Erhard Rahm

Web Usage Mining

Mit Web Usage Mining bezeichnet man die Analyse des Nutzungsverhaltens von Websites. Von besonderer Bedeutung ist hierbei die Anwendung von Data-Mining-Verfahren zur Entdeckung von Mustern im Zugriffsverhalten. Eine Vielzahl weiterer Begriffe wird verwendet für Auswertungen des Web-Nutzungsverhaltens, z.B. Clickstream-Analyse, Web-Traffic-Analyse, E-Analytics, E-Intelligence etc.

Eine umfassende quantitative Bewertung der Web-Zugriffe ist erforderlich für eine aussagekräftige Zustandserfassung und gezielte Optimierung eines Web-Auftritts. Dies ist umso dringlicher, je stärker der Erfolg eines Unternehmens, einer Einrichtung etc. von einer effektiven Web-Präsenz abhängt. So ist z.B. in E-Commerce-Unternehmen bzw. -Abteilungen (Online-Buchhändler, Online-Versandhandel, Online-Kaufhäuser, ...) die Wertschöpfung unmittelbar davon abhängig, dass zahlreiche Benutzer die Website nutzen, die von ihnen benötigten bzw. sie ansprechenden Artikel schnell finden und in einfacher Weise Bestellungen tätigen können. Mit Web Usage Mining kann festgestellt werden, in welchem Umfang Nutzer bestimmte Seiten und Pfade verwenden, woraus sich u.a. Hinweise zur Änderung des Web-Auftritts ableiten lassen. Die Effektivität solcher Anpassungen kann durch erneute Anwendung von Web Usage Mining unmittelbar verifiziert werden. Die hohe Dynamik der Web-Nutzung erfordert ohnehin eine kontinuierliche Erfassung und Auswertung der Web-Zugriffe, um eine aktuelle Bewertung des Nutzungsverhaltens erreichen zu können.

Die stark zunehmende Notwendigkeit einer quantitativen Bewertung der Web-Nutzung führte dazu, dass in den letzten Jahren eine Vielzahl von Werkzeugen für bestimmte Aspekte des Web Usage Mining entwickelt wurde und das Thema auch in der Forschung großes Interesse findet [4, 5, 7]. Die meisten Ansätze stellen dabei die Auswertung der von den Webservern geführten Web-Log-Dateien in den Mittelpunkt. In ihnen wird jeder »Hit« auf eine Webseite und den darin enthaltenen Dateien (Bilder, Skripte etc.) protokolliert, womit nahezu jeder Mausklick der Benutzer registriert wird. Einfache Werkzeuge operieren direkt auf diesen Log-Dateien und erzeugen vordefinierte Berichte mit statistischen Angaben zur Häufigkeit von Hits, vollständig angezeigten Seiten (Page-Views) und Besuchen (Sitzungen), den am meisten referenzierten Seiten, Einstiegs- und Ausstiegsseiten, fehlerhaften Zugriffen (z.B. aufgrund ungültiger Verweise) und den so genannten Referrer-Knoten (von woher kommen die Besucher auf die Website). Nutzerspezifische Auswertungen sind im Wesentlichen auf die IP-Adressen der Nutzerrechner sowie auf Angaben zum verwendeten Betriebssystem und Browser beschränkt.

Data-Warehouse-Einsatz

Weiter gehende Lösungsansätze zum Web Usage Mining sind datenbankbasiert und verwenden neben den Web-Log-Daten zusätzliche Datenquellen. Ein Datenbankansatz ist erforderlich, um eine effiziente und skalierbare Verwaltung der riesigen Datenmengen (große Websites produzieren täglich mehrere Gigabytes an Log-Daten) sowie flexible und interaktive Auswertungen zu ermöglichen. Außerdem erlauben die Angaben in den Web-Log-Dateien alleine oft unzureichende inhaltliche (z.B. geschäftsbezogene) und personenbezogene Auswertungen. Solche Auswertungen verlangen die Kopplung der Web-Log-Aufzeichnungen mit Angaben zur Struktur und den Inhalten (z.B. Produkten) einer Website, ggf. Applikationsserver-Protokollen zu den ausgeführten Anwendungsfunktionen sowie ggf. Informationen über Kunden eines Unternehmens. Diese Kopplung geschieht am besten durch Integration der Daten im Rahmen eines *Data Warehouse*, auf dem dann die Auswertungen erfolgen. Ein Warehouse-Ansatz bietet auch die Möglichkeit zur Einbindung des Web Usage Mining im Rahmen eines unternehmensweiten Customer Relationship Management (CRM), bei dem die Beziehungen zwischen Kunden und Unternehmen über alle Interaktionskanäle (Web, E-Mail, Telefon, klassische Post, ...) hinweg erfasst und genutzt werden [3].

