

OLAP-Auswertung von Web-Zugriffen

Thomas Stöhr, Erhard Rahm, Stephan Quitzsch

Institut für Informatik, Universität Leipzig
Augustusplatz 10/11, 04109 Leipzig, Germany

1 Einleitung

Die Analyse des Zugriffsverhaltens auf Web-Seiten ist von großer Wichtigkeit, um die Attraktivität und Nützlichkeit von Internet-Angeboten zu bewerten und Rückschlüsse auf eine Verbesserung der Web-Präsentation zu erhalten. Im kommerziellen Umfeld (E-Commerce-Angebote wie Online-Buchhandel, -Reisebuchung, etc.) soll die sogenannte "Click Stream"-Analyse wertvolle Informationen über das Zugriffsverhalten von Kunden bzw. potentiellen Kunden liefern, mit dem Ziel den Wünschen der Kunden stärker entgegenzukommen, eine verbesserte Kundenbindung zu erreichen und letztlich den Umsatz zu steigern [8],[3]. Weiterhin ermöglicht die Analyse der Web-Zugriffe Erkenntnisse hinsichtlich von Performance-Anforderungen und daraus abgeleitet hinsichtlich der bereitzustellenden Hardware-Konfiguration, um ein gutes Durchsatz- und Antwortzeitverhalten zu erreichen [1].

Verbreitet sind einfache Analyse-Ansätze wie das Führen von Zählern für die Zugriffe auf Web-Seiten oder die direkte Auswertung einer Web-Log-Datei, welche sämtliche Zugriffe auf von einem Web-Server kontrollierten Seiten sequentiell speichert. Solche Verfahren weisen erhebliche Beschränkungen auf, die insbesondere im kommerziellen Umfeld große Nachteile implizieren:

- Auswertungen auf einer einfachen Datei sind inflexibel. Insbesondere ist keine Auswertung von Ad-Hoc-Anfragen zur Analyse möglich
- Der bei häufig frequentierten Web-Sites rasch wachsende Umfang der Web-Log-Datei ermöglicht meist nur eine Aufbewahrung der Detaildaten für einen relativ kurzen Zeitraum (z.B. 1-3 Monate). Ansonsten führen die Auswertungen auf der Datei meist zu inakzeptablen Antwortzeiten. Notwendig sind jedoch längerfristiger Auswertungen, um Aufschluß über signifikante Änderungen des Zugriffsverhaltens zu erhalten
- Die besuchten Web-Adressen werden üblicherweise in Form ihrer Directory-/Dateinamen abgespeichert. Die Herstellung eines inhaltlichen Bezuges und damit die semantische Aussagekraft und Auswertbarkeit solcher Dateizugriffspfade ist naturgemäß stark eingeschränkt. Erschwert wird die Analyse zusätzlich durch die üblicherweise sehr große Anzahl von zu verwaltenden Web-Seiten.
- In der Regel ist keine oder nur sehr umständliche Verknüpfung mit weiteren Datenquellen, wie z.B. demoskopischen Daten, möglich.

Zur Überwindung bzw. Reduktion dieser Problematik sind folgende wesentlichen Randbedingungen umzusetzen, wie sie teilweise in kommerziellen Systemen [9] zur Analyse des Kundenverhaltens im E-commerce-Bereich bereits verfolgt werden:

- Zum Erreichen einer angemessenen Skalierbarkeit und Verfügbarkeit sowie nicht zuletzt einer Durchführbarkeit von Ad-Hoc-Anfragen sollte eine Datenbank-gestützte Lösung eingesetzt werden
- Häufig werden flexible Read-only-Auswertungen über längere Zeiträume benötigt. Die Menge der verwalteten Clicks erreicht schnell einem Umfang im Giga- oder sogar Terabyte-Bereich [10]. Zur Datenhaltung müssen daher ein oder mehrere konsolidierte Data Warehouse(s) mit bewährter relationaler Technik eingesetzt werden

- Die Auswertung muß flexibel gehalten werden, d.h. anpaßbar an umgebungsspezifische Auswertungsdimensionen. Dafür kann das in der Geschäftswelt weitverbreitete multidimensionale Datenmodell eingesetzt werden. Die Analyse kann dann mit entsprechenden professionellen OLAP [5]-Werkzeugen erfolgen, die zahlreich und stabil auf dem Markt vertreten sind
- Zunehmend unverzichtbar wird eine einfache und plattformunabhängige Nutzung von Analyse-Werkzeugen. Daher ist ihre Einbindung ins Internet/Intranet nahezu unumgänglich
- Weiterhin sind ein möglichst geringer Aufwand, geringe Kosten und ein stabiler Betrieb anzustreben.

Auch für kostenfreie Informationsangebote wie im Hochschulbereich kann eine detaillierte Analyse des Zugriffsverhaltens wertvolle Hinweise über die Nützlichkeit und Verbesserung der bereitgestellten Web-Inhalte bieten. An der Universität Leipzig ist in einem initialen Szenario eine solche innovative, DB-gestützte und Web-basierte Anwendung zur multidimensionalen Analyse von Zugriffen auf die Web-Seiten des Instituts für Informatik entstanden. Mit Hilfe von OLAP-Werkzeugen werden die erhaltenen Daten ausgewertet und entsprechende (dynamische) Reports über Zugriffshäufigkeiten erstellt, wobei die zu analysierenden Web-Seiten inhaltsbezogen organisiert sind. Die Ergebnisse sind im WWW verfügbar und über einen herkömmlichen Web-Browser abrufbar. Als Software kommen kommerziell erhältliche DBS und Werkzeuge zum Einsatz. Unsere Studie zeigt damit auch den State-of-the-Art solcher Werkzeuge zur Realisierung innovativer Web-Anwendungen unter Integration von Datenbanktechnologie auf. Die beschriebene Vorgehensweise sollte in vielen vergleichbaren Projekten anwendbar sein.

