11 Data-Warehouse-Einsatz zur Web-Zugriffsanalyse

Erhard Rahm, Thomas Stöhr

Kurzfassung

Die Analyse des Nutzungsverhaltens von Websites ermöglicht wichtige Hinweise zur Optimierung und Weiterentwicklung eines Web-Auftritts. Skalierbarkeit und Flexibilität der Auswertungen verlangen oft eine datenbankbasierte Realisierung. Wir diskutieren hierzu verschiedene Varianten, insbesondere den Einsatz eines »Web Data Warehouse«, in dem neben den Web-Log-Daten Informationen zu Nutzern/Kunden, Inhalten/Produkten und Anwendungsfunktionen integriert werden. Weiterhin geben wir einen Überblick zu derzeit verfügbaren Werkzeugen für die Web-Zugriffsanalyse.

11.1 Einführung

Eine aussagekräftige quantitative Analyse und Bewertung der Nutzung von Websites wird immer wichtiger. Durch eine derartige *Web-Zugriffsanalyse* sollen die vom Web-Server protokollierten Zugriffe sowie weitere Informationen ausgewertet werden, um ein möglichst genaues Bild vom Zugriffsverhalten der Nutzer einer Website zu erhalten. Mit den Auswertungen wird die Antwort auf eine Vielzahl von technischen, inhaltlichen und nutzerbezogenen Fragestellungen angestrebt, u.a.

- Wie gut ist die Leistung der Website (Durchsatz, Antwortzeit, Datenvolumen)?
- Welche Seiten / Inhalte / Produkte interessieren die Nutzer (bzw. Kunden) am meisten, welche werden dagegen kaum genutzt bzw. nicht gefunden?
- Wie ist die zeitliche Entwicklung beim Zugriff (aktuell im Vergleich zur Vorwoche, Vormonat etc.)?
- Woher kommen die Besucher (Suchmaschine, Portal, Anklicken eines Werbe-Banners etc.)?
- Wie ist das Navigationsverhalten innerhalb der Website (wichtige Pfade innerhalb von Sitzungen, Einstiegs- und Ausstiegsseiten)?
- Wer kommt auf die Website?
- Wie hoch ist die Rückkehrquote von Nutzern?
- Wie hoch ist die Kaufquote?
- Wie zufrieden sind die Nutzer? Warum haben sie die Site verlassen?

- Welche Informationen bzw. Produkte werden häufig zusammen betrachtet bzw. gekauft?
- Wie wirken sich Änderungen in der Strukturierung und im Design einer Website aus?
- Wie wirken sich Marketing-Aktivitäten (z.B. Banner-Werbung, Print-Anzeige etc.) im Zugriffsverhalten aus?
- Wie ist die Wirtschaftlichkeit der Website (Return on Investment)?

Die Ergebnisse der Web-Zugriffsanalyse sollen helfen, die Zielsetzungen eines Web-Auftritts besser zu erreichen. Damit bestimmen diese Zielsetzungen die benötigten Auswertungen oder Bewertungsmetriken und auch den Realisierungsaufwand einer Web-Zugriffsanalyse. Ein umfassender Bewertungsansatz ist besonders kritisch für E-Commerce-Unternehmen (Online-Buchhändler, Online-Versandhandel, Online-Kaufhäuser, ...). Sie benötigen möglichst genaues Wissen über die Interessen und Bedürfnisse ihrer Nutzer und Kunden, um diese optimal bedienen zu können (Gewinnung neuer Kunden, Ausbau der Kundenbindung). Aber auch andere Unternehmen und nicht-kommerzielle Einrichtungen können eine Web-Zugriffsanalyse nutzen, um generelle Ziele wie einfache und schnelle Bereitstellung relevanter Informationen, Gewinnung neuer Nutzer, Erhöhung der Rückkehrquote von Nutzern etc. besser zu erreichen.

In den letzten Jahren hat das Thema der Web-Zugriffsanalyse in Forschung und Praxis große Bedeutung erlangt [SCDT00, Spil00, WKDD01]. Eine Vielzahl von Werkzeugen für bestimmte Aspekte ist bereits verfügbar, insbesondere zur Auswertung der von den Web-Servern geführten *Web-Log-Dateien,* in denen nahezu jeder Maus-Klick der Website-Nutzer registriert wird. Neben der von uns verwendeten Bezeichnung der Web-Zugriffsanalyse findet man viele weitere Begriffe, u.a. *Web Usage Mining*, *Clickstream-Analyse*, *Web Traffic-Analyse*, *E-Analytics*, *E-Intelligence* etc. Diese Namen orientieren sich dabei teilweise an den ausgewerteten Daten (z.B. Clickstream-Log), den Analyseverfahren (z.B. einfache »Traffic«-Bewertungen, »Business Intelligence«-Auswertungen oder Data-Mining-Analysen) und dem Anwendungsschwerpunkt (z.B. E-Commerce).

In diesem Kapitel geben wir einen Überblick zu wesentlichen Realisierungsaspekten einer umfassenden Web-Zugriffsanalyse und den derzeit verfügbaren Werkzeugen. Dazu stellen wir im nächsten Kapitel einfache dateibasierte Ansätze einer Datenbank-Lösung gegenüber und plädieren für den Einsatz eines Data-Warehouse-Ansatzes, der durch Integration unterschiedlicher Datenquellen auch inhaltliche und benutzerbezogene Auswertungen ermöglicht. Zur Umsetzung des Warehouse-Ansatzes diskutieren wir zunächst die relevanten Datenquellen sowie wichtige Teilaufgaben bei der Datentransformation (Bereinigung von Log-Daten, Bestimmung von Sitzungen, Nutzerzuordnung). Anschließend stellen wir ein einfaches Datenbankschema (Stern-Schema) vor, das vielfältige Auswertungen zulässt, und geben einen Überblick über wesentliche Auswertungsansätze, insbesondere Data-Mining-Verfahren. Nach einer Diskussion von Möglichkeiten zur Umsetzung der Analyseergebnisse folgt der Überblick zu derzeitigen Werkzeugen.

11.2 Datei- vs. datenbankbasierte Realisierungsansätze

Einfache Werkzeuge und Verfahren zur Web-Zugriffsanalyse werten lediglich die Web-Log-Dateien des Web-Servers aus. Damit erstellt man im Wesentlichen vordefinierte statistische Auswertungen auf den Daten, die in diesen Log-Dateien pro Zugriff vermerkt werden (s.u.). Beispiele sind zeitliche Entwicklung der Zugriffshäufigkeit und -verteilung auf bestimmte Seiten (URLs) und Rechner (IP-Adressen), transferierte Datenmengen, Häufigkeit fehlerhafter Zugriffe etc. Die dateibasierten Ansätze ermöglichen zwar bereits eine Reihe von Erkenntnissen, weisen jedoch wesentliche Beschränkungen auf:

- *Datenmengen:* Die Auswertungen auf Dateiebene erfordern das sequenzielle Lesen der kompletten Log-Daten und sind somit nur für kleinere Datenmengen ausreichend schnell erstellbar. Größere Websites produzieren jedoch sehr große Mengen an Daten – bis zu mehrere Gigabytes pro Tag. Für zeitliche Vergleichsanalysen sollten diese Daten über Zeiträume von einigen Jahren genutzt werden können, was leicht zu Datenmengen im höheren Terabyte-Bereich führt.
- *Flexibilität der Auswertungen:* Die vordefinierten Auswertungsberichte können nur Basisinformationen liefern, nicht jedoch die gezielte Analyse einzelner Aspekte. Erforderlich ist ein Spektrum an Analysemöglichkeiten von vordefinierten Berichten, interaktiven Abfragen bis hin zu Data-Mining-Analysen.
- *Inhaltliche und benutzerbezogene Auswertungen:* Die Web-Log-Daten unterstützen primär technische Auswertungsaspekte, jedoch keine ausreichenden inhaltlichen (z.B. produktbezogene) und nutzer/kundenspezifischen Auswertungen. Diese erfordern die Verknüpfung mit weiteren Daten wie Angaben zu Struktur und Inhalten einer Website sowie Nutzer-/Kundenmerkmalen.

Die Behebung dieser Nachteile verlangt eine leistungsfähige Datenbank-Lösung, da nur so eine hohe Skalierbarkeit und Flexibilität der Auswertungen erreicht werden kann. Je nach dem angestrebten Auswertungsumfang kommen hierbei wiederum mehrere Varianten in Frage, u.a.:

- Verwendung einer *»einfachen« Datenbank* zur Speicherung und Auswertung der Web-Log-Daten (ohne Integration weiterer Daten),
- Verwendung eines *dedizierten Data Warehouse* zur Web-Zugriffsanalyse, wobei neben den Web-Log-Daten weitere Datenquellen für inhaltliche und nutzerspezifische Auswertungen integriert werden,
- Einbettung der Web-Zugriffsanalyse in ein *Unternehmens-Data-Warehouse*.

Der erstgenannte Ansatz verursacht den geringsten Realisierungsaufwand und unterstützt bereits größere Datenmengen und bessere Auswertungsmöglichkeiten als dateibasierte Verfahren. Allerdings können die Auswertungsziele nur durch die Integration mehrerer Datenquellen umfassend erreicht werden, wie von Data-Warehouse-Lösungen ermöglicht. Der Data-Warehouse-Ansatz [BaGü01] hat sich in den letzten Jahren unabhängig von dem hier betrachteten Anwendungsge-

biet als Schlüsseltechnologie für entscheidungsunterstützende Analysen insbesondere in größeren Unternehmen etabliert. Kennzeichnende Merkmale sind, dass die auszuwertenden Daten aus mehreren Datenquellen extrahiert und im Rahmen einer separaten Datenbank integriert werden. Dieses Data Warehouse wird periodisch (z.B. täglich) aus den Quelldaten aktualisiert und ist von seiner Struktur und Realisierung auf Analyseerfordernisse ausgerichtet. Im Rahmen so genannter *OLAP*-Auswertungen (Online Analytical Processing) können interaktiv auch auf sehr großen Datenmengen eine Vielzahl von Auswertungen erfolgen. Daneben werden weitere Auswertungen wie Erstellung vordefinierter Berichte oder Data-Mining-Analysen unterstützt.

Durch Integration verschiedener Datenquellen sowie die umfassenden Auswertungsmöglichkeiten sind Data-Warehouse-Lösungen für die Web-Zugriffsanalyse besonders geeignet [KiMe00, StRQ00, ScNL01]. Eine Variante ist dabei die Einbettung der Web-Nutzungsdaten innerhalb eines Unternehmens-Data-Warehouse zur Realisierung eines umfassenden *Customer Relationship Management (CRM)*, bei dem die Beziehungen zwischen Kunden und Unternehmen über alle Interaktionskanäle (Web, E-Mail, Telefon, klassische Post, ...) hinweg erfasst, analysiert und genutzt werden [Schr00]. Die CRM-Spezifika sollen jedoch im Weiteren nicht vertieft werden.

