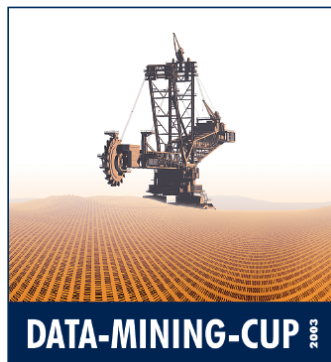

**Vortrag auf den 4. DATA-MINING-CUP Anwendertagen
(Chemnitz, 18.-20. Juni 2003)**



<http://www.data-mining-cup.de>

Copyright-Hinweis:

Das Urheberrecht des folgenden Vortrags liegt beim Author. Verbreitung, Vervielfältigung und Kopie, auch auszugsweise, ist nur mit schriftlicher Genehmigung des Authors erlaubt.

Data-Warehouse-Einsatz zur kundenorientierten Web- zugriffsanalyse

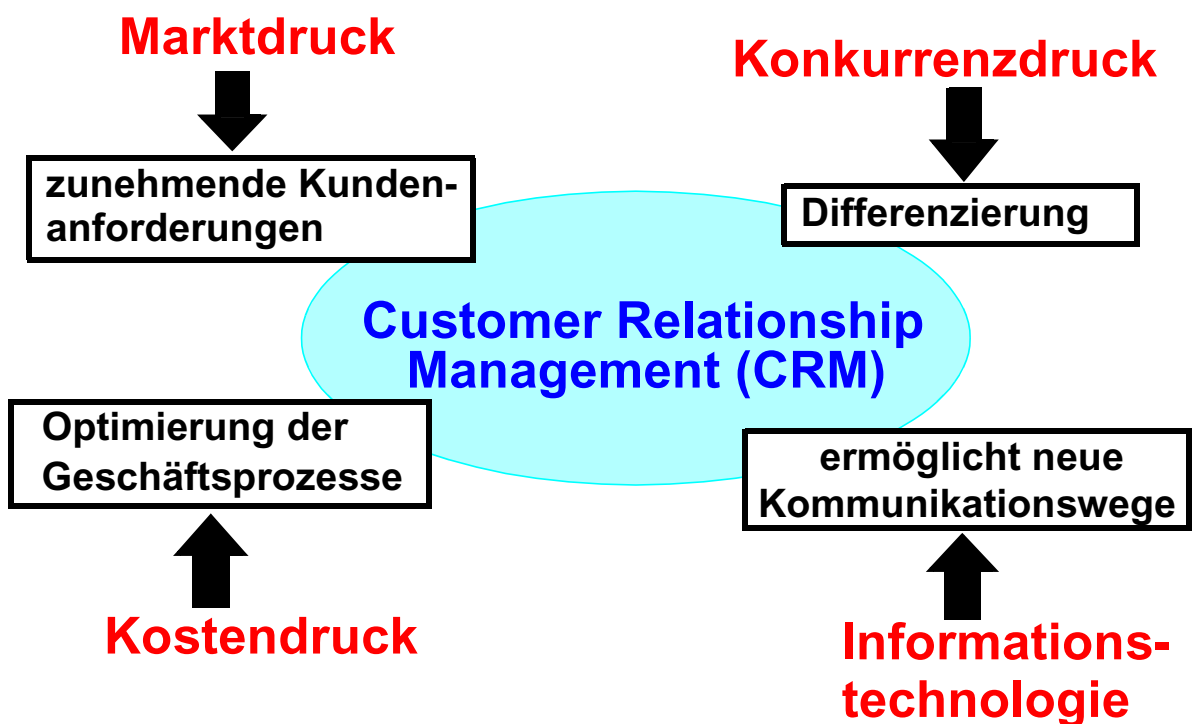
Prof. Dr. Erhard Rahm

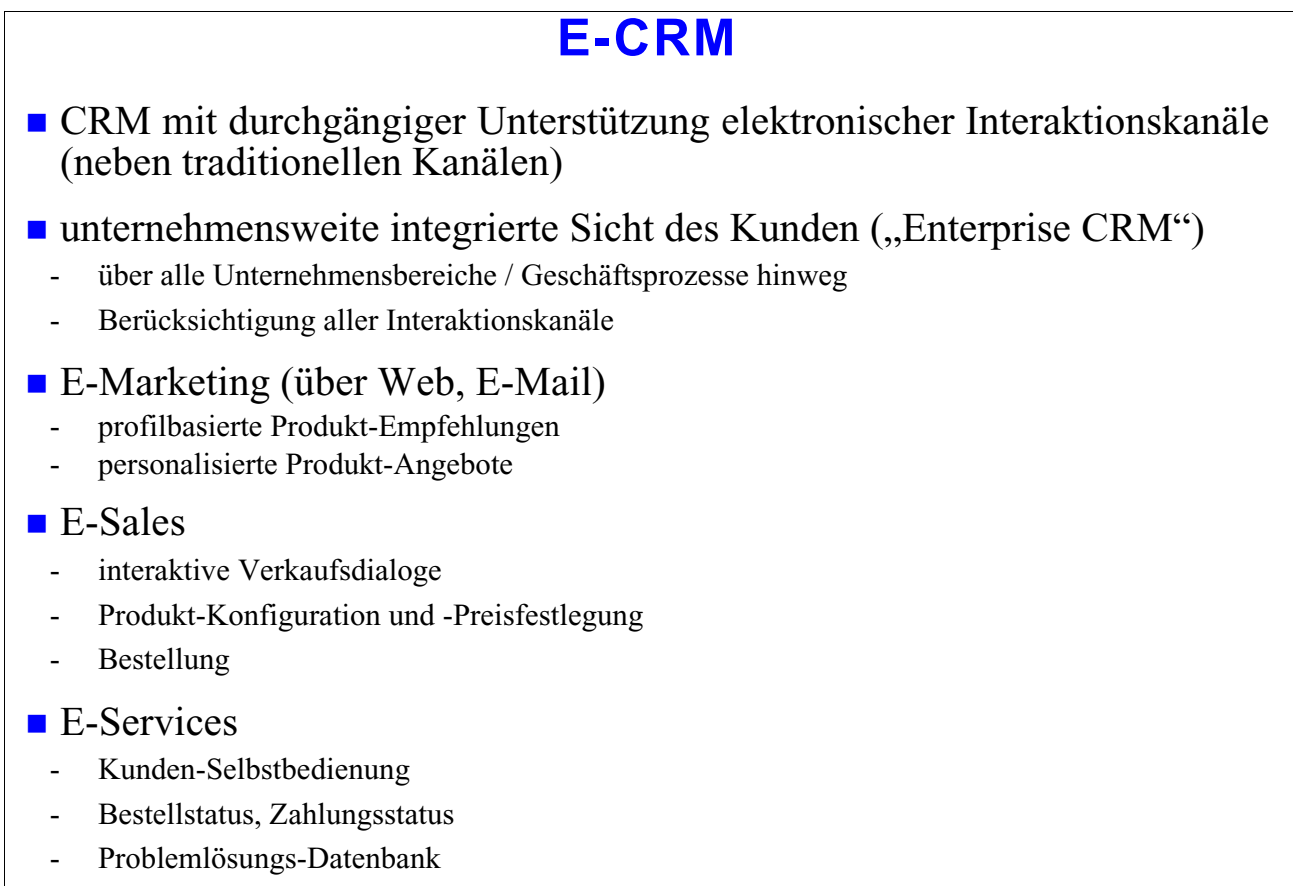
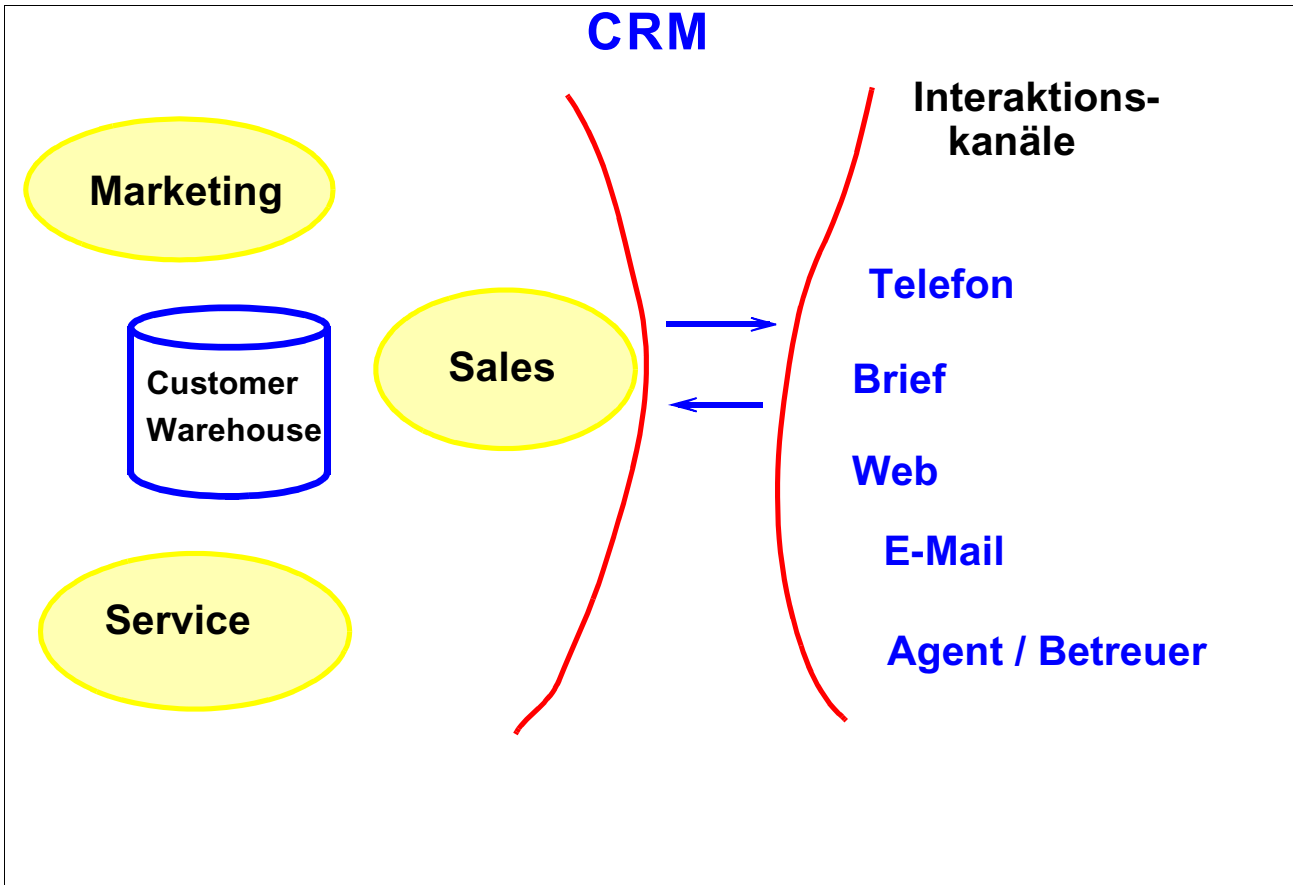
Universität Leipzig