Die Umsetzung eines Data-Warehouse-basierten Konzepts zum Web Usage Mining [2, 6] erfordert die Festlegung der Auswertungsziele und davon abgeleitet die Auswahl der Datenquellen, die Modellierung des Data Warehouse (Schemaentwurf) sowie die Festlegung der Verfahren (bzw. Werkzeuge) zur Datentransformation und -integration sowie zur Analyse. Wie bei den meisten Data-Warehouse-Projekten liegt ein Großteil des Aufwands bei der Übernahme und Transformation der Daten aus den einzelnen Quellen in das Data Warehouse. So sind auf den Web-Log-Dateien umfassende Bereinigungen erforderlich, z.B. um Zugriffe von Roboter- und Indexierungsprogrammen sowie Zugriffe auf Hilfselemente wie Skripte, Applets, Multimediadateien etc. zu identifizieren und ggf. zu eliminieren.

Die Zugriffe sind ferner einzelnen Benutzern und Sitzungen zuzuordnen, um das Navigationsverhalten feststellen zu können. Diese Aufgaben können auf Basis der Web-Log-Dateien nur mit Abstrichen gelöst werden, da insbesondere die IP-Adresse oft keine Rückschlüsse auf bestimmte Nutzer zulässt. Personenbezogene Aussagen zum Navigationsverhalten über mehrere Sitzungen hinweg können nur getroffen werden, wenn die Nutzer bereit sind, sich explizit zu identifizieren (etwa über eine Anmeldung) oder – mit geringerer Sicherheit – wenn sie Cookies akzeptieren. Die Bestimmung von Sitzungen und des Navigationsverhaltens wird ferner erschwert durch die Pufferung von Webseiten durch Browser oder Proxy-Rechner, welche Zugriffe beim Webserver einspart und damit zu »Lücken« in den Log-Aufzeichnungen führt. Zur inhaltlichen Bewertung der Zugriffe ist eine Zuordnung der Web-Adressen (URLs) zu anwendungsorientierten Begriffen erforderlich, wie dies über eine hierarchische Inhaltsstrukturierung einer Site oder Konzepthierarchien erreicht werden kann [6, 5]. Eine inhaltliche Einordnung der Web-Zugriffe ermöglicht auch eine bessere Auswertbarkeit bei häufigeren Änderungen im Aufbau einer Site bzw. bei dynamisch generierten Webseiten.

Bewertungsmetriken und Analyseansätze

Die Analysen betreffen einerseits statistische Aussagen zu Zugriffshäufigkeiten und Verweildauer hinsichtlich unterschiedlicher Dimensionen (Zeit, Inhaltsbereiche, Benutzermerkmale, ...), wie sie von Data Warehouses über OLAP-Auswertungen gut unterstützt werden können. Als Bewertungsmetriken sind ferner unterschiedliche *Konversionsraten* von Interesse, die angeben, welche Anteile der Nutzer bestimmte Webseiten aufgesucht bzw. bestimmte Aktivitäten durchgeführt haben. So interessieren im E-Commerce-Umfeld vor allem die Konversionsraten von Besuchern zu Kunden (Käufern) sowie der damit generierte Umsatz. Die Detailauswertung dieser Metriken kann dann genutzt werden zur Bestimmung profitabler (bzw. unprofitabler) Referrer, Produkte, Kundengruppen etc. Zur Erklärung und Optimierung der Konversionsraten ist wichtig, Navigationspfade im Rahmen von Besuchen näher zu bewerten, insbesondere die Übergangswahrscheinlichkeiten (Mikrokonversionsraten) für Zwischenstationen auf dem Weg von einer Einstiegsseite bis zur Bestellung oder anderen Ausstiegsseiten.

Für Data-Mining-Auswertungen zur Entdeckung relevanter Nutzungsmuster kommen unterschiedliche Techniken in Frage, insbesondere Clustering-Verfahren, Klassifikationsansätze sowie Assoziations- und Sequenzregeln [1]. So können Clustering-Verfahren genutzt werden, um ähnliche Sitzungen und damit Nutzergruppen mit ähnlichen Interessen zu entdecken. Die Tatsache, dass Webseiten eines Clusters häufig zusammen genutzt werden, kann ferner Anhaltspunkte zur Optimierung der Site-Strukturierung geben. Klassifikationsansätze erlauben die Zuordnung von Nutzern zu vordefinierten Kategorien, z.B. nach Nutzungshäufigkeit und -dauer oder Interessensgebieten. Damit können dann gruppenspezifische Auswertungen und z.B. die Generierung gezielter Empfehlungen (Verweise) auf bestimmte Inhalte oder Produkte erfolgen. Assoziations- und Sequenzregeln ermitteln, welche Inhalte oder Aktionen oft zusammen im Laufe einer Sitzung oder in aufeinander folgenden Sitzungen eines Benutzers aufgerufen werden. Solche Muster lassen sich ebenfalls für gezielte Hinweise auf potenziell interessante Webseiten nutzen, z.B. um Cross-Selling-Effekte zu erzielen.