Abb. 11-1 zeigt die Grobarchitektur eines datenbankbasierten Ansatzes zur Web-Zugriffsanalyse mit dediziertem Data Warehouse (»Web Data Warehouse«). Die protokollierten Zugriffe auf der Website werden kontinuierlich transformiert und in das Warehouse überführt. Daneben erfolgt die Integration von Angaben zu Kunden bzw. Nutzern, der Site-Struktur, zu Produkten/Inhalten etc. Auf dem integrierten und periodisch aktualisierten Datenbestand des Web Data Warehouse erfolgen dann unterschiedliche Auswertungen sowie Data-Mining-Analysen, z.B. zur Entdeckung von Mustern im Navigationsverhalten, zur Analyse des Kaufverhaltens und zur Segmentierung von Kunden. Die Umsetzung der Analyseerkenntnisse kann zu Anpassungen im Web-Auftritt oder zur Durchführung bestimmter Marketing-Aktivitäten führen, deren Auswirkungen dann wieder im Rahmen der Web-Zugriffsanalyse überprüfbar sind. Dieser geschlossene Kreislauf (»closed loop«) ist wesentlich für den Erfolg und verdeutlicht die Vorteile des Interaktionskanals »Web«, der eine schnelle Reaktion auf geänderte Verhältnisse und eine kontinuierliche Kontrolle und Analyse der Effektivität gestattet.

Die Vorteile einer Data-Warehouse-Lösung verlangen allerdings auch einen relativ hohen Realisierungsaufwand, insbesondere bei der Aufbereitung und Integration der Daten. Auf die dabei anfallenden Probleme wird im Folgenden näher eingegangen.

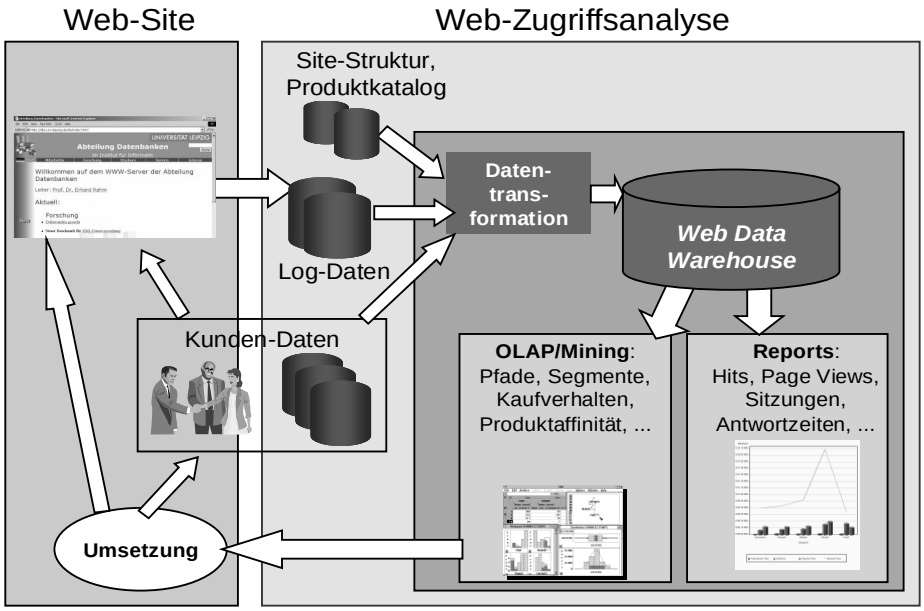

Abb. 11-1: *Grobarchitektur einer Data-Warehouse-basierten Web-Zugriffsanalyse*

11.3 Datenquellen und Datentransformation

Zur Umsetzung eines Warehouse-Ansatzes zur Web-Zugriffsanalyse betrachten wir in diesem Abschnitt die wesentlichen Datenquellen sowie die wichtigsten Schritte der Datentransformation. Hierzu diskutieren wir zunächst den Aufbau der Web-Log-Dateien sowie die Verwendung weiterer Datenquellen (Site-Struktur, Nutzerinformationen, Inhaltskategorisierung etc.). Zur Datentransformation werden Aufgaben der Datenbereinigung, Sitzungs- und Nutzeridentifikation behandelt.

11.3.1 Aufbau der Web-Log-Dateien

Die Zugriffe auf die Inhalte einer Website werden von Web-Servern verwaltet, welche jeden von ihnen bearbeiteten Dateizugriff oder »Hit« (Aufruf einer HTML-Seite, Zugriff auf Bilddatei, Skript-Ausführung, ...) in einer Log-Datei protokollieren. Jeder Zugriff wird durch einen Log-Satz in einem bestimmten Format repräsentiert, wobei vor allem das Common Log Format (CLF) sowie dessen Erweiterungen weit verbreitet sind. Ein Beispieleintrag dafür ist

```
green.dresdnerbank.de - - [11/Jul/2002:15:33:13 +0200]        // Host, Zeitpunkt
"GET /skripte/WMS/inhalt2.html HTTP/1.0" 200 7885             // Request, Status, #Bytes
"http://www.google.de/search?q=workflow&hl=de&meta=" "        // Referrer
Mozilla/4.0 (compatible; MSIE 5.01; Windows NT; drebaIE-ZE)"  // Nutzer-Agent
```

Wesentliche Komponenten der Einträge sind:

- *Host:* Name bzw- IP-Adresse des Rechners, von dem der Zugriff erfolgte,
- *Zeitpunkt* des Zugriffs (Datum, Uhrzeit, Zeitzone),
- *HTTP-Request*, der insbesondere die URL der aufgerufenen Datei enthält sowie möglicherweise Aufrufparameter im Fall von Skripten und dynamischen Webseiten,
- *Status*: standardisierte Kennzeichnung des Zugriffsstatus, z.B. 200 (ok), 304 (not modified since last retrieval) oder 404 (not found),
- *Bytes*: übertragene Datenmenge (bei Status 304 in der Regel 0, da eine aktuelle Version des Objekts in einem Cache vorliegt),
- *Referrer:* URL der Seite, von der aus der Nutzer zugegriffen hat; diese Seite kann auf derselben Website oder extern liegen sowie (für dynamische Seiten) Aufrufparameter enthalten, z.B. die Suchbegriffe bei der Ergebnisseite einer Suchmaschine (s. Beispieleintrag),
- *Nutzer-Agent:* Kennzeichnung des Browsers und des Betriebssystems des Nutzers.

Die beiden letzten Angaben sind nicht im CLF-Standard enthalten, werden jedoch von den meisten Web-Servern unterstützt. Daneben können diese noch für eine Protokollierung weiterer Angaben wie Cookies (s.u.) oder anwendungsspezifische Erweiterungen konfiguriert werden.

11.3.2 Weitere Datenquellen

Zur strukturellen, nutzerbezogenen und inhaltlichen Auswertung der Web-Zugriffe ist eine Verknüpfung mit Informationen weiterer Datenquellen erforderlich, u.a:

- Angaben zur *Struktur* der Website, insbesondere zur Vernetzung der Webseiten (Hypertext-Struktur) sowie von Navigationsmöglichkeiten aufgrund dynamischer Anfragen.
- *Nutzerbezogene Informationen*, wie bereits angelegte Nutzerprofile oder Kundenangaben aus Unternehmensdatenbanken. Möglicherweise können auch allgemeine demografische Informationen herangezogen werden, z.B. um aus Angaben wie Adressen Rückschlüsse auf Einkommen etc. zu gewinnen.
- Angaben zur *Anwendungsfunktion* einzelner Webseiten bzw. Seitenbereiche. Dies erfordert zum einen eine auf die Auswertungserfordernisse ausgelegte Kategorisierung von Anwendungsfunktionen und zum anderen eine Zuordnung der Webseiten bzw. Seitenbereiche zu diesen Kategorien. Die Kategorisierung kann eine einfache Aufzählung sein oder eine mehrstufige (hierarchische) Aufgliederung. So könnte z.B. eine Kategorisierung von Webseiten nach Navigations- und Inhaltsseiten erfolgen und bei den Inhaltsseiten eine Unterteilung nach Funktionen wie Produktansicht, Bestellseite etc.

- Angaben zu den *Inhalten* einzelner Webseiten bzw. dynamischer Anfragen. Analog zur Festlegung der Seitenfunktion ist hier eine Kategorisierung der Inhalte und die Zuordnung zu Webseiten (Web-Zugriffen) erforderlich. Bei einem Online-Shop empfiehlt sich z.B. eine mehrstufige Kategorisierung der Produkte analog zu einem Produktkatalog, um Auswertungen bezüglich unterschiedlicher Produktgruppen zu unterstützen (z.B. Anteil der Zugriffe auf Fachbücher vs. Belletristik, Bestellungshäufigkeit für Computer-Bücher gegenüber allen Fachbüchern etc.). Die inhaltliche Zuordnung von Web-Zugriffen zu den Inhaltskategorien erfordert oft die Berücksichtigung dynamischer Zugriffsparameter (z.B. Produktnummer).
- Werden die aufgerufenen Funktionen weitgehend unter Kontrolle von Applikations-Servern durchgeführt, sind zur funktionalen und inhaltlichen Auswertung der Website-Nutzung die *Log-Dateien der Applikations-Server* als Datenquelle zusätzlich auszuwerten.

Als Alternative zu den Web-Log-Dateien kommt die Aufzeichnung der Datentransfers auf Netzwerkebene durch so genannte *Packet Sniffer* als Datenquelle in Frage. Dieser Ansatz kann auch bei größeren Websites mit mehreren Web-Servern sämtliche Zugriffe in der chronologischen Reihenfolge erfassen und somit ein aufwändiges Mischen mehrerer Log-Dateien umgehen. Andererseits gibt es Beschränkungen, z.B. können verschlüsselte Nachrichtenpakete nicht interpretiert werden und Web-Server-Techniken wie die Auswertung von Cookies werden erschwert.

Die von den Web-Servern registrierbaren Seitenzugriffe repräsentieren das Zugriffsverhalten nicht vollständig. Denn zur Reduzierung des zu übertragenden Datenvolumens sowie zur Unterstützung schneller Antwortzeiten erfolgt an mehreren Stellen auf dem Weg zwischen dem Browser eines Benutzers zum Web-Server eine Pufferung (Caching) von Webseiten, insbesondere durch die Browser selbst sowie bei vermittelnden Rechnerknoten, z.B. Proxy-Rechnern in Unternehmen oder beim Internet Service Provider (ISP). Entsprechende Aufzeichnungen der Zugriffe bei den Browsern bzw. Proxy-Rechnern stellen somit prinzipiell weitere Datenquellen zur Web-Zugriffsanalyse dar. Ein Problem dabei ist die potenziell große Anzahl solcher Datenquellen, die zudem weitgehend außerhalb der Kontrolle eines Website-Betreibers liegen. Insbesondere dürften nur wenige Benutzer bereit sein, ihre Web-Zugriffe z.B. durch ein entsprechendes Browser-Plugin aufzeichnen zu lassen und diese Zugriffsdaten für externe Auswertungen zur Verfügung zu stellen. Eine Abmilderung des Problems besteht darin, Heuristiken zur Erkennung von Zugriffslücken im Log zu verwenden und eine so genannte *Pfadvervollständigung* durchzuführen (s.u.).