<http://dbs.uni-leipzig.de>

- E-CRM
- Vorgehensweise Webzugriffsanalyse
- Umsetzung
 - Datenaufbereitung
 - Data-Warehouse-Schema
- Nutzung der Analyseergebnisse

Notwendigkeit der Kunden-Fokussierung





Bewertung von Websites

Web-Investition

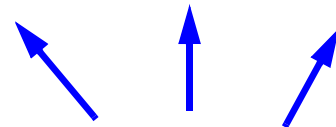
- Geld
- Know How
- Zeit

Web-Nutzung

- Navigation
- Suche
- Transaktionen

Web-Ergebnis

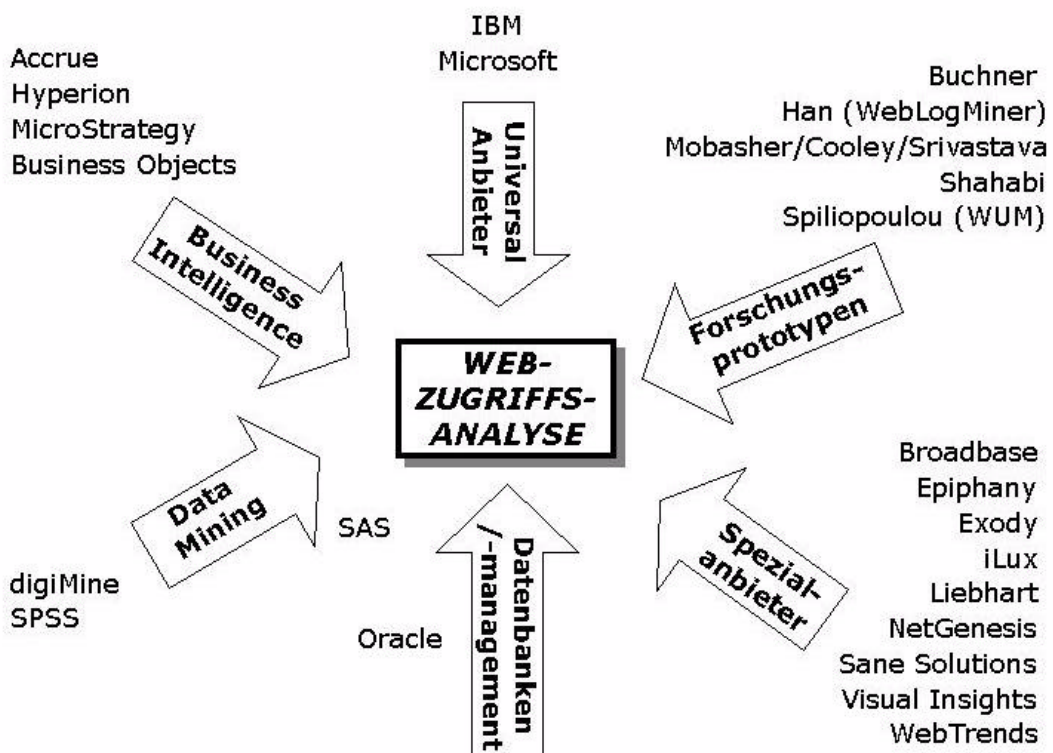
- Wer, Wann, Was, Wo, Warum (nicht)
- ROI (Return of Investment) ???



Web Usage Mining

Nutzung von Web-Logs
sowie ergänzenden Informationsquellen

Werkzeug-Anbieter



Vorgehensweise zur Webzugriffsanalyse (1)

■ Ziele des Web-Auftritts

- Wünsche der angestrebten Zielgruppen vs. Wünsche des Unternehmen / Anbieters
- Bsp.: Informationbereitstellung, Gewinnung neuer Kunden, Ausbau bestehender Kundenbeziehungen, Service-Angebote (Entlastung eigener Mitarbeiter) ...

■ Wie sollen Ziele erreicht werden ?

- Wie werden Kunden / Interessenten auf Website gebracht ?
- Inhalte / Gestaltung der Website

■ Welche Ergebnisse soll Zugriffsanalyse bringen ?