Der Beitrag ist wie folgt aufgebaut: in Abschnitt 2 wird die unserer Web-Zugriffsanalyse zugrundeliegende Architektur und die Vorgehensweise bei der Generierung der Analysedaten beschrieben. Abschnitt 3 erläutert das multidimensionale Datenmodell, welches als Grundlage für die Auswertung der Zugriffe dient. Anschließend wird die Web-Browser-basierte Benutzerschnittstelle für OLAP präsentiert (Abschnitt 4). Abschnitt 5 zeigt Erweiterungsmöglichkeiten unseres Ansatzes auf. Abschließend folgt eine Zusammenfassung und ein Ausblick (Abschnitt 6).

2 Bereitstellung der Analysedaten

Gemäß der beschriebenen Leistungsanforderungen an eine professionelle Web-Analyse entspricht die zugrundeliegende Architektur (Abb. 1) unserer Beispielanwendung derjenigen moderner Data Warehouses/Data Mart¹-Umgebungen. In solchen Umgebungen müssen Daten von oftmals zahlreichen und heterogenen **operativen Systemen** (z.B. Sachbearbeiter- oder Außendienstdaten etc.) integriert werden. In existierenden Warehouses sind häufig noch handprogrammierte, fehleranfällige und wartungsunfreundliche Eigenlösungen zur Übernahme dieser Daten in das Warehouse anzutreffen. Aufgrund der Komplexität der Problematik werden allerdings immer häufiger spezielle Werkzeuge, sog. ETL²-Werkzeuge, propagiert. Diese extrahieren die operativen Daten, transformieren sie durch (meist graphisch entwickelte) komplexe Abbildungen, korrigieren, bereinigen und vereinheitlichen sie, soweit möglich, um sie dann weitgehend konsolidiert in ein relationales Datenhaltungssystem bzw. das **Data Warehouse** zu laden.

Basis für die multidimensionale Analyse stellen oftmals relationale Star-Schemata dar, welche die relevanten Auswertungsdimensionen und -metriken in entsprechenden Tabellen repräsentieren und aggregierte Daten vorhalten. Diese Schemata werden in einem zweiten ETL-Vorgang befüllt. Einige Analysewerkzeuge können direkt auf solchen relationalen Schemata arbeiten (ROLAP); eine andere Werkzeugklasse (MOLAP) arbeitet allerdings auf davon abgeleiteten, **multidimensionalen Strukturen** (sog. Würfel, Cubes), die im Dateisystem abgelegt werden.

In unserer konkreten Web-Log-Anwendung wird zur Zeit nur eine Datenquelle auf der operativen

1. Multidimensional organisierte Ausschnitte eines Data Warehouse werden oft als Data Marts bezeichnet.
 2. ETL = Extraction, Transformation, Load.

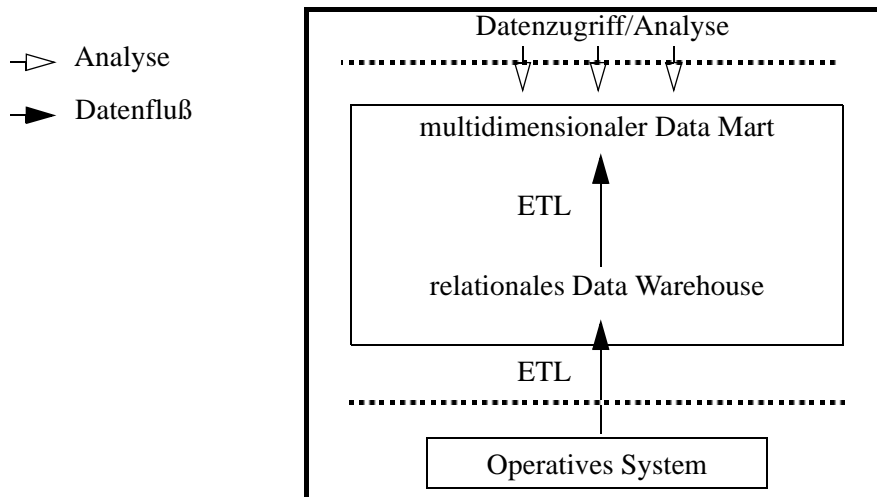


Abb. 1: Architektur der Web-Log-Anwendung

Ebene genutzt, nämlich die vom Web-Server geführte Log-Datei, welche sämtliche Zugriffe auf die verwalteten Web-Seiten sequentiell speichert. Daraus werden täglich mit einem einfachen Skriptprogramm diejenigen Einträge, welche die neuen Zugriffe seit dem letzten Update darstellen, extrahiert und in eine Datei gespeichert (in unserem Anwendungsbeispiel bis zu 50.000 neue Einträge pro Tag). Dateieinträge umfassen *Quelle* (Rechner), *Ziel* (Web-Seite), *Zeitpunkt* und *Umfang* (Bytes) des Zugriffs, welche sich im Analyseverfahren als Auswertungsdimensionen wiederfinden (s. Abschnitt 3). In unserem Fall befinden sich Web-Server und die generierte Web-Log-Datei auf einer UNIX-Maschine. Das Warehouse dagegen wurde auf einem Windows NT-Server (mit MICROSOFT SQL Server 7.0 [7]) eingerichtet, so daß die Web-Log-Daten auf die NT-Plattform transferiert werden müssen.

Im zweiten Schritt werden die aktuellen Zugriffsdaten durch einen ersten **ETL-Vorgang** mit geringfügigen Fehlerkorrekturen quasi 1:1 in eine relationale Tabelle (Teil des Data Warehouse) übertragen. Bei diesem Vorgang wird z.Z. das Werkzeug MICROSOFT DATA TRANSFORMATION SERVICES (DTS) eingesetzt, da es im Lieferumfang des SQL SERVER 7.0 enthalten ist. Trotz funktioneller Einschränkungen gegenüber spezialisierten ETL-Werkzeugen (wie z.B. ARDENT DATASTAGE, ETI EXTRACT, INFORMATICA POWERMART etc. [4],[2]) ist DTS im Kontext der Web-Log-Applikation z.Z. ausreichend. Dateien werden mit diesem Werkzeug über den sog. FLATFILE PROVIDER (nutzt die MICROSOFT OLE DB-Schnittstelle) ausgelesen, die Zieltabelle werden über ODBC befüllt. Metadaten über die strukturellen Beschreibungen des Quell- und Zielsystems sowie die Extraktions- und Transformationsvorgänge werden über die gleichen Schnittstellen an das Extraktionswerkzeug übermittelt und von diesem über die OPEN INTERFACE MODEL-Schnittstelle (OIM) im MICROSOFT-Repository zentral abgelegt.