11.3.3 Datentransformationen

Bei der Übernahme und Integration der Daten aus den verschiedenen Quellen in das Web Data Warehouse sind umfassende Datentransformationen erforderlich.

Dies betrifft insbesondere die Web-Log-Einträge, auf deren Behandlung wir uns in diesem Abschnitt konzentrieren wollen. Die Angaben aus den weiteren Datenquellen können vergleichsweise einfach in das Schema des Web Data Warehouse (Kapitel 11.4) integriert werden. Zur Datentransformation von Web-Log-Einträgen besprechen wir im Weiteren die Datenbereinigung, die Bestimmung von Sitzungen, die Nutzeridentifikation sowie die Pfadvervollständigung. Diese Aufgaben sollten durch ein Werkzeug behandelt werden, das die gängigen Web-Log-Formate unterstützt. Bei mehreren Web-Servern sind die Log-Dateien zu mischen. Dies erfolgt am einfachsten über die Zeitstempel der Log-Einträge, vorausgesetzt die Uhrzeiten der Web-Server sind eng synchronisiert [KiMe00].

Die Notwendigkeit der Datentransformationen erkennt man bereits daran, dass die Anzahl der Log-Einträge und damit der Hits nur eine sehr grobe Angabe zur Nutzungsintensität einer Website und ihrer Bereiche entsprechen, obwohl diese Metrik noch häufig Verwendung findet. Dies hängt u.a. damit zusammen, dass diese Zahl sehr stark von Zugriffen z.B. durch Roboter-Programme verfälscht sein kann. Ausserdem besteht eine große Abhängigkeit zum jeweiligen Web-Design, da jedes in eine Seite eingebettete Bild (z.B. Button), Skript etc. einen eigenen Hit verursacht. Diese Effekte müssen im Rahmen der Datentransformation bereinigt werden, um auch Kennzahlen wie die Anzahl der vollständig angezeigten Seiten (Page Views), die Anzahl der Sitzungen (Besuche) sowie die Anzahl der Nutzer (Besucher) bestimmen zu können. Abb. 11-2 verdeutlicht, dass diese Größen in 1:n-Beziehungen zueinander stehen, d.h., pro Besucher gibt es eine oder mehrere Sitzungen, pro Sitzung einen oder mehrere Page Views und pro Page View ein oder mehrere Hits. Weiterhin ist die Anzahl der Sitzungen und Nutzer deutlich aussagekräftiger für die Nutzungsintensität als die Anzahl der Hits.

Abb. 11-2: *Von vielen Hits zu wenigen Besuchern*

Zur Verdeutlichung stellt Abb. 11-3 diese und einige weitere aus den Web-Log-Daten abgeleitete technische Größen für zwei einfache Websites gegenüber. Man erkennt, dass damit schon einige Beurteilungen der Websites unterstützt werden. So weist der erste Server (Lehrstuhl) pro Tag fast die fünffache Anzahl von Hits, jedoch weniger Besuche als der zweite Server (Dokumenten-Server) auf, da in ersterem Fall (aufgrund des Designs) durchschnittlich mehr als doppelt so viele Hits pro Seite und fast dreimal so viele Seiten pro Sitzung anfallen. Die beim zweiten Server primär abgerufenen, relativ voluminösen Dokumente führen dort zu einem höheren Gesamtdatenvolumen. Die Anzahl der übertragenen Dateien ist generell kleiner als die Anzahl der Hits, wegen der Nutzung von Caches (Status 304) und Fehlern beim Zugriff (Datei nicht vorhanden etc.). Für den ersten Server war der

Caching-Effekt besonders ausgeprägt (begünstigt durch die große Zahl von Hits pro Sitzung), wo nur für 70% der Hits eine Dateiübertragung stattfand.

	dbs.uni-leipzig.de (Lehrstuhl Datenbanken)	dol.uni-leipzig.de (Dokumenten-Server)
#Hits pro Tag	10000	2100
#übertragene Dateien pro Tag	7000	1900
#Seiten (page views) pro Tag	3200	1400
#Sitzungen pro Tag	430	540
Datenvolumen in MB pro Tag	77	190
#Hits / Seite	3,1	1,5
#Seiten / Sitzung	7,4	2,6
KB / Datei	11	100

Abb. 11-3: *Bewertungskennzahlen zweier Beispiel-Websites (Jan. 2002)*

Datenbereinigung (Data Cleaning)

Hierbei sind alle für die Auswertungen irrelevanten bzw. verfälschenden Zugriffe zu identifizieren und ggf. zu eliminieren, wodurch der Datenumfang oft signifikant reduziert werden kann. Dies betrifft typischerweise die Zugriffe auf in den explizit angeforderten Seiten eingebetteten Hilfsdateien (*auxiliary resources*) wie Bilder und andere Multimedia-Inhalte, Stylesheet-Dateien, Applets, Skripte etc. Weiterhin sind die Zugriffe von Roboter-Programmen (Crawlern), die z.B. zur Unterstützung von Suchmaschinen (Kap. 7) das Web ständig durchkämmen, zu identifizieren. Hierzu gibt es mehrere Möglichkeiten, z.B. das Vorliegen bestimmter IP-Adressen und Browser-Bezeichnungen oder das Erkennen einer für menschliche Nutzer unerreichbar hohen Zugriffsfrequenz. Auch von Site-spezifischen Indexierungsprogrammen verursachte Log-Einträge sowie Zugriffe lokaler Benutzer sind möglicherweise zu extrahieren. Die jeweils notwendigen Filteraktionen sind site- und auswertungsabhängig und sollten somit durch das Transformationswerkzeug konfigurierbar sein. Weitere Bereinigungsaktionen betreffen die Vereinheitlichung und Ergänzung von Datensätzen wie zur Namensauflösung für numerische IP-Adressen (*reverse DNS lookup*), um z.B. 139.18.2.117 durch db1.informatik.uni-leipzig.de zu ersetzen.

Sitzungs- und Nutzeridentifikation

Um das Nutzungsverhalten von Websites bewerten zu können, ist die Bestimmung von Sitzungen unabdingbar. Eine Sitzung besteht dabei aus der Sequenz von Web-Zugriffen eines Nutzers. Somit erfordert die Bestimmung der Sitzungen bereits eine Zuordnung von Log-Einträgen zu Nutzern. Dann rechnet man üblicherweise alle Zugriffe eines Nutzers, deren zeitlicher Abstand einen Grenzwert (z.B. 30 Minuten) nicht übersteigt, derselben Sitzung zu. Mit dieser Vorgehensweise wird die Aufgabe der Sitzungsidentifikation im Wesentlichen auf das Problem der Nutzeridentifikation abgebildet. Diese ist jedoch im Web-Kontext ein

großes Problem, da die überwiegende Zahl der Zugriffe anonym erfolgt und das verwendete HTTP-Protokoll zustandslos arbeitet, d.h. jeder Web-Zugriff eines Nutzers wird unabhängig von seinen vorhergehenden abgewickelt.

Die Nutzeridentifikation ist offenbar über die Sitzungsidentifikation hinaus von großer Bedeutung für die Web-Zugriffskontrolle, da nur die Berücksichtigung möglichst aller Sitzungen eines Nutzers ein aussagekräftiges Bild über seine Interessen und Nutzungsgewohnheiten gestattet. Dies erfordert insbesondere, dass ein Nutzer bei einer neuen Sitzung wiedererkannt werden kann. Die weitreichendsten Erkenntnisse werden möglich, wenn der Nutzer darüber hinaus auch personifiziert werden kann, d.h. seine personenbezogenen Daten wie Name oder E-Mail-Adresse bekannt sind (z.B. ein wiederkehrender Kunde). Wir unterscheiden somit drei Stufen der Nutzeridentifikation mit zunehmender Aussagekraft, aber auch wachsender Schwierigkeit der Umsetzung:

- temporäre Nutzeridentifikation für die Dauer einer Sitzung,
- sitzungsübergreifende Nutzeridentifikation, insbesondere Erkennung wiederkehrender Nutzer,
- Personifizierung.

In den ersten beiden Fällen bleibt der Nutzer namentlich unbekannt und somit anonym.

Ansatz	sitzungsbezogene Nutzeridentifikation	sitzungsübergreifende Nutzeridentifikation	Personifizierung	Abdeckungsgrad
IP-Adresse + Agent + Referrer	o / +	–	– –	++
temporäre Cookies	+	n.a.	n.a.	+
Session-IDs in URL / versteckten Feldern	+	n.a.	n.a.	+
persistente Cookies	+	+	o	+
Nutzer-Registrierung	++	++	++	–

Abb. 11-4: *Grobbewertung von Ansätzen zur Nutzeridentifikation*
(++ sehr gut, + gut, o mittel, -- schlecht, -- sehr schlecht, n.a. nicht anwendbar)

Im Folgenden diskutieren wir Realisierungsansätze für die drei Arten der Nutzeridentifikation. Abb. 11-4 zeigt hierzu eine zusammenfassende Bewertung hinsichtlich der Eignung für die einzelnen Identifizierungsvarianten (Genauigkeit der Identifizierung). Als weitere Bewertungsgröße wird zur Nutzbarkeit der Ansätze noch der erreichbare Abdeckungsgrad abgeschätzt, d.h., welcher Anteil der Zugriffe ist aufgrund von notwendiger Nutzerkooperation bzw. erforderlichen Browser-Einstellungen mit einem Ansatz überhaupt behandelbar. Man erkennt bereits, dass eine korrekte Identifizierung aller Nutzer unmöglich ist und dass die

Verfahren (und damit die Werkzeuge, die sie einsetzen) sich in Genauigkeit und Abdeckungsgrad teilweise stark unterscheiden. Die drei erstgenannten Ansätze eignen sich nur zur sitzungsbezogenen Nutzeridentifikation.

Temporäre Nutzeridentifikation / Sitzungsidentifikation

Einige einfache Auswertungsansätze verwenden lediglich die Host-Angabe (IP-Adresse) zur Nutzeridentifikation, d.h., alle Log-Sätze mit derselben Host-Adresse werden einem Nutzer zugerechnet. Dies ist jedoch selbst für die temporäre Nutzeridentifikation sehr ungenau, da – gerade zur Anonymisierung – IP-Adressen zunehmend für jeden Zugriff dynamisch generiert werden (z.B. von Internet-Service-Providern). Eine wesentlich bessere Zuordenbarkeit wird erreicht, wenn neben der IP-Adresse (ggf. reduziert auf den Teil, der nicht dynamisch variiert wird) noch die Nutzer-Agent-Angabe (Browser und Betriebssystem) sowie die Referrer-Information hinzugenommen werden. Damit würden im Beispiel von Abb. 11-5 die (in eine relationale Repräsentation überführten) Einträge 27101, 27103 und 27104 demselben Nutzer und derselben Sitzung zugeordnet, trotz der dynamischen Variation im Präfix der Host-Angabe.