- Wissen über Nutzungsverhalten gewinnen
- Wissen über Besucher / Kunden gewinnen
- **einfache statistische Bewertungs-Metriken:** Zugriffshäufigkeiten, #Besucher, Verweilzeiten ...
- **Referrer-Analyse:** Von woher kommen die Besucher (Effektivität von Werbemaßnahmen)
- **Konversionsraten:** Anteil der Besucher, die in bestimmten Zustand wechseln (Kauf, Angebots-einholung, Kontaktaufnahme mit persönlichem Vermittler, etc.)
- **Return on Investment (ROI):** Umsatz- und Gewinn-Summe der Besucher

Vorgehensweise zur Webzugriffsanalyse (2)

■ Festlegung des Realisierungsansatzes zur Webzugriffsanalyse

- einfacher Ansatz
- Datenbank-basierter Ansatz
- Data-Warehouse-Ansatz

■ Datenaufbereitung und Datenintegration

■ Auswertung / Analyse der Ergebnisse

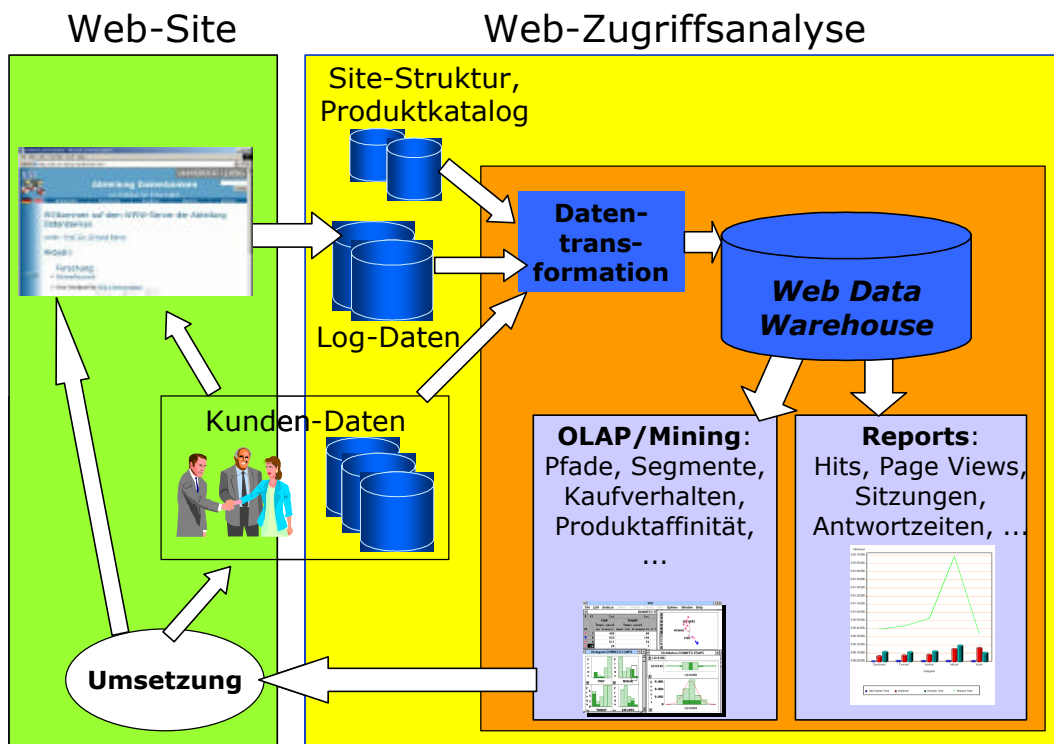
■ Reaktion auf Ergebnisse

- Umgestaltung der Website
- Empfehlungen / Personalisierung
- Marketing-Aktivitäten, ...

Realisierungsalternativen

- einfache Lösungen zur Web-Log-Auswertung (ohne Datenbank)
 - direkte, unflexible Auswertung auf einer Datei
 - Performance-Probleme aufgrund riesiger Datenmengen (viele MKlicks, GKlicks)
 - offline erstellte vordefinierte Reports
 - keine bzw. umständliche Verknüpfung mehrerer Datenquellen
 - fehlende inhaltliche Kategorisierung (welche Produkte?)
 - unzureichende Nutzerauswertung (kundenbezogene Auswertung)
- Datenbank-gestützte Lösung
 - Skalierbarkeit, Verfügbarkeit, Erweiterbarkeit
 - interaktive Auswertungsmöglichkeiten
- Data Warehouse-basierter Ansatz
 - flexible Integration mehrerer Datenquellen
 - Nutzung mächtiger OLAP-/Data Mining-Werkzeuge
 - inhaltliche Kategorisierung der Zielseiten
 - kundenbezogene Auswertungen
 - Kopplung mit anderen Kundeninformationen -> CRM

Warehouse-Einsatz zur Webzugriffsanalyse



Datenquellen

- Aufzeichnungen von Website-Zugriffen
 - (standardisierte) Log-Dateien der Web-Server (Web-Log)
 - Aufzeichnung der Netzwerk-Kommunikation (TCP/IP Packet Sniffing)
- Applikations-Server-Aufzeichnungen
 - Aufruf von Anwendungsfunktionen
 - keine standardisierte Protokollformate; Eigenentwicklung erforderlich
- Besucher-/Kunden-Informationen
 - Kunden-DB
 - Registrierungsdaten
- Produktinformationen / -kataloge
- Website-Topologie
- Weitere Datenquellen
 - Hilfstabellen für Datentransformation
 - durchgeführte Marketing-Aktionen ...

