnlichkeitsberechnung zweier Dokumente kann z. B. mit statistischer Gewichtung von Schlagwörtern als auch über einen aufwendig errechneten Feature-Vektor aus einer linguistischen Analyse erfolgen. Hierbei werden zumeist die partitionierenden den hierarchischen Verfahren vorgezogen, da die Distanzberechnungen hierarchischer Verfahren sehr schnell an Komplexitätsgrenzen stoßen [6].

Auch die Untersuchung der Verbindungen der Seiten untereinander kann helfen, die inhaltliche Verwandtschaft von Webseiten abzuleiten. Dem Entwickler der Webseiten darf unterstellt werden, dass sich sein Wissen über semantische Zusammenhänge zweier Seiten in den entsprechenden Verbindungen ausdrückt. Von Wang und Liu [7] werden Ansätze beschrieben, die das Web als gerichteten Graph modellieren, wobei die Dokumente auf die Knoten und die Links auf die Kanten abgebildet werden. Mittels Graphentransformationen und Bewertung wird nach Mustern in der Struktur des Webs gesucht, die auch Aussagen über den Inhalt der vernetzten Dokumente zulassen.

Neben der Klassifikation von Webseiten aufgrund des Textinhaltes können auch die Ergebnisse der Klassifikation der „Nachbardokumente“ herangezogen werden [8]. Als Techniken werden sowohl klassische statistische Verfahren (K-Nächste-Nachbarn, Diskriminanzanalyse) als auch Verfahren des maschinellen Lernens eingesetzt (künstliche neuronale Netze, Entscheidungsbäume) [9].

## Web Usage Mining

Die Interaktion des Benutzers mit dem Internet ist Zentrum der Fragestellungen des Web Usage Minings. Wie navigiert er durch die Netzstruktur? Welche Inhalte bewegen ihn zu welchem Handeln?

Sofern sich der Benutzer nicht freiwillig zu erkennen gibt (z. B. Benutzer-Login in eine Site), spricht man von nichtpersonalisiertem Web Mining, da der Benutzer anonym seine Seitenanfragen im Hitlog hinterlässt bzw. die Aktionen, die er auf den Seiten auslöst, keiner realen Person zugeordnet werden können. Im anderen Fall spricht man von personalisiertem Web Mining, da der Benutzer namentlich bekannt ist und sogar zusätzliche Datenquellen (z. B. demographische Daten) zur Analyse hinzugezogen werden können.