Nr	Host	Zeitpunkt	URL	Referrer	Browser	Betriebssystem
27101	cw06.a1.srv.t-online.de	... 16:27:13	A	google.de	MSIE 5.5	NT 5.0
27102	proxy2.sbs.de	... 16:27:14	A	-	MSIE 6.0	Windows98
27103	cw08.a1.srv.t-online.de	... 16:27:43	B	A	MSIE 5.5	NT 5.0
27104	cw10.a1.srv.t-online.de	... 16:28:02	C	B	MSIE 5.5	NT 5.0

Abb. 11-5: *Sitzungsidentifikation mit dynamischen IP-Adressen (Beispiel)*

Alternativen bezüglich der temporären Nutzeridentifikation bzw. Sitzungsidentifikation sind die Generierung einer expliziten Sitzungs-ID durch den Web-Server und deren Austausch mit dem Browser bei jeder folgenden Interaktion. Hierzu sind folgende Ansätze gebräuchlich:

- Ablage der ID innerhalb eines temporären Cookies (Session Cookie), das vom Browser auf dem Rechner des Nutzers gespeichert und vom Web-Server bei jedem erneuten Zugriff gelesen wird. Die Lebensdauer dieser Cookies ist eng begrenzt und läuft mit Beendigung einer Browser-Nutzung ab.
- Speicherung der ID innerhalb der an den Browser zurückgelieferten HTML-Seiten, entweder in versteckten Eingabefeldern oder durch dynamische Modifikation aller in der Seite vorkommenden URLs.

Die ID-Werte werden vom Web-Server beim erneuten Zugriff des Nutzers identifiziert und im Web-Log protokolliert und können somit zur temporären Nutzeridentifikation herangezogen werden. Wesentlicher Nachteil dieser Ansätze ist der vom Web-Server zu leistende Zusatzaufwand, der auch die Zugriffszeiten für den

Benutzer verlängert. Ausserdem gibt es Abhängigkeiten zu Browser-Einstellungen (z.B. Akzeptanz von temporären Cookies).

Sitzungsübergreifende Nutzeridentifikation

Der am weitesten verbreitete Ansatz hierbei ist der Einsatz *persistenter Cookies*. Die Cookies enthalten insbesondere eine eindeutige Nutzer-ID und werden dauerhaft auf der Festplatte des Nutzer-Rechners gespeichert. Über das vom Web-Server gelesene Cookie kann somit ein wiederkehrender Nutzer erkannt und z.B. eine zugeschnittene Nutzung der Website ermöglicht werden (keine wiederholte Eingabe von Präferenzen, Kennwörtern etc.). Dies erfordert seitens der Website die Verwaltung von über Cookies identifizierten Nutzerprofilen, die jedoch die Anonymität der Nutzer wahren, solange keine personenbezogenen Angaben erfasst werden. Eine Einschränkung bei Cookies ist, dass sie nur einen Browser auf einem bestimmten Rechner, also nicht unbedingt eine bestimmte Person kennzeichnen (Nutzung des Rechners durch mehrere Personen). Ausserdem kann von Personen, die mehrere Rechner bzw. Browser nutzen, nur ein Teil der Zugriffe korrekt zugeordnet werden. Schließlich wird der Grad der erreichbaren Nutzeridentifikation dadurch begrenzt, dass die Cookies seitens der Nutzer abgelehnt bzw. gelöscht werden können. Derzeit akzeptieren die weitaus meisten Web-Nutzer offenbar das Anlegen persistenter Cookies, jedoch mehren sich die Bedenken aus Datenschutzgründen. Die Website-Betreiber sollten daher zur Vertrauensbildung im eigenen Interesse die Inhalte und Verwendung der angelegten Nutzerprofile offenlegen, auch wenn keine personenbezogenen Angaben gespeichert werden.

Eine zuverlässigere Alternative zur sitzungsübergreifenden Nutzeridentifikation ist die *Nutzer-Registrierung*, welche anonym (unter einem Pseudonym) oder personenbezogen erfolgen kann. Dabei muss sich der Nutzer registrieren und bei jeder neuen Sitzung durch Angabe des Registrierungskennworts explizit identifizieren (während einer Sitzung kann der Web-Server die Kennung wie im Falle der Sitzungs-IDs mit dem Browser austauschen). Die explizite Identifizierung ist für die Nutzer sehr lästig und wird somit meist nur von einem kleinen Anteil (potenzieller) Nutzer akzeptiert. Auch ist bei Verwendung von Pseudonymen eine Mehrfach-Registrierung möglich.

Personifizierung

Die personenbezogene Identifizierung wird meist über eine Nutzer-Registrierung mit Angabe persönlicher Informationen wie Name, Anschrift, E-Mail-Adresse etc. erreicht. Sie ist notwendig für Geschäftsbeziehungen wie Online-Bestellungen, Online-Banking etc., bei der die Nutzer also einen Kundenstatus haben. Der Vorteil liegt in der eindeutigen Identifizierung und der weitestgehenden Nutzbarkeit personenbezogener Angaben (Alter, Geschlecht, Beruf, Wohnort, Kaufverhalten etc.) bei der Web-Zugriffsanalyse. Auf der anderen Seite sind viele Nutzer

zur Preisgabe der Anonymität nur bereit, wenn dies zur Abwicklung von Online-Diensten unabdingbar ist bzw. deutliche Vorteile daraus resultieren. Somit wird auf Grundlage einer Nutzerregistrierung nur ein geringer Abdeckungsgrad bezogen auf alle Web-Zugriffe erreicht werden können.

Auch persistente Cookies können zur Personifizierung genutzt werden, wenn der Nutzer-ID personenbezogene Daten zugeordnet werden, z.B. nach einer erfolgten Bestellung oder einer personenbezogenen Registrierung. Damit kann auch ohne explizite Anmeldung ein wiederkehrender (registrierter) Kunde erkannt und eine personenbezogene Erfassung des Nutzungsverhaltens erreicht werden. Solche Verwendungsformen personenbezogener Daten erfordern in Deutschland aus Datenschutzgründen die explizite Zustimmung des Nutzers.

Pfadvervollständigung

Wie erwähnt spart die Pufferung von Webseiten durch Browser oder Proxy-Rechner Zugriffe beim Web-Server ein, so dass es zu »Lücken« in den Log-Aufzeichnungen kommt. Um eine Verfälschung der Analyseergebnisse zu begrenzen, kann versucht werden, solche Lücken bei der Log-Transformation zu erkennen und im Rahmen der so genannten Pfadvervollständigung zu korrigieren [CoMS99]. Die Erkennung der Lücken basiert einerseits auf der Strukturinformation einer Website und zum anderen auf der durch die Referrer-Information dokumentierten Sequenz der protokollierten Zugriffe. So muss im Beispiel von Abb. 11-6 der im Log erkannte Pfad A-B-C-D eine Lücke zwischen C und D enthalten, da gemäß der Site-Struktur keine direkte Verbindung zwischen diesen Seiten besteht.

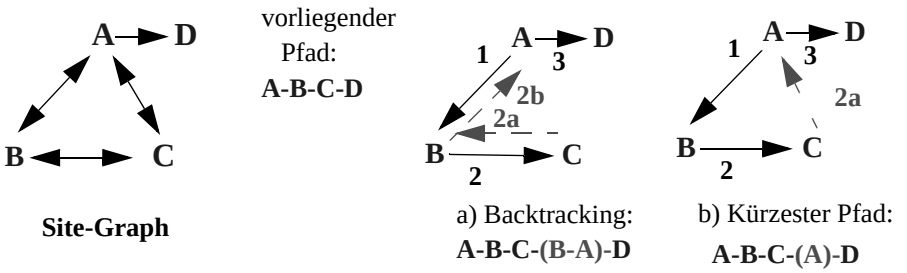

Abb. 11-6: *Pfadvervollständigung (Beispiel)*

Die Schließung erkannter Lücken kann nur in Annäherung durch Einsatz von Heuristiken verfolgt werden. Ein Ansatz ist ein so genanntes *Backtracking*, bei dem unterstellt wird, dass die Lücken primär durch die Rückwärtsnavigation über die Back-Funktion des Browsers enstanden sind. Dies ist insbesondere dann wahrscheinlich, wenn der Log-Eintrag zu einer unverbundenen Zielseite (D im Beispiel) in der Referrer-Information eine Seite angibt, die bereits in der jüngsten Vergangenheit referenziert wurde (z.B. A). Zur Korrektur werden in diesem Fall Zugriffe zur Rückwärtsnavigation auf die entsprechende Referrer-Seite eingefügt, im Beispiel zwei Zugriffe auf B und A (Reaktion a in Abb. 11-6). Alternativ kann

unter allen Pfaden, die gemäß Site-Struktur möglich sind, eine Auswahl unter bestimmten Optimalitätskriterien erfolgen, z.B. kürzester oder häufigster Pfad. Im Beispiel könnte so mit nur einem zusätzlichen Zugriff die Lücke geschlossen werden (Reaktion b in Abb. 11-6).

11.4 Data-Warehouse-Schema

Kennzeichnend für auf relationalen Datenbanken basierende Data Warehouses ist eine spezielle Strukturierung der Daten zur Unterstützung mehrdimensionaler Auswertungen. Große Bedeutung hat hierbei der Einsatz so genannter Stern-Schemata gefunden, deren Einsatz sich auch zur Web-Zugriffsanalyse eignet [KiMe00]. Dabei ist die Masse der auszuwertenden Daten im Rahmen von Faktentabellen gespeichert. Die zur Auswertung benötigten beschreibenden Eigenschaften der Fakten sind durch Dimensionstabellen repräsentiert; die Verbindung zwischen den Tabellen erfolgt durch Fremdschlüssel der Faktentabelle, die sich auf jeweils eine Dimension beziehen.