Anschließend wird das **Sternschema**, welches die relevanten multidimensionalen Analysestrukturen bereits repräsentiert (s. Abschnitt „Datenmodell“) mit den extrahierten Einträgen über aktuelle Web-Zugriffe in einem weiteren ETL-Schritt aktualisiert. Während MICROSOFT SQL Server unmittelbar als Datenhaltungssystem für die relationalen Data Warehouse-Daten fungiert, werden zur multidimensionalen Datenhaltung die neben DTS ebenfalls im Lieferumfang enthaltenen OLAP SERVICES genutzt.

Zur Ableitung des **multidimensionalen Cubes** aus dem relationalen Schema war weiterhin ein (Visual-Basic-) Programm zu entwickeln, welches die benötigten Funktionen im API der OLAP SERVICES aufruft. Die Ausführungsplanung des Programms obliegt dem SQL SERVER AGENT und wird im MICROSOFT Repository gespeichert. Die Daten werden anschließend über die OLE DB-Schnittstelle des SQL Server 7.0 transportiert. Auf die strukturellen Metadaten des Starschemas wird bei der Konstruktion des Cubes über OLE DB zugegriffen. Der Cube selbst wird im proprietären Repository der OLAP SERVICES verwaltet und kann bei Bedarf in das MICROSOFT Repository exportiert werden.

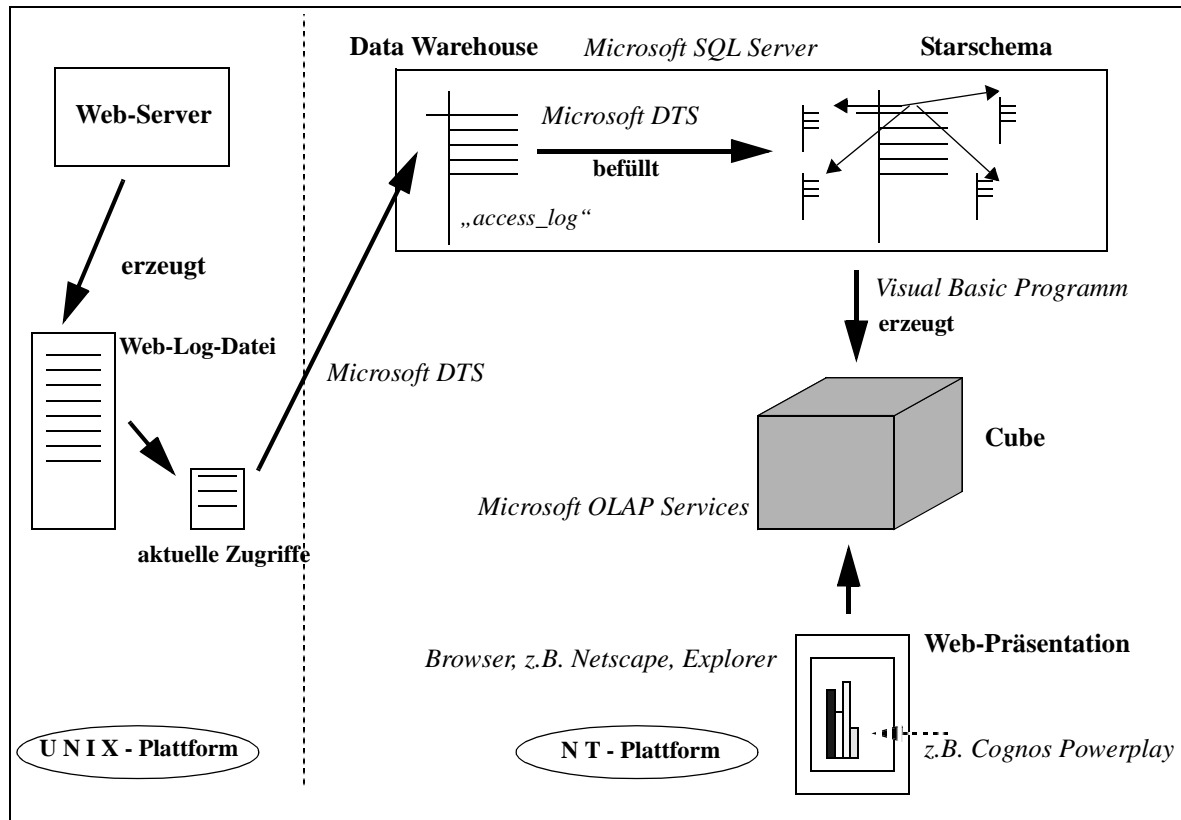


Abb. 2: Technische Architektur und Datenfluß

In unserer beschriebenen Konstellation sind zur Auswertung beliebige OLE DB-kompatible **OLAP-Werkzeuge** auf NT-Plattformen nutzbar (s. Abschnitt 4). Einen Gesamtüberblick über die technische Web-Log-Architektur sowie den Datenfluß gibt Abb. 2.