Umsetzung der Analyseerkenntnisse

Die aus dem Web Usage Mining gewonnenen Erkenntnisse lassen sich vielfach verwenden, auch ohne dass die Identität der Nutzer bekannt sein muss. Dies betrifft Hinweise zur generellen oder Nutzergruppen-bezogenen Anpassung von Websites: Beseitigung fehlerhafter Verweise, verbesserte Benutzerführung durch angepasste Such- und Navigationsmöglichkeiten, gezielte Empfehlungen, Ergänzung und Aktualisierung der Inhalte und Dienste etc. Die Referrer-Analyse kann für Marketingzwecke genutzt werden, um zu entscheiden, auf welchen Knoten Werbehinweise auf die eigene Site die besten Erfolgsaussichten versprechen (der Erfolg kann danach mit Web Usage Mining bewertet werden). Falls Nutzer z.B. aufgrund einer Anmeldung identifiziert werden können, lassen sich personenbezogene Nutzungsprofile erstellen. Dies kann für personalisierte Empfehlungen und Produktangebote genutzt werden, wie es etwa im One-to-One-Marketing verfolgt wird. Ferner lassen sich aus den Nutzungsprofilen nach bestimmten Kriterien Adressaten für Marketingkampagnen (z.B. E-Mail- oder Briefwerbung) bestimmen. In Deutschland verlangen die Datenschutzbestimmungen die ausdrückliche Zustimmung der betroffenen Personen zu solchen personalisierten Nutzungsformen. Dies gilt natürlich erst recht für weiter gehende Verwendungsformen, wie Verkauf personenbezogener Daten an andere Unternehmen oder Austausch der Daten zur Erstellung unternehmensübergreifender personalisierter Nutzungsprofile.

Ausblick

Derzeit werden die sich aus dem Web Usage Mining ergebenden Möglichkeiten noch kaum genutzt. Dies liegt auch daran, dass insbesondere die Umsetzung eines Data-Warehouse-Ansatzes sowie die Nutzung von Data-Mining-Werkzeugen meist noch einen hohen Aufwand verursachen. Allerdings ist mit einer deutlichen Verbesserung der Situation aufgrund der Weiterentwicklung der Werkzeuge und durch die sich abzeichnende Unterstützung des Web Usage Mining im Rahmen von kommerziellen Datenbanksystemen und Applikationsserver-Software zu rechnen. Auch aus Sicht der Forschung gibt es zahlreiche noch unzureichend untersuchte Fragestellungen, z.B. bei der Datenaufbereitung und -integration, geeigneten anwendungsspezifischen Bewertungsmetriken, der Ausgestaltung spezifischer Data-Mining-Verfahren sowie der Umsetzung der Analyseergebnisse.

Literatur

[1] *Ester, M.; Sander, J.:* Knowledge Discovery in Databases: Techniken und Anwendungen. Springer-Verlag, 2000.

[2] *Kimball, R.; Merz, R.:* The Data Webhouse Toolkit: Building the WebEnabled Data Warehouse. Wiley, 2000.

[3] *Schroeck, M. J.:* E-Analytics – The Next Generation of Data Warehousing. DM Review, Aug. 2000, http://www.dmreview.com

[4] *Srivastava, J.; Cooley, R.; Deshpande, M.; Tan, P.:* Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations 1(2), 12-23, 2000.

[5] *Spiliopoulou, M.:* Web Usage Mining for Web Site Evaluation. Comm. ACM 43 (8), 127-134, 2000.

[6] *Stöhr, T.; Rahm, E.; Quitzsch, S.:* OLAP-Auswertung von Web-Zugriffen. Proc. GI-Workshop Internet-Datenbanken, Sep. 2000, http://dol.uni-leipzig.de/pub/2000-23

[7] WEBKDD-Workshops zu Web Mining (1999, 2000, 2001): http://www.acm.org/sigkdd/proceedings/webkdd99/toconline.htm, http://robotics.Stanford.EDU/~ronnyk/WEBKDD2000, http://robotics.Stanford.EDU/~ronnyk/WEBKDD2001

Erhard Rahm
Universität Leipzig
Institut für Informatik
Augustusplatz 10-11
04109 Leipzig
rahm@informatik.uni-leipzig.de
http://dbs.uni-leipzig.de