Zur Illustration zeigt Abb. 11-7 die Grobstruktur eines möglichen Stern-Schemas zur Web-Zugriffsanalyse mit einer Faktentabelle und acht Dimensionstabellen, welche die meisten der besprochenen Daten integrieren. Als Granularität der Faktentabelle und damit der Auswertungen werden einzelne Seitenzugriffe (Page Views) verwendet, d.h., für jeden nach der Bereinigung verbliebenen Zugriff auf eine Webseite gibt es einen Satz in der Faktentabelle. Für die Auswertung relevante Merkmale aus den Log-Sätzen wie Tag und Zeit der Zugriffe, Referrer, Seite (URL) und Nutzerangaben (IP-Adresse, Browser, Betriebssystem, Cookie-Inhalte) sind direkt durch entsprechende Dimensionstabellen repräsentiert. Bei den Seiten kann auch eine Kategorisierung nach Funktion (Navigation vs. Inhalt, Produktinformation, Bestellung ...) abgebildet werden. Zur inhaltlichen Kategorisierung wird, wie v.a. für E-Commerce-Anwendungen bedeutsam, die Produkttabelle verwendet, welche einen Produktkatalog repräsentiert. Somit kann für jeden Web-Zugriff, z.B. abgeleitet aus dynamischen Parametern wie der Produkt-ID, eine gezielte Zuordnung zu einzelnen Produkten erfolgen. Die Dimension Session-Typ erlaubt eine Kategorisierung von Sitzungen, z.B. nach Dauer, Länge oder zur Unterscheidung verschiedener Arten von Informations- oder Kaufbesuchen. Die Dimensionstabellen sind i.A. hierarchisch organisiert, so dass Auswertungen auf unterschiedlichen Granularitätsstufen möglich sind (z.B. zeitliche Auswertungen auf Tages-, Monats-, Quartals- oder Jahresebene; Referrer-Analyse auf Seiten- oder Site-Ebene, Produktauswertungen für Produktgruppen oder einzelne Produkte etc.).

Die Nutzertabelle soll nur Angaben zu anonymen Besuchern erfassen und kann auch aus den Sitzungen ableitbare Nutzerprofile enthalten (Häufigkeit und Intensität der Nutzung, Interessensschwerpunkte etc.). Personifizierte Nutzer wie Kunden eines Unternehmens sollen in einer eigenen Kundentabelle verwaltet werden, um besondere Auswertungen für diese wichtige Personengruppe zu unter-

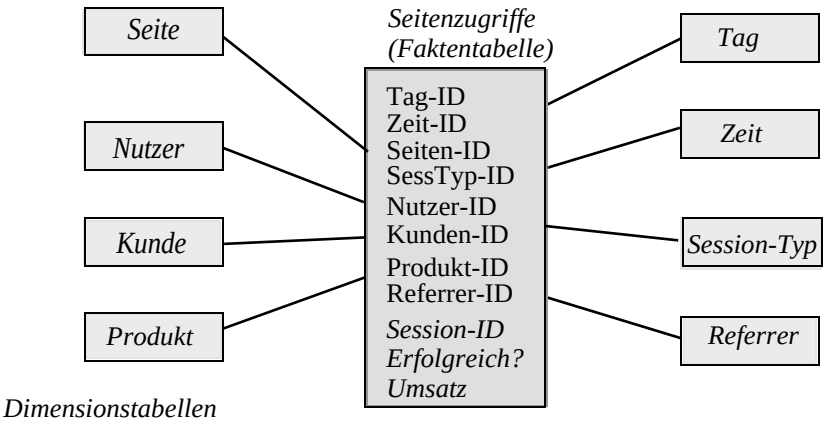

Abb. 11-7: *Stern-Schema zur Web-Zugriffsanalyse (Beispiel)*

stützen. Für die Auswertung interessieren dabei u.a. die Zuordnung der Kunden zu Kategorien bezüglich Wohnort/Region, Geschlecht, Altersgruppe, Familienstand, Berufsgruppe, Umsatz etc.

Die Faktentabelle enthält neben den Verweisen (Fremdschlüssel) auf die Dimensionseinträge noch so genannte Kennzahlen (kursiv dargestellt). In dem Beispiel werden sie genutzt, um die Zugehörigkeit von Seitenzugriffen zu einer bestimmten Sitzung zu dokumentieren und um anzuzeigen, ob der Zugriff durch den Web-Server erfolgreich durchgeführt wurde. Für E-Commerce-Anwendungen kann ferner der z.B. durch eine Bestellung entstehende Umsatz abgelegt werden, dessen Analyse natürlich von besonderer Bedeutung ist.

Es versteht sich von selbst, dass je nach Zielsetzung einer Website und ihrer Analyse weitere bzw. andere Merkmale Berücksichtigung finden können. Soll z.B. die Effektivität bestimmter Werbemaßnahmen oder Design-Änderungen analysiert werden, müssten solche Maßnahmen auch innerhalb von entsprechenden Dimensionstabellen repräsentiert werden [KiMe00]. Auch können zusätzliche Faktentabellen verwendet werden, etwa für eine genauere Sitzungsanalyse oder zur Vorberechnung häufig benötigter Auswertungen (z.B. monatliche oder kundenbezogene Zusammenfassungen).

11.5 Analyse

Mit den im Web Data Warehouse integrierten Daten wird eine Vielzahl von Analysen unterstützt, insbesondere zur Beantwortung von technischen, inhaltlichen und nutzerbezogenen Fragestellungen, wie sie bereits in der Einleitung angesprochen wurden. Als Auswertungsverfahren können die umfangreichen von Data Warehouses unterstützten Ansätze genutzt werden, wie die Erstellung vordefinierter Berichte und Statistiken, die interaktive Durchführung von Anfragen und mehrdimensionaler OLAP-Auswertungen sowie komplexere Data-Mining-Ana-

lysen. Insbesondere bietet die Warehouse-Lösung eine hohe Auswertungsflexibilität, da unterschiedliche Auswahlkriterien auf verschiedenen Detaillierungsstufen nahezu beliebig kombinierbar sind. So kann die Referrer-Analyse, Nutzeranalyse oder inhaltliche Analyse unter Berücksichtigung der anderen Dimensionen erfolgen. Beispiele für solch kombinierte Anfragen sind: Von welchen Referrern kommen die Interessenten bestimmter Produkte, wie ist die zeitliche Entwicklung bei der Nutzungsintensität von Kunden vs. anonymen Nutzern, welche Produkte werden von jungen weiblichen Kunden besonders nachgefragt etc.

Zur Illustration zeigt Abb. 11-8 die grafische OLAP-Ausgabe zur inhaltlichen Auswertung von Zugriffen auf die Website eines Lehrstuhls. Die Ausgabe bezieht sich auf die in [StRQ00] vorgestellte Warehouse-Realisierung, bei der eine mehrstufige inhaltliche Kategorisierung der Webseiten zunächst nach Lehre, Forschung etc. erfolgt, bei der Lehre erfolgt eine weitere Unterteilung nach Vorlesungen, Übungen, Praktika etc. Die Beispielausgabe gibt die Zugriffshäufigkeit auf die Materialien zu einzelnen Vorlesungen verschiedenener Quartale an. Wie für OLAP-Auswertungen üblich können zur Verfeinerung/Vergröberung interaktiv Drill-down- bzw. Roll-up-Auswertungen entlang der Inhaltsdimension erfolgen oder neue Parameter für andere Auswertungsdimensionen (Zeitraum, Nutzer) gewählt werden.

Generell sind als Kennzahlen bzw. Bewertungsmetriken statistische Aggregierungen zu absoluten und relativen Zugriffshäufigkeiten, Verweilzeiten und Übertragungsvolumen hinsichtlich der unterschiedlichen Dimensionen relevant. Zur Fehlererkennung interessieren die Seiten und Wege, die zu einem nicht erfolgreichen Zugriff oder einem Sitzungsabbruch geführt haben. Wichtig sind ferner unterschiedliche *Konversionsraten*, z.B. welche Anteile der Nutzer bestimmte Inhalte aufgesucht bzw. bestimmte Aktivitäten durchgeführt haben. So interessieren

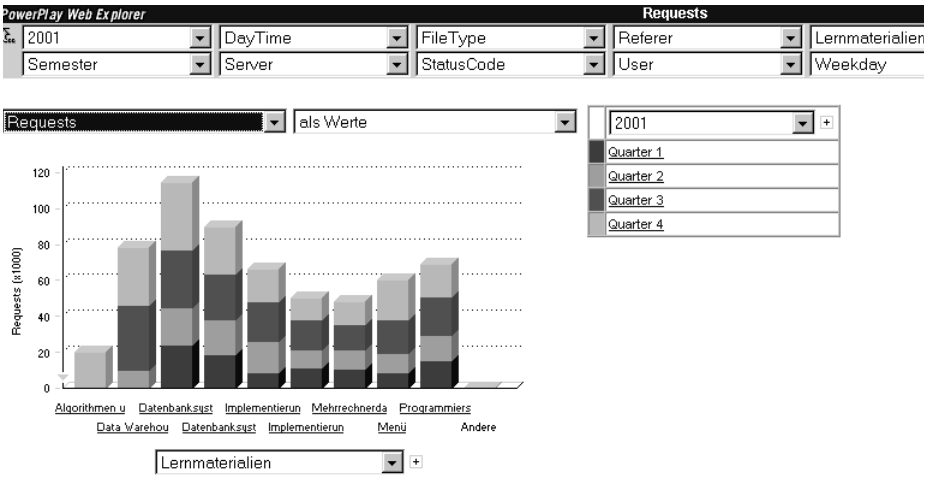

Abb. 11-8: *OLAP-Ausgabe zur inhaltlichen Zugriffsanalyse (Zugriffshäufigkeit auf Vorlesungsunterlagen einer Lehrstuhl-Website)*

im E-Commerce-Umfeld vor allem die Konversionsraten von Besuchern zu Kunden (Käufern), die Rückkehrquote von Kunden (auch ein Indikator für die Zufriedenheit) sowie der pro Kunde generierte Umsatz und Gewinn. Die Detailauswertung dieser Metriken kann dann genutzt werden, um profitable (bzw. unprofitable) Referrer, Produkte, Kunden etc. zu bestimmen. Zur Erklärung und Optimierung der Konversionsraten ist es 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. Die Referrer-Analyse kann helfen, die Effektivität von Marketing-Aktivitäten wie Werbung auf externen Websites zu bewerten. Idealerweise kann sogar der Return on Investment bestimmt werden, in dem die Kosten einer Werbemaßnahme ins Verhältnis zu den neu generierten Umsätzen und Gewinnen gesetzt wird.

Data-Mining-Auswertungen sollen zusätzliche Erkenntnisse ermöglichen, indem nicht nur festgelegte Auswertungen berechnet, sondern relevante Muster im Nutzungsverhalten selbstständig entdeckt werden. Von besonderem Interesse sind hierzu vor allem Ansätze zur Segmentierung von Kunden und Nutzern hinsichtlich Kriterien wie z.B. Interessensgebieten und Kaufverhalten. Die Identifikation von Inhalten/Web-Bereichen mit ähnlichen Nutzungsmerkmalen kann zur Optimierung des Web-Auftritts genutzt werden. Für solche Mining-Aufgaben kommen unterschiedliche Techniken der Statistik und der künstlichen Intelligenz in Frage, insbesondere Cluster-Verfahren, Klassifikationsansätze sowie Assoziations- und Sequenzregeln [EsSa00]. So erlauben Cluster-Verfahren die Bestimmung von Nutzergruppen mit ähnlichen Interessen oder von ähnlichen Sitzungen, sofern geeignete Ähnlichkeitsmaße hierfür gefunden werden. Klassifikationsansätze ermöglichen z.B. die Vorhersage unbekannter Kundenmerkmale bzw. des künftigen Kundenverhaltens, indem das Wissen zu bekannten Kunden verallgemeinert wird. Assoziationsregeln unterstützen eine Warenkorbanalyse bzw. allgemein die Bestimmung von häufig zusammen aufgerufenen Inhalten (zusammen gekauften Produkten). Sequenzregeln analysieren, welche Inhalte oder Aktionen oft in aufeinander folgenden Sitzungen eines Benutzers aufgerufen werden.