3 Datenmodell

Ein Ziel unserer Datenanalyse ist es, das einfache Zählen von Zugriffen auf Web-Seiten und ein ebenso einfaches Ranking von Seiten durch aussagekräftigere Bewertungen über die Attraktivität von *inhaltsorientierten Kategorien* besuchter Web-Seiten abzulösen. So soll statt dessen die Zugriffsfrequenz auf *Sparten* wie Forschung, Lehrangebot, Diskussionsforen, aktuellen Informationen etc., bewertet werden können. Dazu wurde die inhaltliche Struktur der Web-Seiten des Instituts multidimensional modelliert und ein Zuordnungsmechanismus Web-Seite \Leftrightarrow Kategorie, basierend auf den Adressen der Web-Seiten und relevanter Sparten von Web-Seiten, implementiert. Im folgenden wollen wir unser Modell insbesondere am Beispiel der Kategorisierung der Web-Seiten der Abteilung Datenbanken erläutern, deren Organisation eine entsprechende Detailtiefe erreicht. Diese Festlegungen sind natürlich in anderen Umgebungen an die jeweiligen Gegebenheiten anzupassen.

3.1 Ausgangssituation

Basis der Modellierung sind die sich aus der Struktur des Web-Log ergebenden Dimensionen und Metriken. Abb. 3 zeigt dazu einen Ausschnitt der Web-Log-Datei. Jede Dateizeile enthält als relevante Information über einen einzelnen Click folgende 5 Angaben:

- *Quelle* des Zugriffs als *IP-Adresse* oder *Name eines Rechners* (z.B. *rechnername.domäne.land* oder *rechnername.domäne.organisation*)
- sekundengenaue *Zeitpunkt* des Zugriffs in der Form *dd/mm/yyyy:hh:mm:ss* (inkl. Angabe der Zeitverschiebung)

```

abe09e64.ipt.aol.com - - [21/Jan/2000:18:37:20 +0100] "GET /bilder/ifi_logo2.gif HTTP/1.1" 200 1268
abe09e64.ipt.aol.com - - [21/Jan/2000:18:37:20 +0100] "GET /bilder/uni_logo2.gif HTTP/1.1" 200 4746
abe09e64.ipt.aol.com - - [21/Jan/2000:18:37:20 +0100] "GET /ifi/lehre/lernserver.html HTTP/1.1" 200 27614
ilabpc19.informatik.uni-leipzig.de - - [21/Jan/2000:18:37:23 +0100] "GET /~apel/LinAlg/serie1.ps HTTP/1.1" 200 82337
abe09e64.ipt.aol.com - - [21/Jan/2000:18:37:31 +0100] "GET /ifi/abteilungen/db/skripte/PPS/inhalt.html HTTP/1.1" 200 3095
tipc015.informatik.uni-leipzig.de - - [21/Jan/2000:18:37:37 +0100] "GET / HTTP/1.0" 304 0
tipc015.informatik.uni-leipzig.de - - [21/Jan/2000:18:37:37 +0100] "GET /bilder/ifi_logo2.gif HTTP/1.0" 304 0
tipc015.informatik.uni-leipzig.de - - [21/Jan/2000:18:37:38 +0100] "GET /bilder/uni_logo2.gif HTTP/1.0" 304 0
news.uni-leipzig.de - - [21/Jan/2000:18:37:41 +0100] "GET /ifi/lehre/Heyer9900/kap23/img020.GIF HTTP/1.0" 304 0

```

Abb. 3: Ausschnitt der Web-Log-Datei

- *Ziel* des Zugriffs, also die Adresse der Web-Seite, auf die zugegriffen wurde (relativ zur root-Adresse des Web-Servers)
- *Returncode*, welcher den Erfolg bzw. Mißerfolg des Zugriffs repräsentiert
- gelesene *Datenmenge* in Bytes.

Die Hinzunahme weiterer Informationen aus anderen Datenquellen bereitet konzeptionell wenig Schwierigkeiten und ist insbesondere für personenbezogene Auswertungen (z.B. bezüglich Kundenverhalten) erforderlich. Hier liegt ein Hauptproblem in der mit einer i.d.R. erforderlichen Preisgabe der Identität des Nutzers verbundenen Datenschutzgewährleistung, worauf hier jedoch nicht näher eingegangen werden soll.

3.2 Auswertungsdimensionen und Hierarchien

Aus o.g. Dateistruktur ergeben sich als *Dimensionen* der Auswertung entsprechend *Quelle*, *Ziel* und *Zeit* sowie als *Bewertungsmetriken* die *Anzahl der Zugriffe* nebst übertragenen *Datenumfang* in Bytes. Abb. 4 zeigt das der Analyse zugrundeliegende Datenmodell inklusive der nachfolgend erläuterten Hierarchien der Dimensionen, welche eine inhaltsbasierte Kategorisierung von Web-Seiten erlauben.

Dimension Zeit

Gemäß der in der Web-Log-Datei sekundengenau gespeicherten Zugriffszeit ergibt sich für diese Dimensionen die Aufschlüsselung

Jahr → *Monat* → *Tag* → *Stunde* → *Minute* → *Sekunde*

Allerdings erscheint eine solch feine Hierarchie nur in Ausnahmefällen sinnvoll. Somit wurden, auch aus Speicherplatzgründen, die Daten lediglich auf das Granulat *Tag* aggregiert und zusammen mit *Monats*- und *Jahres*werten gespeichert.

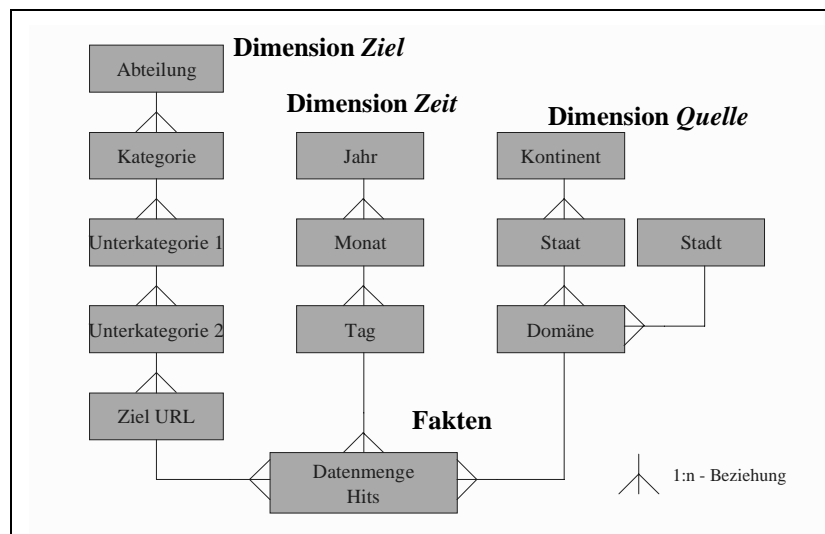


Abb. 4: Datenmodell für Web-Zugriffe

Dimension Quelle

Die Quelle des Web-Zugriffs leitet sich ebenfalls weitestgehend von der in der Web-Log-Datei gespeicherten zusammengesetzten Rechnernamen ab. Die Aufschlüsselung ist hier