Aufbau der Log-Einträge

green.dresdnerbank.de - - [11/Feb/2003:15:33:13 +0200] "GET /skripte/WMS/inhalt2.html HTTP/1.0" 200 7885
"http://www.google.de/search?q=state+activity+chart+workflow&hl=de&meta=" "Mozilla/4.0 (compatible;
MSIE 5.01; Windows NT; drebalE-ZE)"

Common Log Format

Host:	Client-Hostname, IP-Adresse
Ident:	Identität (Benutzername)
Authuser:	Authorization user ID
Date-Time:	dd/mm/yyyy:hh:mm:ss
Zone:	Zeitzone: +dddd oder -dddd
Request:	Erste Zeile einer Anfrage, z.B. "GET / http/1.0"
Status:	Antwortstatus-Code des Servers
Bytes:	Anzahl der übertragenen Bytes

zusätzlich im Extended Common Log Format

Referring URL:	URL der Seite, von der Client kommt
User Agent:	Browser, Betriebssystem des Clients
Cookie:	Cookie-Eintrag
Selbstdefinierte Felder	

HTML-Seite mit mehreren Dateien

Adresse Wechsell zu

3 x JavaScript **UNIVERSITÄT LEIPZIG** **GIF**

Database Group **GIF**

within Department of Computer Science

People **GIF** Research **GIF** Study **GIF** Service **GIF** Internals **CSS**

Welcome to the WWW-Server of the Database Group at the University of Leipzig

Head: [Prof. Dr. Erhard Rahm](#)

News:

Research **JPG**

- new Benchmark for [XML data management](#)
- New Working Group "[Web and Databases](#)"

GIF **PNG** **GIF** **New**

Log-Auszug (dbs.uni-leipzig.de)

```
crawl1.googlebot.com -- [01/Mar/2003:08:10:35 +0100] "GET /seminararbeiten/semSS99/arbeit3/kapitel5.html HTTP/1.0" 200 17816
crawl2.googlebot.com -- [01/Mar/2003:08:11:50 +0100] "GET /de/buecher/DBSI-Buch HTTP/1.0" 301 356
crawl1.googlebot.com -- [01/Mar/2003:08:13:00 +0100] "GET /de/Research/webusage.html HTTP/1.0" 200 11792
crawl2.googlebot.com -- [01/Mar/2003:08:14:03 +0100] "GET /de/vorlesungen/DBS1/WS9900/uebungen.html HTTP/1.0" 200 9653
ilabpc07.informatik.uni-leipzig.de -- [01/Mar/2003:08:14:16 +0100] "GET /vorlesungen/DBS1/WS0001/menueVorDBS1.html HTTP/1.1" 200 10856
ilabpc07.informatik.uni-leipzig.de -- [01/Mar/2003:08:14:16 +0100] "GET /js/menuDef.de.js HTTP/1.1" 200 5298
ilabpc07.informatik.uni-leipzig.de -- [01/Mar/2003:08:14:16 +0100] "GET /js/hierMenus.js HTTP/1.1" 200 23266
ilabpc07.informatik.uni-leipzig.de -- [01/Mar/2003:08:14:16 +0100] "GET /js/addFunc.js HTTP/1.1" 200 3949
ilabpc07.informatik.uni-leipzig.de -- [01/Mar/2003:08:14:16 +0100] "GET /css/db_style.css HTTP/1.1" 200 2103
ilabpc07.informatik.uni-leipzig.de -- [01/Mar/2003:08:14:16 +0100] "GET /bilder/db_logo3.gif HTTP/1.1" 200 6489
ilabpc07.informatik.uni-leipzig.de -- [01/Mar/2003:08:14:16 +0100] "GET /bilder/logouni_small.gif HTTP/1.1" 200 4058
ilabpc07.informatik.uni-leipzig.de -- [01/Mar/2003:08:14:16 +0100] "GET /bilder/german.gif HTTP/1.1" 200 71
ilabpc07.informatik.uni-leipzig.de -- [01/Mar/2003:08:14:16 +0100] "GET /bilder/british.gif HTTP/1.1" 200 1453
ilabpc07.informatik.uni-leipzig.de -- [01/Mar/2003:08:14:16 +0100] "GET /bilder/transparent_dot.gif HTTP/1.1" 200 43
ilabpc07.informatik.uni-leipzig.de -- [01/Mar/2003:08:14:16 +0100] "GET /bilder/db_left.png HTTP/1.1" 200 193
ilabpc07.informatik.uni-leipzig.de -- [01/Mar/2003:08:14:16 +0100] "GET /bilder/db_left_top.png HTTP/1.1" 200 116
ilabpc07.informatik.uni-leipzig.de -- [01/Mar/2003:08:14:17 +0100] "GET /bilder/menu/tri.gif HTTP/1.1" 200 67
crawl1.googlebot.com -- [01/Mar/2003:08:15:02 +0100] "GET /seminararbeiten/semSS99/arbeit3/kapitel6.html HTTP/1.0" 200 13481
info002.informatik.uni-leipzig.de -- [01/Mar/2003:08:15:15 +0100] "GET /en/index.html HTTP/1.0" 200 11027
info002.informatik.uni-leipzig.de -- [01/Mar/2003:08:15:15 +0100] "GET /js/menuDef.en.js HTTP/1.0" 304 -
info002.informatik.uni-leipzig.de -- [01/Mar/2003:08:15:15 +0100] "GET /js/hierMenus.js HTTP/1.0" 304 -
info002.informatik.uni-leipzig.de -- [01/Mar/2003:08:15:15 +0100] "GET /js/addFunc.js HTTP/1.0" 304 -
info002.informatik.uni-leipzig.de -- [01/Mar/2003:08:15:16 +0100] "GET /css/db_style.css HTTP/1.0" 304 -
info002.informatik.uni-leipzig.de -- [01/Mar/2003:08:15:16 +0100] "GET /bilder/db_left.png HTTP/1.0" 304 -
info002.informatik.uni-leipzig.de -- [01/Mar/2003:08:15:16 +0100] "GET /bilder/db_logo3.gif HTTP/1.0" 304 -
info002.informatik.uni-leipzig.de -- [01/Mar/2003:08:15:16 +0100] "GET /bilder/logouni_small.gif HTTP/1.0" 304 -
info002.informatik.uni-leipzig.de -- [01/Mar/2003:08:15:16 +0100] "GET /bilder/german.gif HTTP/1.0" 304 -
info002.informatik.uni-leipzig.de -- [01/Mar/2003:08:15:16 +0100] "GET /bilder/british.gif HTTP/1.0" 304 -
info002.informatik.uni-leipzig.de -- [01/Mar/2003:08:15:16 +0100] "GET /bilder/transparent_dot.gif HTTP/1.0" 304 -
info002.informatik.uni-leipzig.de -- [01/Mar/2003:08:15:16 +0100] "GET /bilder/db_left_top.png HTTP/1.0" 304 -
info002.informatik.uni-leipzig.de -- [01/Mar/2003:08:15:16 +0100] "GET /bilder/db_logo_back_ns4.jpg HTTP/1.0" 304 -
info002.informatik.uni-leipzig.de -- [01/Mar/2003:08:15:18 +0100] "GET /bilder/menu/tri.gif HTTP/1.0" 304 -
```