11.6 Nutzung

Die aus den Analysen gewonnenen Erkenntnisse können vor allem zur Umstrukturierung einer Website verwendet werden, um erkannte Defizite zu beheben bzw. erkannte Optimierungspotenziale umzusetzen. Bereits die Auswertung anonymer Web-Zugriffe, also ohne Personifizierung, ermöglicht viele Verbesserungen, etwa Beseitigung fehlerhafter Verweise, verbesserte Nutzerführung durch angepasste Such- und Navigationsmöglichkeiten, Ergänzung und Aktualisierung der Inhalte und Dienste etc. Eine wichtige Rolle zur Nutzerführung nehmen gezielte Empfehlungen (Links auf andere Webseiten) und Produktangebote sowie deren Platzierung ein. Während eine statische Platzierung für alle Nutzer einfach realisierbar

ist, wird durch eine dynamische Generierung in Abhängigkeit vorhergehender Zugriffe oder erkannter Nutzermerkmale (z.B. für wiederkehrende Nutzer) eine weit größere Flexibilität erzielt. Bei Personifizierung der Nutzer und erstellten Nutzungsprofilen kann dies am weitestgehenden durch personenbezogene (individuelle) Empfehlungen und Angebote umgesetzt werden, wie z.B. im Rahmen von One-to-One-Marketing angestrebt. Generell sollen im E-Commerce-Umfeld durch die Empfehlungen/Angebote vor allem Anreize für erhöhten Umsatz und Gewinn (Cross-Selling und Up-Selling) gesetzt werden.

Diese Maßnahmen beziehen sich auf die Nutzer, welche die Website bereits erreicht haben. Darüber hinaus liegt eine wichtige Nutzungsart der Web-Zugriffsanalyse in verbesserten Marketing-Maßnahmen, um überhaupt mehr Nutzer auf die Website zu bringen und um neue Kunden zu gewinnen. Hierzu kann insbesondere die Referrer-Analyse wichtige Hinweise geben, auf welchen externen Websites Werbehinweise die besten Erfolgsaussichten versprechen. Ferner lassen sich aus personalisierten Nutzungsprofilen nach bestimmten Kriterien Adressaten für Marketing-Kampagnen (z.B. E-Mail- oder Briefwerbung) bestimmen.

Die Verwendung personenbezogener Daten zur Personalisierung einer Website oder für Marketing-Zwecke sollten nur bei expliziter Zustimmung der betroffenen Personen erfolgen. In Deutschland ist diese Forderung auch in den Datenschutzgesetzen festgeschrieben, wenngleich dies schwer zu überwachen ist. Allerdings sollte es auch im Interesse der Unternehmen liegen, ihren Kunden zu zeigen, dass sie keinen Datenmissbrauch zu befürchten haben. Hierzu gehört die Bekanntgabe von detaillierten Datenschutzrichtlinien bezüglich der Web-Nutzung (Cookie-Einsatz, was wird ausgewertet und wie genutzt etc.), das Einholen der Nutzereinwilligung bezüglich der Verwendung personenbezogener Daten für die Website-Personalisierung oder für Werbemaßnahmen sowie das jederzeitige Widerrufsrecht bezüglich der Zustimmung zu diesen Verwendungsformen.

11.7 Werkzeug-Markt

Der Markt für Werkzeuge zur Web-Zugriffsanalyse ist umfangreich und unübersichtlich. Eine Vielzahl von Anbietern aus den Bereichen Daten(bank)-Management, Business Intelligence oder Data Mining haben ihre Produktpalette um spezifische Web-Zugriffsanalyse-Werkzeuge, -Lösungen oder -Funktionalitäten ergänzt. Daneben existieren zahlreiche Spezialanbieter sowie Prototypen aus Forschungsgruppen, die z.B. spezielle Data-Mining-Algorithmen zur Pfadanalyse realisieren. Eine Auswahl der Anbieter zeigt Abb. 11-9.

Die Bandbreite der Funktionalitäten von Web-Zugriffsanalyse-Werkzeugen ist ebenfalls weit gefächert. Die einfachsten Werkzeuge liefern lediglich tabellarische Statistiken, die sie aus der Log-Datei eines Web-Servers erstellen, z.B. die Anzahl der Zugriffe/Besucher über die Zeit, die am häufigsten referenzierten Seiten/Pfade etc. Dadurch sind einfache Aussagen über die Anzahl durchgeführter »Clicks«, den Zeitverlauf der Web-Server-Last, oder »beliebte« Seiten zu treffen.

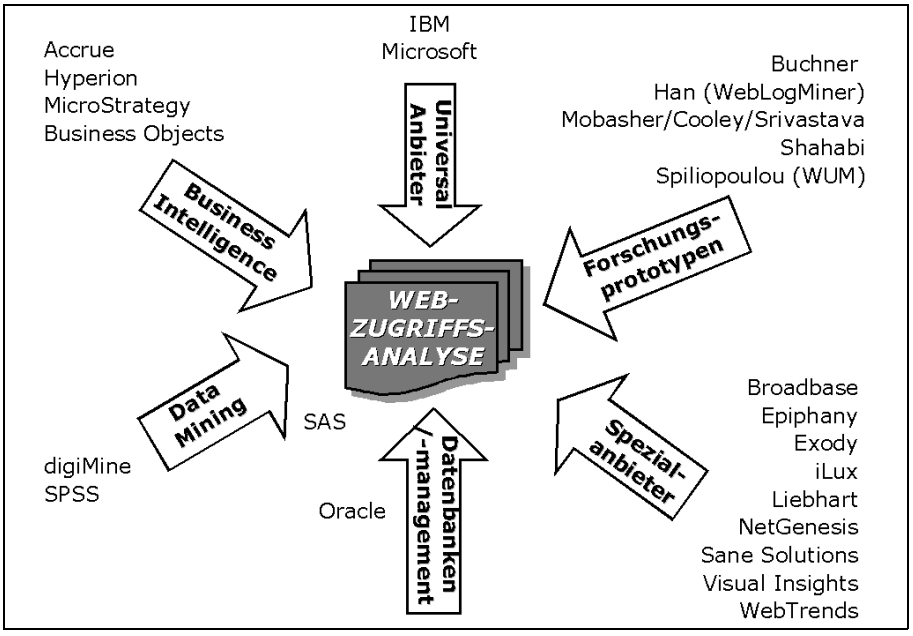

Abb. 11-9: *Anbieter von Web-Zugriffsanalyse-Produkten (Auswahl)*

Einige komplexe Werkzeuge dagegen ermöglichen eine Integration von Zugriffs- und Website-Struktur-Daten aus den Log-Dateien mehrerer Web- und Applikations-Server und Daten aus dem Unternehmens-Data-Warehouse. Zusätzlich kann eine Bewertung der Zugriffsverhalten, teilweise teil-automatisiert, über geschäftsorientierte Metriken erfolgen. Manche Produkte bieten außerdem eine Umsetzung der Ergebnisse in personenbezogene Marketing-Aktionen oder (automatische) Website-Anpassungen, inklusive einer Erfolgskontrolle der durchgeführten Maßnahmen und Verwaltung der auf einer Website angebotenen Produkte und deren Käufer.

Zur Diskussion unterteilen wir die Menge der Web-Zugriffsanalyse-Werkzeuge in drei Klassen:

- Werkzeuge für einfache Web-Zugriffsstatistiken (1)
- Data-Warehouse-basierte Werkzeuge mit OLAP-Funktionalität (2)
- Data-Mining-Werkzeuge (3)

Bei dieser Unterteilung zählt nicht nur die Art der Analysefunktionalität; vielmehr spielen Qualität, Umfang und Sicherheit der zur Verfügung stehenden Daten sowie die Fähigkeiten zur Umsetzung der erzielten Ergebnisse und deren Kontrolle eine wesentliche Rolle (vgl. Abschnitte 11.2, 11.3). Abb. 11-10 zeigt wichtige Unterkriterien wie *Art der Datenquellen*, *Datenhaltung*, *Qualität der Datentransformation*, *Art der Datenanalyse* und *Umsetzung/Kontrolle der Ergebnisse* sowie deren Ausprägungen. Als Beispiele zur Abgrenzung von Werk-

Kriterium	Ausprägungen
Art der Datenquellen	Website-Daten: Web-Server-, Browser-, Proxy-Logs, selbstdefinierte Logs (Web-Server-Plug-ins), Website-Struktur, Kundentransaktionen im Web
	Application Server Logs
	Unternehmensdaten: Kunden-/Produktinformationen, Unternehmens-Warehouse/-Data Marts
Datenhaltung	Datei(en), Datenbank, Data Warehouse
Qualität der Datentransformation	Qualität der Identifikation von: Page Views, Benutzer, Sessions, Suchbegriffe
	Filterungen: Roboterzugriffe, Dateiformate, ...
	Inhaltliche Kategorisierung von URLs
Art der Datenanalyse	tabellarische Zugriffsstatistiken, OLAP-Analyse, Data-Mining-Techniken
	(Geschäfts-)Metriken (Kaufrate, Mikrokonversionsraten, ROI)
Umsetzung/Kontrolle der Ergebnisse	Website-Anpassung
	Personalisierung
	Kampagnen-Management
	erreichter Automatisierungsgrad

Abb. 11-10: *Unterscheidungskriterien für Web-Zugriffsanalyse-Werkzeuge*

zeugen seien die emminenten Leistungs- und Verfügbarkeitsunterschiede zwischen der Datei- bzw. Datenbankhaltung der zu analysierenden Daten genannt, Aspekte der Datentransformation, bei der die schwierige, aber auch Ergebnis-kritische Bestimmung der »Besucher« und »Sessions« aus den Web-Log-Daten vorgenommen wird, oder Möglichkeiten zur Personalisierung der Auswertung.

Zur Auswahl und Planung des Einsatzes von Werkzeugen für ein Unternehmen ergeben sich zusätzliche Kriterien. Insbesondere die »*Passgenauigkeit*« des Werkzeuges in die Unternehmenslandschaft (Interoperabilität, Software-/Hardware-Gegebenheiten, Anbieter-«Politik« des Unternehmens, evtl. Outsourcing an einen ASP[1] vs. In-House-Nutzung) oder die *Marktsituation* sowohl des Produkts als auch des Anbieters sind hier zu nennen. Weiterhin sind die *Performanz* (insbesondere der Datenaufbereitung und der Anfragebearbeitung auf grossen Datenmengen), *Ergonomie* und nicht zuletzt das recht unterschiedliche *Pricing* zu nennen.