Kontinent → **Staat** → **Domäne** → **Rechnername**, wobei aus der Domäne ggf. der Name der assoziierten **Stadt** automatisch ableitbar und speicherbar ist (z.B. *uni-leipzig.de* → *Leipzig*).

Aus Platz- (und auch Datenschutzgründen) wurde auch hier auf das zu feine Zugriffsgranulat *Rechnername* verzichtet und somit die Zugriffe auf Domänen-Level aggregiert. Bei Quellangaben, welche nicht im sog. *Domain Name Service* registriert sind (z.B. IP-Adressen) erfolgt die Sammlung unter der Domäneninstanz *Not Registered*. Desweiteren werden nur Zugriffe aus Ländern registriert - und somit Staaten und Kontinenten zugeordnet - die eine eindeutig einem Staat zuweisbare Top-Level-Domäne besitzen. Beispiele sind *.de* für Deutschland oder *.mx* für Mexiko. Internationale Top-Level-Domänen wie *.com*, *.edu* oder *.org* werden unter der Domäneninstanz *International* zusammengefaßt.

Dimension Ziel

Zur inhaltsorientierten Auswertbarkeit der Web-Seiten wurde neben Verwendung der Informationen ihrer physischen Organisation (Dateipfade) die inhaltliche Hierarchie möglichst generisch abgebildet. Die Hierarchie der Web-Seiten für Lehrstühle innerhalb des betrachteten Institutes für Informatik benötigt maximal 5 Stufen. Sie ergibt sich folgt

Lehrstuhl → **Kategorie** → **Unterkategorie 1** → **Unterkategorie 2** → **Ziel URL**

Abb. 5 zeigt beispielhaft einen Ausschnitt der Organisation der Web-Seiten für die Abteilung Datenbanken, innerhalb derer die Anwendung entwickelt wurde. Die Ebene **Lehrstuhl** bezeichnet als Wurzel