Transformations-Aufgaben

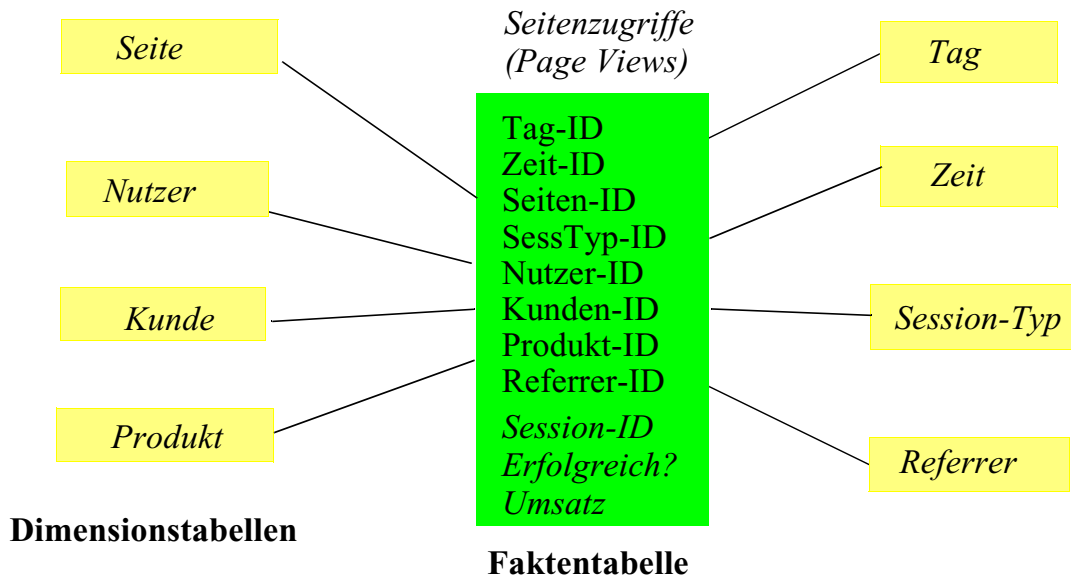
- Eliminieren irrelevanter Log-Einträge
 - eingebundene Bilder und Grafiken (.gif, .jpg, .png ...)
 - Applets, Skripte (.jar, .class, .js ...), Style-Sheets (.css) ...
- Erkennung und Eliminierung von **Roboter-Zugriffen**
- evtl. auch Beseitigung von Zugriffen lokaler Nutzer, Indexerstellung etc.
- Namensauflösung für numerische IP-Adressen (reverse DNS lookup)
- **Nutzeridentifikation:**
 - Zuordnen von Log-Einträgen zu einzelnen Benutzern
- **Session-Identifikation**
 - Zuordnung der Zugriffe eines Benutzers zu Sitzungen (Sessions)
 - Sitzungsidentifikation erfordert (temporäre) Nutzeridentifikation
 - maximaler Zeitabstand aufeinander folgender Zugriffe unter eingestelltem Grenzwert (z.B. 20-30 Minuten)