Im Folgenden konzentrieren wir uns bzgl. der Definition der Werkzeugklassen und der Abgrenzung der Werkzeuge auf die Kriterien gemäß Abb. 11-10. Wir skizzieren die wichtigsten Eigenschaften der betrachteten drei Werkzeugklassen, benennen jeweils eine Auswahl aus der umfangreichen Menge von Web-Zu-

1. ASP = Application Service Provider

griffsanalyse-Werkzeugen sowie spezielle Eigenschaften einiger Vertreter. Die Werkzeugauswahl und die abgeleiteten Aussagen orientieren sich an den verfügbaren Materialien auf den Internet-Seiten der Anbieter bzw. eigenen Erfahrungen mit den (wenigen) frei zugänglichen Demo-Versionen (Stand 4/2002). Es wurden keine Evaluierungen oder Testinstallationen unter Mitwirkung der Anbieter durchgeführt. Detaillierte Informationen finden sich unter der in Abb. 11-11 bereitgestellten Sammlung von Internet-Links zu den genannten Anbietern und Produkten.

11.7.1 Werkzeuge für einfache Web-Zugriffsstatistiken

In dieser Werkzeugklasse erfolgt die Datenhaltung üblicherweise in Dateien oder proprietären Datenbanken. Die Auswertung beschränkt sich auf (Offline-)Web-Log-Dateien, die Darstellung der Analyseergebnisse erfolgt tabellarisch oder mit rudimentären Grafiken auf der Basis fester Dimensionen (Zeit, Besuch, Nutzer). Die Datentransformation ist einfach (Ausschluss bestimmter Sites oder Dateiendungen, einfache Identifikation von Benutzern, keine inhaltliche Kategorisierung von URLs). Eine Umsetzung oder Kontrolle von Ergebnissen muss manuell erfolgen. Einfache »geschäftsorientierte« Metriken sind teilweise vorhanden, wie z.B. die Ermittlung einer Click-through-Rate als Maß für den Erfolg eines Werbebanners. Personalisierung erfolgt nicht. Vertreter dieser Kategorie sind (u.a):

- Exody WebSuxess
- IBM SurfAid for Reporting
- Liebhart systems WebFeedback
- MrUnix Webalizer
- Sane Solutions NetTracker
- WebTrends Reporting Server

Das einfachste Werkzeug dieser Gruppe stellt der *Webalizer* dar. Es produziert einen überwiegend tabellarischen Report mit allgemeinen Leistungs- und Zugriffsstatistiken (z.B. Zugriffe bzgl. Seiten, Übertragungsmenge, Besuche, fehlerhafte Seiten, zugreifende Rechnernamen etc.), deren zeitliche Entwicklung sowie Rankings (Top-Referrer/-Seiten/-Pfade/-Rechner/-Browser etc.). Sämtliche Auswertungsdimensionen sind statisch, Einstellungen werden über eine Datei vorgenommen. Quelle ist die Web-Log-Datei eines Servers.

NetTracker, *SurfAid for Reporting*, *Reporting Server* und *WebSuxess* liefern eine vorgegebene Menge von Analyseberichten, welche Zugriffsstatistiken bzgl. der Zeit, Besuche oder Pfaden umfassen. *SurfAid for Reporting*, *Reporting Server* und *WebSuxess* bieten neben Tabellenausgaben auch eine grafische Darstellung der Ergebnisse (z.B. den zeitlichen Verlauf). Eine einfache Wahl eines Abschnitts der Zeitdimension (z.B. ein bestimmte(r) Tag/Woche/Monat) ist i.d.R. möglich. Auch werden Rankings bzgl. einfacher Konversionsraten angeboten, wie man sie z.B. aus dem Verhältnis der Zugriffe über ein Werbebanner zu denen auf einer

Produktseite ablesen kann. Alle genannten Werkzeuge bedienen sich einer Datenhaltung in einer proprietären Datenbank.

IBM bietet Web-Zugriffsanalyse-Reports generell als Application Service an. *IBM* erstellt ein kundenspezifisches Warehouse und liefert Reports und/oder einen Zugang zum Data Warehouse. Die einfachste Variante und damit einen Vertreter der einfachen Kategorie stellt *SurfAid for Reporting* dar.

WebFeedback liefert eine URL-Zugriffsstatistik auf Basis der Struktur einer bis zu einer bestimmten Verzweigungstiefe zu analysierenden Website. Diese umfasst z.B. die grafische Darstellung der Verknüpfungen der URLs, die von einer Seite abgehen bzw. eine Seite erreichen, und deren Besuchshäufigkeit. Reports können in XML im- und exportiert werden.

Die Preise dieser Werkzeuge erreichen drei- bis vierstellige Euro-Beträge. Viele Anbieter stellen Demo-Versionen zur Verfügung (s.u.). Einige Werkzeuge sind sogar kostenfrei (z.B. *Webalizer*).

11.7.2 Data-Warehouse-basierte Werkzeuge mit OLAP-Analyse

Die Werkzeuge dieser Kategorie basieren auf einem Data Warehouse und meist mehreren daraus abgeleiteten Data Marts über Web-Zugriffe. Die Werkzeuge nutzen die Infrastruktur der Data-Warehouse-Lösung desselben Anbieters zur Datenaufbereitung, -haltung und -analyse; sie sind jedoch selbst zusätzlich zu erwerben.

Bezüglich der in Abb. 11-10 genannten Kriterien sind als vorherrschende Eigenschaften dieser Produktklasse zu nennen:

- Als Datenquellen fungieren neben den Log-Dateien eines oder mehrerer Web-Server häufig zusätzliche Web-Zugriffsdaten. So werden durch Web-Server-Plug-ins, »Packet Sniffer«, Browser-Plug-ins, Application Server Logs, Analyse der Website-Struktur etc. umfangreichere bzw. zeitnähere Daten gewonnen. Die Werkzeuge arbeiten auf einem Data Warehouse, welches neben den Web-Zugriffsdaten häufig auch andere Datenquellen (z.B. Kundendaten, -transaktionen, Data Marts) integrieren kann.
- Die Datentransformation erfolgt über eine grafische Oberfläche und ist üblicherweise ausgefeilter als die der einfachen Werkzeuge. So können z.B. zur Reduktion der Web-Log-Daten verfügbare Listen von Roboter-Sites integriert werden oder Suchbegriffe aus dynamischen Anfragen extrahiert werden. Eine hierarchische, inhaltliche Kategorisierung von Webseiten wird nicht unterstützt; allerdings ist es häufig möglich, Metadaten zur Beschreibung von URL(-Gruppen) zu definieren.
- Die Datenhaltung erfolgt mit gängigen (häufig allerdings ausschließlich proprietären) relationalen und/oder multidimensionalen DBS.
- Die Auswertung wird mittels Eigen- oder Fremd-OLAP-Werkzeugen mit entsprechender Roll-up/Drill-down-Funktionalität durchgeführt. Die Auswer-

tungsdimensionen sind i.d.R. vorgegeben. Ergänzt wird die Analyse durch die für einfache Web-Zugriffsanalyse-Werkzeuge typischen Statistiken und Diagramme. In der Regel können Benutzergruppen definiert werden, bzgl. derer Auswertungen durchgeführt werden.

- Die meisten Werkzeuge bieten inhaltsorientierte Metriken (z.B. Konversionsraten) an; einige unterstützen auch (personalisiertes) Kampagnen-Management.

Vertreter dieser Kategorie sind u.a:

- Accrue Hitlist/Insight
- Hyperion Web Analysis SuiteIBM SurfAid for Analysis/Business (ASP)
- Microsoft Commerce Server
- NetGenesis NetGenesis*5*
- Oracle Clickstream Intelligence
- SAS WebHound
- TeaLeaf Technologies TeaCommerce Suite

NetGenesis5 bietet eine vergleichsweise umfangreiche Menge von – auch geschäftsorientierten – Metriken zur Bewertungen des Web-Auftritts. Es können eigene Metriken und Benutzergruppen definiert werden. *NetGenesis* bietet eine zeitlich begrenzte, kennwortgeschützte Nutzung seiner Demo-Installationen (z.B. Demo-Data Mart zum Online-Buchhandel / Online-Aktienhandel). *NetGenesis5* nutzt die OLAP-Engine von *MicroStrategy*, die allerdings auch ein eigenes Web-Zugriffsanalyse-Modul *Web Traffic Analysis Module* anbieten.

Microsoft liefert mit den *Commerce Server* eine Suite, die viele Aspekte der Durchführung von Online-Shops abdeckt. Zusätzlich zur Web-Zugriffsanalyse können Webseiten generiert werden, Kunden, Produkte, Umsätze einer Website verwaltet und gezielte B2C-Kampagnen durchgeführt sowie deren Effektivität ausgewertet werden. Die Lösung beschränkt sich aber auf die Microsoft-Produktwelt (Web-Server, DBS, Analyse).

Oracle Clickstream Intelligence ermöglicht die nutzerspezifische Erweiterung des Data-Warehouse-Schemas zur Anpassung an Analyseanforderungen. Oracle liefert der immanenten Datenbanktechnologie (Parallelität, materialisierte Sichten, Datenpartitionierung) entsprechende Tuning-Möglichkeiten.

Die *TeaLeaf*-Suite bietet mittels Plug-ins für Web-Clients, -Server und einige Application Server (z.B. von SAP) die wohl weitreichendste Akquise und Darstellung von Web-Zugriffsdaten. Sie ermöglicht z.B. die Ermittlung der tatsächlichen Verweildauer von Nutzern auf Webseiten, dezidierter Fehlerursachen oder Nutzereingaben durch Browser-Plug-ins. Mittels eines Visualisierungswerkzeugs können besuchte Seiten, durchgeführte Eingaben etc. in einer Art »Slide-Show« en detail dargestellt werden. Bei der Verwendung dieses Werkzeugs sollte der Einhaltung des Datenschutzes daher besondere Beachtung geschenkt werden. Der Anbieter weist auf seine diesbezüglichen Verkaufsrichtlinien und entsprechende Software-Unterstützung (opt-in/opt-out) hin.

Bei der Auswahl einer kommerziellen Web-Zugriffsanalyse-Lösung aus dieser Werkzeugklasse sollte insbesondere die Integration in die Unternehmensproduktwelt genau beachtet werden, da die Interoperabilität mit Fremdanbietern häufig nicht oder nur beschränkt gewährleistet ist. Die (Neu-)Anschaffungskosten können für diese Werkzeuggruppe wegen der notwendigen Infrastruktur leicht im fünf- bis sechsstelligen Euro-Bereich liegen.