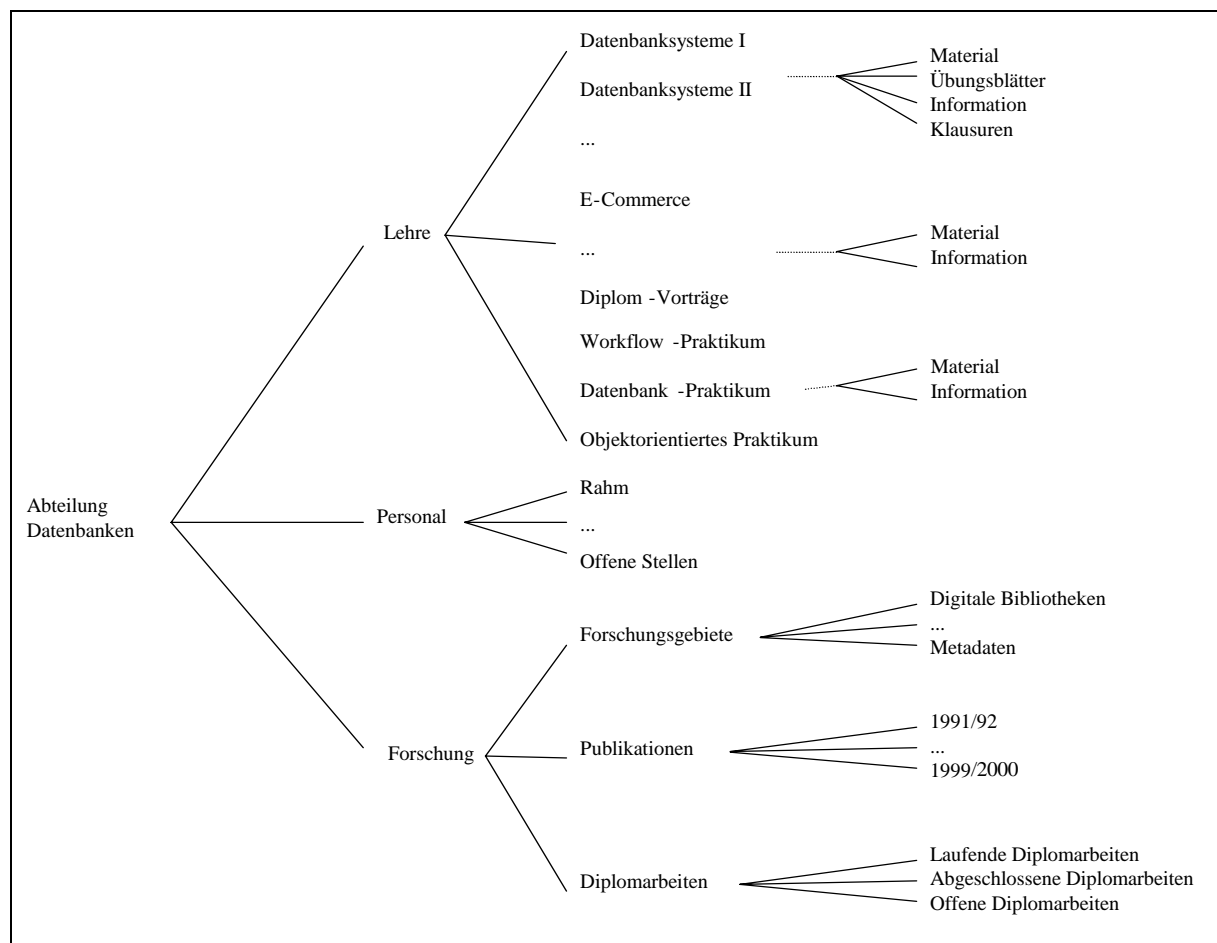


Abb. 5: Ausschnitt der Kategorisierung der Web-Seiten der Abteilung Datenbanken

den einer Web-Seite assoziierten Lehrstuhl des betrachteten Institutes. Jede Abteilung umfaßt mehrere, oftmals von Abteilung zu Abteilung unterschiedliche inhaltliche Kategorien von Web-Seiten (z.B. in Abb. 5 instantiiert durch *Lehre, Forschung, Personal*,...), welche wiederum Unterkategorien umfassen können (z.B. die Vorlesung *Datenbanksysteme1* als Unterkategorie der *Lehre* innerhalb der Abt. Datenbanken) etc. Auf der untersten Ebene ist schließlich der Pfadname der Datei (Ziel URL) gespeichert.

Die Hierarchieebenen der meisten betrachteten Lehrstühle entwickeln einen geringen Detailgrad (teilweise nur eine Hierarchiestufe unterhalb der Wurzelseite), so daß in diesem Falle lediglich die Gesamtanzahl der Zugriffe auf ein Abteilung sinnvoll bewertet werden kann.

3.3 Kategorisierung und Abbildung von Web-Zugriffen auf Dimensionen

Die Abbildung von Web-Zugriffen auf Quell-, Ziel- und Zeitdimensionen und deren Hierarchieebenen wird über DTS-Skripte bei der Befüllung des relationalen Starschema aus der Basistabelle *access_log* (vgl. Abb. 2) im Warehouse abgewickelt. Für jedes Tupel (entspricht einem Satz in der Web-Log-Datei) werden die unter den bereits dimensionsbezogenen Attributen abgelegten Werte mittels einer umfangreichen Menge von SQL-Anweisungen ihrer Hierarchie-Ausprägung zugeordnet. Zur Zuordnung werden Pattern-Matching- und String-verarbeitende Funktionen auf den jeweiligen Dateipfad einer Web-Seite angewendet. Dabei werden die inhärenten inhaltlichen Informationen ausgenutzt (z.B. *.../db/lehre/..* wird abgebildet in die „Ausprägung *Lehre der Unterkategorie1* Datenbanken, Dimension *Ziel*“, *...uni-leipzig.de.* wird abgebildet in „Ausprägung *Leipzig* der Hierarchie *Stadt*, Dimension *Quelle*“ etc.). Für nicht per Skript automatisch assoziierbare Seiten existieren entsprechende „Sammel“-Hierarchieausprägungen: ist z.B. eine Web-Seite den Ziel-Hauptkategorien *Lehre, Forschung* oder *Personal* nicht per Skript zuzuordnen, so wird sie in der Kategorie *Unbekannt* abgelegt.

Zur Bewertung der Zugriffsfrequenz auf die so inhaltlich kategorisierten Seiten wird die einem Eintrag assoziierte Byte-Menge, mit passenden Fremdschlüsseln versehen, in der Faktentabelle abgelegt, wobei Aggregationen zur Reduktion des Speicherbedarfs durchgeführt werden (z.B. Speicherung der Summe aller Einträge eines Tages pro Hierarchie-Ebene bezogen auf alle anderen Dimensionen).

4 Benutzerschnittstelle

Als Benutzerschnittstelle kommen Business-Intelligence-Werkzeuge verschiedener Anbieter in Frage. Einzige Voraussetzung ist, daß die eingesetzten Werkzeuge mittels OLE DB for OLAP auf die multidimensionalen Daten, welche durch die MICROSOFT OLAP Services verwaltet werden, zugreifen. Eine wichtige Anforderung war daneben, daß die Auswertung über einen Web-Browser möglich sein sollte, um die Enbindung unserer DB-gestützten Analyse in das WWW zu ermöglichen.

Z.Z. wird COGNOS POWERPLAY 6.6 [6] am Lehrstuhl eingesetzt. Vergleichbar anderen Werkzeugen dieser Kategorie bietet POWERPLAY dynamische und statische Report-Funktionalität, verschiedenste Präsentationsmodi (unterschiedliche Typen von Balken-, Kreisdiagrammen etc.) und flexible Drill-Down, Slice and Dice-Mechanismen auf den unterschiedlichsten multidimensionalen Datenquellentypen. POWERPLAY ist zudem in den gängigen Web-Browsern (Netscape, Internet Explorer) nutzbar. Frontends wie MICROSOFT EXCEL sind dagegen funktional eingeschränkt (z.B. keine Reports, beschränkt auf eine Datenquelle), erfordern die Installation spezieller Zugangskomponenten und sind nicht auf allen Browsern ablauffähig.

Die folgenden zwei Abbildungen zeigen beispielhaft Reports, welche mit COGNOS POWERPLAY erstellt wurden. Abb. 6 präsentiert die Aufteilung der Zugriffe auf die Hauptkategorien *Lehre, Personal, Forschung*, sowie allgemeine bzw. nicht zuordnenbare Zugriffe (*Unknown*) auf die Seiten der Abteilung Datenbanken, aufgeschlüsselt nach Monaten (Juli 1999-Juli 2000). Oberhalb der Graphik kann die entsprechende Hierarchie-Ausprägung gewählt werden (*Source, Datenbanken* etc.) Ein Click auf die vergleichbar HTML-Hyperlinks formatierten Ausprägungen unterhalb der Balken (z.B. „Juli“) führt zur dynamischen Aufschlüsselung der täglichen Zugriffsfrequenzen für diesen Monat.