Nutzeridentifikation

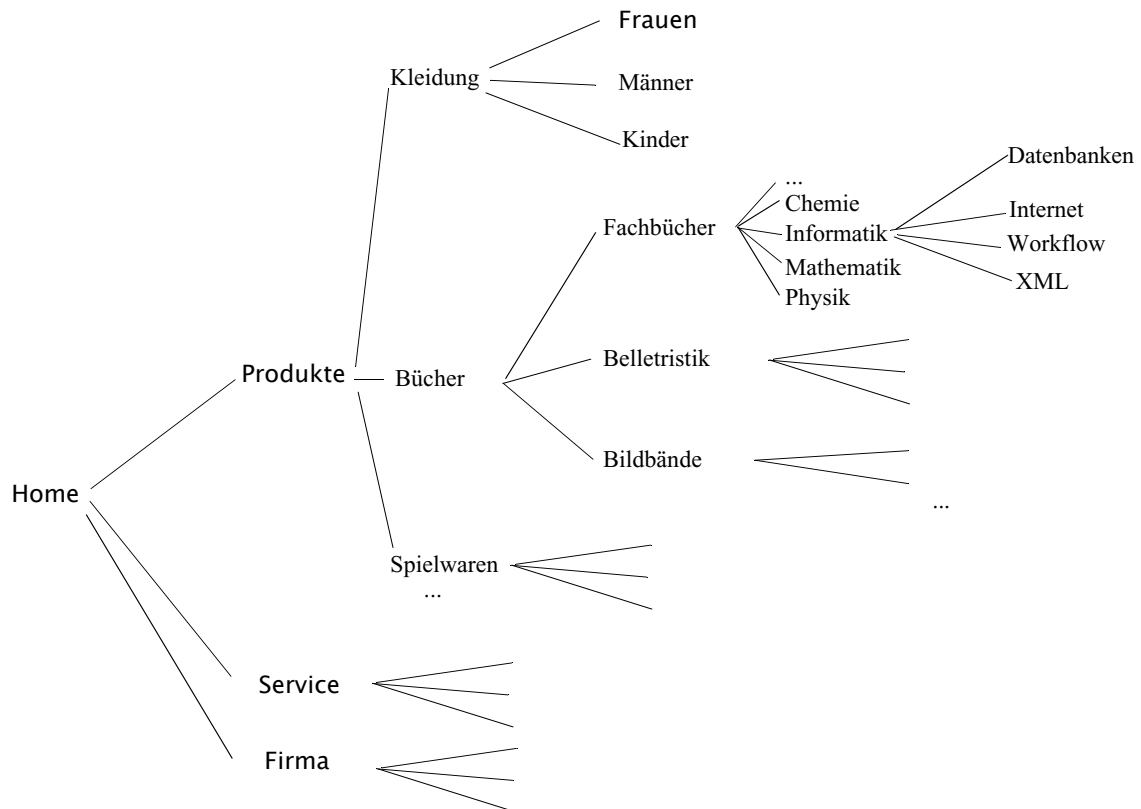
- 3 Stufen der Nutzeridentifikation
- temporäre Nutzeridentifikation
 - innerhalb einer Sitzung
- sitzungsübergreifende Nutzeridentifikation
 - Wiedererkennung eines Benutzers
- Personifizierung

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

Data Warehouse-Schema zur Webzugriffsanalyse (Beispiel)

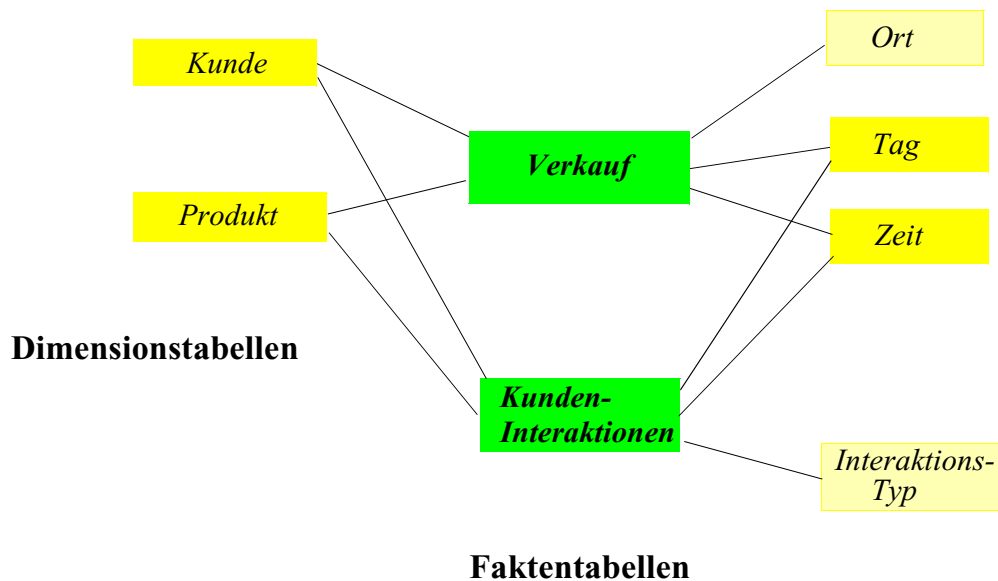


Hierarchische Inhaltskategorisierung



CRM-Erweiterungen

- CRM-Unterstützung über weitere Faktentabellen
- Verknüpfung über *gemeinsame Dimensionen*



Umsetzung von Analyse-Erkenntnissen

- Reorganisation der Website
 - verbesserte Navigation
 - einfachere Nutzung
 - erhöhte Verweilzeit auf bestimmten Bereichen . . .
- Segmentierung der Kundenbasis für gezielte Ansprache über unterschiedliche Kanäle
- Bestimmung neuer Marketing-Strategien
 - Ermittlung geeigneter Produkte für Sonderangebote / Bundling ...
 - Ermittlung der aussichtsreichsten Referrer-Knoten zur Werbung
 - Ermittlung der aussichtsreichsten Kunden- / Nutzergruppen
- optimierte Preisgestaltung
- automatische Bestimmung von Empfehlungen (Recommendations)

Beispiel für Kaufempfehlung

amazon.de

BUCH-INFO

Mehr zu diesem Buch

Überblick

[Amazon.de-Redaktion](#)

Mehr von ...