11.7.3 Data-Mining-Werkzeuge

Eine weitere Werkzeugklasse im Rahmen der Web-Zugriffsanalyse stellen Data-Mining-Werkzeuge dar. Insbesondere im kommerziellen Umfeld stellt Web Usage Mining im Wesentlichen ein Anwendungsgebiet für die »klassischen« Data-Mining-Verfahren zur Klassifikation, Cluster-, Assoziations- und Sequenzanalyse dar. So müssen die zu analysierenden Daten aus den vorliegenden Data Warehouses oder Dateien für die meist datei- oder tabellenorientierten Data-Mining-Werkzeuge vom Benutzer, mit größerer oder geringerer Unterstützung durch das Werkzeug, abgeleitet und aufbereitet werden. Als bekannte kommerzielle Vertreter sind u.a. zu nennen:

- digimine (ASP)
- IBM Intelligent Miner for Data
- NCR/Teradata TeraMiner
- Oracle Data Mining Suite (Darwin)
- SAS Enterprise Miner
- SPSS Clementine Workbench

Diese Werkzeuge bieten in der Regel mehrere Klassifikations-, Cluster- und Prognose-Algorithmen, die auf Web-Zugriffsdaten z.B. zur Navigationspfad-Analyse angewendet werden können. Für die typischen Web-Zugriffsanalysen offerieren die Hersteller meist zusätzliche Produkte (z.B. *SAS Webhound*, *IBM SurfAid als ASP*). SPSS Clementine allerdings bietet für sein Data Mining-Werkzeug das sog. Clementine Application Template (CAT) for Web Mining an, also eine Prozessfolge zur Aufbereitung und Analyse von Website-Besuchen bzw. -sequenzen. Daneben existiert mit CAPRI noch ein Plugin zur Sequenzerkennung. Neben der kommerziellen Produkten gibt es eine Reihe von auf Web (Usage) Mining-Verfahren spezialisierte Forschungsprototypen [SCDT00], von denen z.B. WebSIFT [CoTS99], WebLogMiner [ZaXH98] oder WUM [Spil00] zu nennen sind.

Ein generelles Manko derzeitiger Data-Mining-Werkzeuge ist, dass sie häufig auf Dateien arbeiten, selten auf (einfachen) Datenbanken und gar nicht auf Data Warehouses. Zur Vermeidung redundanter, aus den Datenbanken abgeleiteter Dateien sowie aus Leistungsgründen ist somit eine engere Anbindung der Werkzeuge an die Data Warehouses dringend geboten, um das Potenzial der Mining-Analysen voll nutzen zu können.

11.7.4 Informationen im WWW

Viele Anbieter stellen Einstiegsinformationen über ihre Web-Zugriffsanalyse-Lösungen im Internet zur Verfügung, welche erfreulicherweise häufig über die üblichen Marketing-orientierten Data Sheets hinausgehen. Abb. 11-11 zeigt eine Auswahl der Adressen von 20 der wichtigsten der über 80 kommerziellen Produkte, die während der Recherchen ausgemacht werden konnten. Die Tabelle enthält auch Verweise zu Webseiten, auf denen freie Werkzeug-Demonstrationen zur Eigeninstallation oder Direktausführung zu finden sind. Einige Anbieter allerdings liefern (mit Wissensstand 4/02) leider nur sehr wenig Online-Information über ihre wahrscheinlich interessanten Produkte wie z.B. *Elytics Analysis Suite, E.piphany E-Commerce Reporting & Analysis, iLux Enterprise CampaignManager, MicroStrategy WebTraffic Analysis Module, Quadstone Customer Conversion*.

Anbieter	Tool	Internet-Adresse	Bem. / Demos
Accrue	Hitlist/Insight	http://www.accrue.com/ Company/Contact_Us/ reports_and_white_papers.html	
digiMine	Ebusiness Analytics (ASP)	http://www.digimine.com	http://www.digimine.com/solutions/samplereports-1.asp
Elytics	Analysis Suite	http://www.elytics.com	
E.piphany	E-commerce Reporting & Analysis	http://www.epiphany.com/ products/campmgr.html	
Hyperion	Website Analysis Suite	http://www.hyperion.com/ solutions/index.cfm	Link »Online Demos«
IBM	SurfAid (ASP)	http://www-3.ibm.com/ e-business/solution/28176.html	
IBM	SpeedTracer	http://www.alphaworks.ibm.com	http://www.downloadsafari.com/Files/utilshtmlaccs/S/SpeedTracer.html
iLux Enterprise	Campaign Manager	http://www.ilux.com	http://www.ilux.com/ products/sample_reports.html
Microsoft	Commerce Server	http://www.eu.microsoft.com/ germany/produkte/ overview.asp?siteid=10733	http://www.microsoft.com/ downloads (Microsoft download-Center)
MicroStrategy	Web Traffic Analysis Module	http://www.microstrategy.com/ Software/Applications/WTAM	
Mr Unix	Webalizer	http://www.mrunix.net/webalizer	
NCR Teradata	Teraminer	http://www.ncr.com/products/ software/teradata_mining.htm	
net.Genesis	NetGenesis5	http://www.netgen.com (download von White Papers)	

Abb. 11-11: *Auswahl von Web-Zugriffsanalyse-Tools im WWW (Stand 4/02)*

Anbieter	Tool	Internet-Adresse	Bem. / Demos
Oracle	Intelligence 1.0	http://otn.oracle.com/products/clickstream/htdocs/ocsi_faq.htm http://otn.oracle.com/products/clickstream/content.html	
Quadstone	Customer Conversion	http://www.quadstone.com	
Sane Solutions	NetTracker	http://www.sane.com/products/NetTracker/	http://www.sane.com/demo/de/NetTracker/web/index.html
SAS Institute	Web Hound, Enterprise Miner	http://www.sas.com/products/webhound/index.html http://www.sas.com/service/library/onlinedoc/webhound (Doku)	
SPSS	Clementine	http://www.spss.com/spssbi/clementine/CATs.htm http://www.spss.com/spssbi/capri/index.htm	
TeaLeaf Technology	TeaCommerce Suite	http://www.tealeaf.com/solutions/tcsuite.asp	
WebTrends	Reportings Server, Commerce Trends	http://www.webtrends.com/products/wrc/ent.htm	www.upenn.edu/computing/web/webteam/webtrends/cmplte.htm

Abb. 11-11 *(Fortsetzung): Auswahl von Web-Zugriffsanalyse-Tools im WWW (Stand 4/02)*

Neben den Produktinformationen der einzelnen Anbieter existieren zahlreiche Websites mit Informationen und Listen von Web-Zugriffsanalyse-Werkzeugen. Dort sind Verweise zu den Anbietern zu finden sowie – meist rudimentäre – Werkzeugbeschreibungen und einige wenige Werkzeugvergleiche. Beispiele sind (Stand 4/2002):

- ECRMGUIDE.COM:
  www.ecrmguide.com
- INTELLIGENT ENTERPRISE:
  www.intelligententerprise.com/000929/b_guide.shtml
  www.intelligententerprise.com/000929/feat5.shtml
- INTERNET PRODUCT WATCH: ipw.internet.com/analysis/recent1.html
- KDCENTRAL.COM:
  www.kdcentral.com/Software/Data_Mining/Web_Log_Mining
- KDNUGGETS: www.kdnuggets.com/software/web.html
- ZDNET: xlink.zdnet.com (Suchanfrage »web mining tool«)
- YAHOO:
  http://dir.yahoo.com/Computers_and_Internet/Software/Internet/World_Wide_Web/Servers/Log_Analysis_Tools

11.8 Zusammenfassung

Die Nutzungsmöglichkeiten der Web-Zugriffsanalyse sind sehr vielfältig und nicht nur für kommerzielle Websites von immer größerer Wichtigkeit, um eine gute Bedienung der Nutzer und die Erfüllung eigener Zielsetzungen für den Web-Auftritt zu erreichen. Derzeit herrschen noch einfache statistische Auswertungen auf Basis der Web-Log-Dateien vor, die jedoch viele Optimierungsmöglichkeiten ungenutzt lassen. Wir haben gezeigt, dass die Verwendung eines Data-Warehouse-basierten Ansatzes wesentliche Vorteile bringt, insbesondere hohe Skalierbarkeit, einfache Erweiterbarkeit und flexible Auswertungsmöglichkeiten. Von Bedeutung hierzu ist vor allem die Integration von Informationen zu Nutzern/Kunden und Inhalten/Produkten, um zu aussagekräftigen Ergebnissen und konkreten Verbesserungsmöglichkeiten zu gelangen.

Die Umsetzung eines Data-Warehouse-Ansatzes sowie die Nutzung von Data-Mining-Werkzeugen verursacht derzeit noch einen hohen Aufwand. Dieser sollte aufgrund der Weiterentwicklung entsprechender Werkzeuge, auch im Rahmen von kommerziellen Datenbanksystemen, zunehmend reduziert werden können. Auch aus Sicht der Forschung gibt es noch interessante, unzureichend untersuchte Fragestellungen, z.B. bei der Datenaufbereitung, der Ausgestaltung spezifischer Data-Mining-Verfahren sowie der automatisierten Umsetzung der Ergebnisse (dynamische Website-Modifikationen, Personalisierung).

Literatur

[BaGÜ01] Bauer, A., Günzel, H.: Data-Warehouse-Systeme. dpunkt.verlag 2001.

[CoMS99] Cooley, R., Mobasher, B., Srivastava, J.: *Data Preparation for Mining World Wide Web Browsing Patterns*. Knowledge and Informations Systems 1 (1), Springer-Verlag, 1999.

[CoTS99] Cooley, R., Tan, P.-N., Srivastava, J.: *WebSIFT: The Web Site Information Filter System*. Proc. Web Usage Analysis and User Profiling Workshop (WEBKDD'99), San Diego, CA, August 1999.

[EsSa00] Ester, M., Sander, J.: *Knowledge Discovery in Databases: Techniken und Anwendungen*. Springer-Verlag, 2000.

[HaKa00] Han, J., Kamber, M.: *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2000.

[KiMe00] Kimball, R., Merz, R.: *The Data Webhouse Toolkit: Building the WebEnabled Data Warehouse*. Wiley 2000.

[Schr00] Schroeck, M.J.: *E-Analytics – The Next Generation of Data Warehousing*. DM Review, Aug. 2000, http://www.dmreview.com.

[SCDT00] 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.

[Spil00] Spiliopoulou. M.: *Web Usage Mining for Web Site Evaluation*. Comm. ACM 43 (8), 127-134, 2000.

[ScNL01] Schaarschmidt, R., Nowitzky, J., Lufter, J.: *Clickstream Warehousing für e-CRM: Neue Herausforderungen an die Datenhaltung?* Proc. 5. Wirtschaftsinformatik (WI)-Tagung, pp. 117-131, Augsburg, Sep. 2001.

[StRQ00] 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

[WKDD01] WEBKDD-Workshops zu Web Mining: http://robotics.Stanford.EDU/~ronnyk/WEBKDD2000 bzw. WEBKDD2001.

[ZaXH98] Zaiane, O. R., Xin, M., Han, J.: *Discovering Web Access Patterns and Trends by Applying Olap and Data Mining technology on Web Logs*. Advances in Digital Libraries, pp. 19-29, Santa Barbara, CA, 1998.