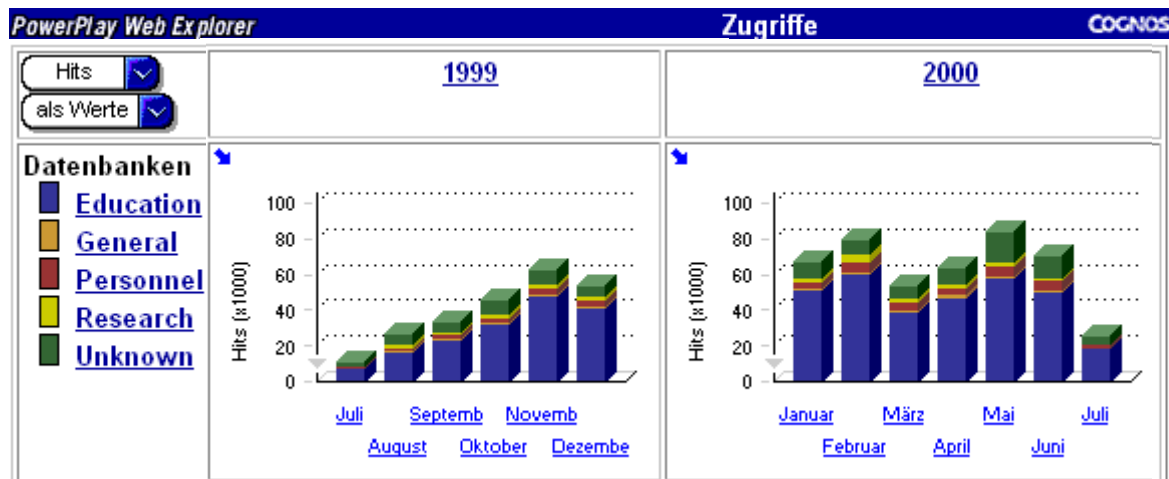


Abb. 6: Report der Web-Zugriffe für Juli 1999 - Juli 2000 (monatliche Aufschlüsselung)

Mit COGNOS POWERPLAY kann die Darstellung von Zugriffshäufigkeiten übersichtlicher gestaltet werden, in dem nur diejenigen Balken angezeigt werden, die zu 80 % der Gesamtsumme aller Zugriffe am stärksten beigetragen haben. Abb. 7 zeigt die Zugriffshäufigkeit auf die den Lehrgebieten der Abteilung Datenbanken zugeordneten Seiten. 80 % Zugriffe fallen somit auf die dargestellten Veranstaltungen *Datenbanksysteme 1*, *Datenbanksysteme 2*, *Implementierung von DBS 1*, ..., *Workflow-Management*, welche gleichzeitig auch am häufigsten frequentiert werden.

Unsere Web-Log-Anwendung läuft stabil seit Januar 2000 und verwaltet derzeit (Juli 2000) nahezu 2 Mill. Web-Zugriffe (Umfang mehrere 100 MB), die seit Juli 1999 auf dem Web-Server des Institutes ausgeführt wurden. Die Anwendung ist verfügbar unter

<http://www.informatik.uni-leipzig.de/ifi/abteilungen/db/frame/db-statistik.html>

oder einfach über die Informatik-Homepage der Universität Leipzig

<http://www.informatik.uni-leipzig.de>, Click auf **Datenbanken** → ...**Web-Zugriffe**....

Zur korrekten Ausführung müssen MS OLAP SERVICES lokal auf dem Nutzerrechner installiert sein.

5 Erweiterungsmöglichkeiten

Speziell für die in anderen Anwendungsfeldern naturgemäß vollkommen unterschiedlichen Ausprägungen der Ziel-Dimension (z.B. Produkthierarchien im E-commerce) ermöglicht die generische Strukturierung der Ziel-Dimension eine einfache Erweiterung auf andere Anwendungsgebiete, ohne daß z.B. das zugrundeliegende Schema zwangsläufig geändert werden müßte. Sollte trotzdem z.B. eine Hierarchieschachtelung nicht ausreichend tief sein, so ist dank der verwendeten flexiblen DB-Lösung eine notwendige Schemaänderung mit verhältnismäßig geringem Aufwand durchführbar.

Eine wesentliche Motivation der Web-Analyse ist die Bewertung der Attraktivität von Web-Angeboten hinsichtlich des Kundenverhaltens. Hier ist eine Anpassung des Schemas an detailliertere Web-Log-Protokolle, welche Kundendaten z.B. mit Nutzeridentifikation über spezifische Abfragen oder Cookies sammeln, denkbar.

Weitere Ausbaufähigkeit besteht durch eine Steigerung der Qualität der Auswertungen. Durch die Entscheidung für ein kommerziell häufig verwendetes Werkzeug wie MICROSOFT SQL Server können z.B. dessen zukünftige DBS-Erweiterungen genutzt werden. So wird Data Mining-Funktionalität Bestandteil des SQL-Server 2000 sein [7].