[Erhard Rahm](#),
[Gottfried Vossen](#)

Kunden kaufen auch

[diese Produkte](#)

Was meinen Sie?

[Ihre Meinung zu diesem Buch](#)

[Empfehlen Sie das Buch per E-Mail weiter](#)

Web & Datenbanken.

Konzepte, Architekturen, Anwendungen.

von [Erhard Rahm](#), [Gottfried Vossen](#)



Preis: EUR 54,00

Neu ab EUR 54,00

Gebraucht ab EUR 40,00

Versandfertig in 3 Tagen.

[Größeres Bild](#)

Kunden, die dieses Buch gekauft haben, haben auch diese Bücher gekauft:

- [XML & Datenbanken. Konzepte, Sprachen und Systeme](#) von Meike Klettke, Holger Meyer
- [Datenbanksysteme. Konzepte und Techniken der Implementierung](#) von Theo Härder, Erhard Rahm
- [XML und Datenbanken. Konzepte und Systeme](#) von Harald Schöning
- [Datenbanken und XML. Konzepte, Anwendungen, Systeme \(Xpert.press\)](#) von Wassilios Kazakos, Il. Il.

► [Entdecken Sie verwandte Produkte](#)

Kunden, die Bücher von Erhard Rahm gekauft haben, haben auch Bücher dieser Autoren gekauft:

- [Meike Klettke](#)
- [Andreas Bauer](#)

Personalisierungsbeispiel (FinanceScout24)



Berufsunfähigkeitsversicherung - ein unterschätztes Risiko

Die private Absicherung des Risikos der Berufsunfähigkeit zählt zu den wichtigsten Versicherungen jedes Arbeitnehmers.

[Informieren, vergleichen und kostenlos online abschließen!](#)

— Welche Versicherungen brauchen Sie wirklich? —

Welche Versicherung brauchen Sie wirklich?

Sie sind unverwechselbar. Deshalb brauchen Sie nicht nur die günstigste Versicherung, sondern auch eine, die individuell zu Ihnen passt. Geben Sie hier einfach Ihre Daten ein und Sie sind bei der Auswahl Ihrer Versicherung auf der sicheren Seite.

Alter	▼
Berufsstatus	▼
Familienstand	▼
Kinder	▼
Einkommen	▼



Persönliche Beratung.

Unsicher? Dann lassen Sie sich von einem neutralen Experten beraten!

[Ich will Beratung!](#)

[Kfz](#)
[Berufsunfähigkeit](#)
[Risiko-Leben](#)
[Kapital-Leben](#)
[Rente](#)
[Kranken Privat](#)
[Kranken Gesetzlich](#)
[Hausrat](#)
[Motorrad](#)
[Wohngebäude](#)
[Private Haftpflicht](#)
[Tierhalterhaftpflicht](#)
[Rechtsschutz](#)
[Unfall](#)
[Fahrrad](#)

[Kündigungsfristen.](#)

ios

Offene Probleme dynamischer Empfehlungen

■ effektive Kombination mehrerer Kriterien

- Nutzerinteressen vs. Anbieterinteressen
- Top-Hits der Vergangenheit vs. neue Produkte/Angebote/Inhalte

■ Robustheit dynamischer Empfehlungen

- Selbstverstärkungseffekte
- Manipulierbarkeit

■ Manipulationsmöglichkeiten: Beispiel 1

- aus beobachtetem Nutzerverhalten abgeleitete Empfehlungen können durch fingiertes Verhalten manipuliert werden
- Beispiel (www.wohnfinder.de): Es werden die Objekte vorgeschlagen, die in den letzten 30 Tagen die meisten Zugriffe auf ihr Exposé hatten.



■ Manipulationsmöglichkeiten: Beispiel 2

- www.amazon.de: Nutzerrezensionen und Nutzervoting, ob Rezension hilfreich war
- *Gute* Rezensenten werden hervorgehoben; Votes aller Nutzer werden jedoch gleich behandelt



Zusammenfassung

■ Webzugriffsanalyse nimmt zentrale Rolle im E-CRM ein

■ Ziele der Web-Präsenz bestimmen zu bewertende Faktoren

- Gewinnung neuer Kunden, Erhöhung der Kundenbindung
- Kostenreduzierung durch Online-Dienste ...

■ Data-Warehouse-Einsatz

- Skalierbarkeit auf große Datenmengen
- Flexible Kombination von Webzugriffsdaten mit weiteren Datenquellen
- Kundenzuordnung sowie inhaltlicher / fachlicher Bezug
- Business-orientierte Bewertungen
- Umfassende und flexible Analysemöglichkeiten (OLAP / Data Mining)

■ Großteil der Arbeit liegt in der Datentransformation und -integration

■ Vielfältige Nutzungsmöglichkeiten, v.a. auch für CRM:

- optimierte Website-Gestaltung, Marketing-Aktivitäten, Personalisierung

■ Ziel: Closed-Loop-Ansatz mit automatisierter Analyse und Reaktion (z.B. dynamische Empfehlungen)