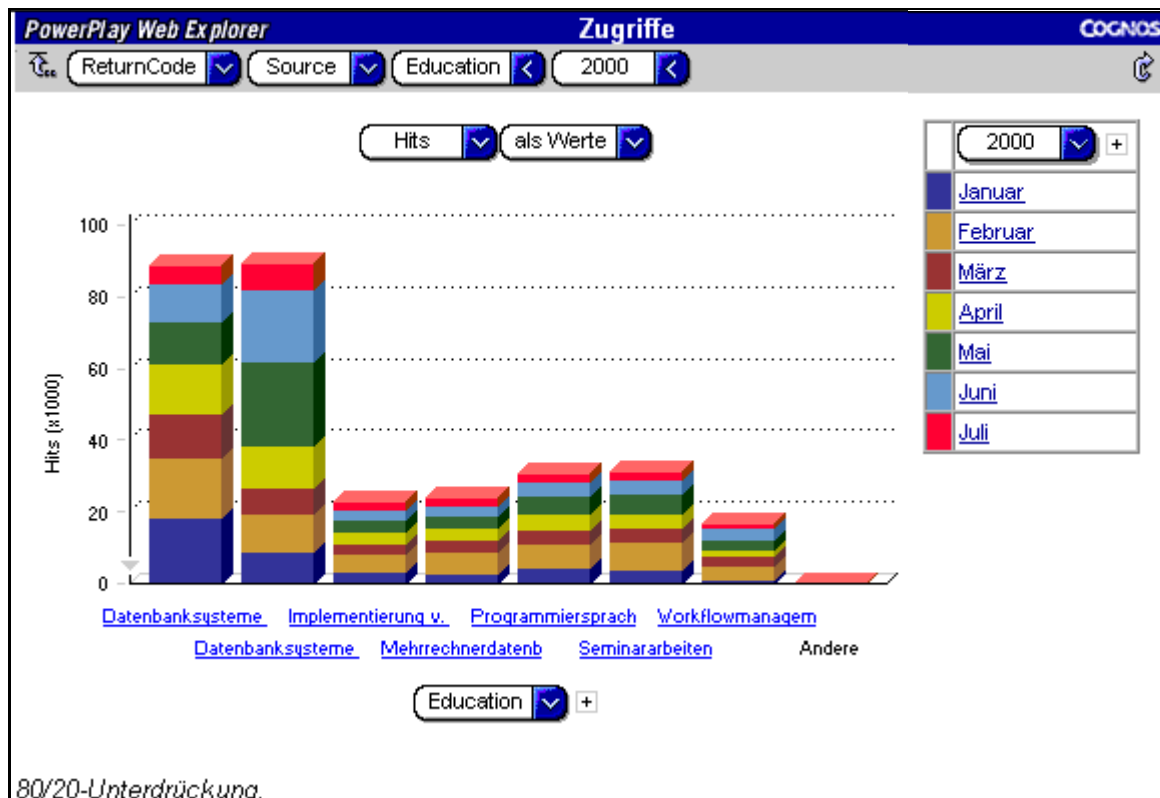


Abb. 7: Report Lehrveranstaltungen der Abt. Datenbanken (80/20-Unterdrückung)

6 Zusammenfassung

Die Auswertung von Web-Zugriffen ist von großer und zunehmender Wichtigkeit für alle Betreiber von Web-Servern, insbesondere zur optimierten Gestaltung von Informations- und Produktangeboten, Erhöhung der Kundenbindung bzw. Gewinnung neuer Kunden, Performance-Optimierung durch adäquate Hardware-Konfigurierung, etc. Zur Erlangung einer ausreichend hohen Flexibilität und Skalierbarkeit auf große Benutzer- und Zugriffsmengen sollte eine DB-gestützte Analyse der Web-Zugriffe erfolgen. Dazu sollte ein auf das jeweilige Umfeld zugeschnittene Data Warehouse-Lösung eingesetzt werden, verbunden mit über Web-Browser nutzbare OLAP-Werkzeugen.

Wie anhand einer an der Universität Leipzig entstandenen Beispiellösung gezeigt wurde, können diese Ziele mit heutigen Werkzeugen relativ schnell und zu moderaten Kosten erreicht werden. Dies gilt insbesondere, da in den meisten Umgebungen relationale DBS ohnehin eingesetzt werden und diese zumindest in den aktuellen Versionen Data Warehouse-Unterstützung bieten. So ist das verwendete DBS SQL-Server unter NT bzw. Windows2000 auch in kleineren Firmen weit verbreitet und beinhaltet Unterstützung für ETL und OLAP. Als Endbenutzer-Tool stehen leistungsfähige Werkzeuge für die gängigen Browser zur Verfügung. Durch die weitgehende Verwendung kommerziell erhältlicher Systeme konnte der Aufwand zur Eigenentwicklung auf ein Minimum beschränkt werden (u.a. Skript zur Datenextraktion aus dem Web-Log und zum Datentransfer). Verbesserungen in neuen Versionen der Produkte können unmittelbar genutzt werden. So können die in SQLServer2000 verfügbaren Data Mining-Funktionen für weitergehende Analysen auf dem bestehenden Datenbestand eingesetzt werden. Die DB-gestützte Vorgehensweise ermöglicht ferner eine relativ einfache Anpaßbarkeit an die umgebungsspezifische Strukturierung der Web-Angebote sowie die Hinzunahme weiterer Datenquellen.

7 Literatur

- [1] Chen, C.; Cochinwala, M. et.al.: *Internet Traffic Warehouse*. Proc. ACM SIGMOD Conf., Dallas, TX, 2000
- [2] *Data Warehouse Tools – DWT*. Band I und II. Projektbericht, Informatikzentrum der Sparkassenorganisation GmbH (SIZ), Univ. Leipzig, Univ. Würzburg, Oktober 1999.
- [3] Dyché, J.: *Click stream analysis*. <http://www.teradatareview.com/spring00/dyche.html>, 2000
- [4] Müller, R.; Stöhr, T.; Rahm, E.; Quitzsch, S.: *Evaluierung und Produktvorschlag für ein technisches Metadaten-Management im R+V-Data Warehouse*. Projektbericht, Inst. für Informatik, Univ. Leipzig, November 1998.
- [5] Pendes, N.: *The OLAP Report - What is OLAP?*; <http://www.olapreport.de/fasmi.htm>, 1999
- [6] *Product information: COGNOS POWERPLAY*; <http://www.cognos.com/powerplay/index.html>, 2000
- [7] *Product information: SQL Server 2000 Technology Guide*, <http://www.microsoft.com/SQL/productinfo/sql2ktec.htm>
- [8] Riggs, S.: *Collecting web data*. <http://www.teradatareview.com/spring00/riggs.html>, 2000
- [9] *The Datawarehousing Information Center: Electronic Commerce Industry Tools*; <http://www.dwinfocenter.org/ecommerce.html>, 2000
- [10] Whiting, R.: *Web data piles up*. <http://www.informationweek.com/shared/printArticle?article=infoweek/785/prdatabase.htm&pub=iwk>, 2000