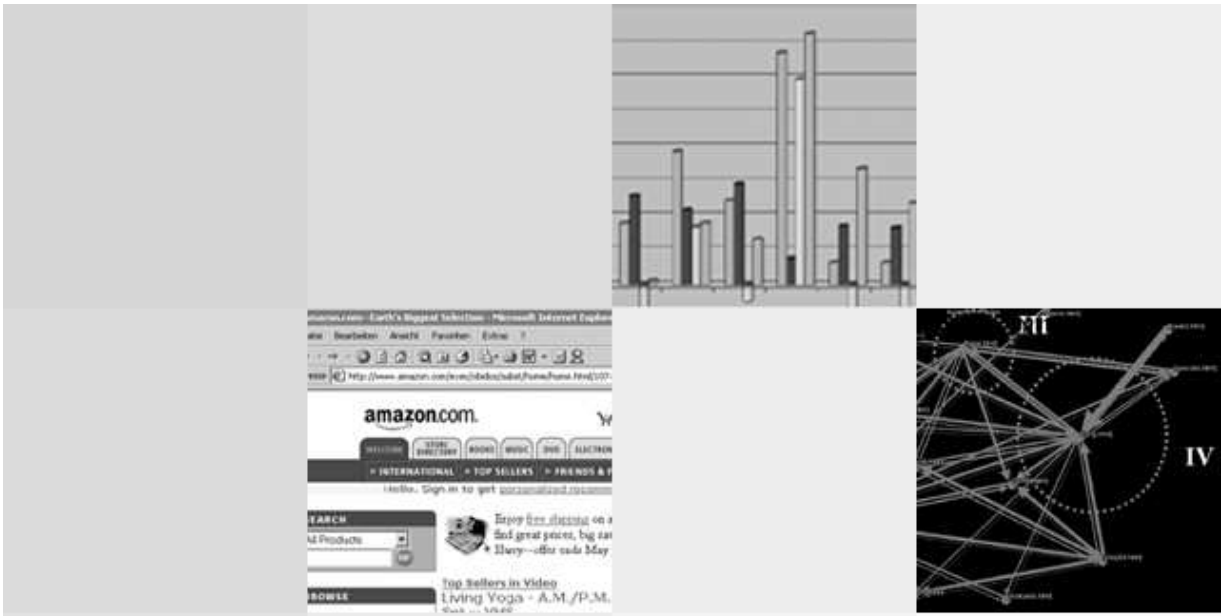


Web Log Mining

21.05.2001 - Martin Klossek

Seminar „WWW und Datenbanken“ - SS2001

Web Log Mining

- Im World Wide Web werden täglich unzählbar viele Dateien von Servern an Arbeitsplatzrechner geliefert...

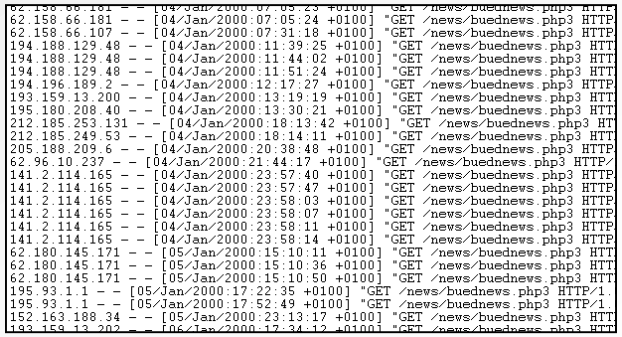
Anfragen

Webserver

Dateien

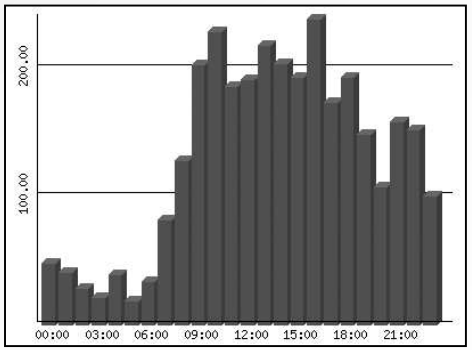
Web Log Mining

- Die Webserversoftware protokolliert dabei jede gelieferte Datei wie
  - statische + dynamische Htmlseiten
  - Bilder (.png, .gif, .jpg, ...)
  - Java-Applets, Flash, ...
- Speicherung der Zugriffe in **Logdatei**



Web Log Mining

- Erzeugte Logfiles werden sehr groß!
- Auswertung und Visualisierung
- Informationen wie
  - **# Zugriffe pro Monat / Woche / Tag**
  - **# Zugriffe pro Datei / Seite**
  - **Zugreifende Rechner**
  - **Browsertypen**
  - **...**



Web Log Mining

eCommerce

Web Log Mining

- Komplexe Sites wie Portale und Online-Shops verlangen Informationen über ihre Besucher
- Können Logfiles mehr als nur die bloßen Verbindungsdaten lieferen?

Ja! ...mit Web Log Mining

Web Log Mining

- Web Log Mining beantwortet Fragen von Websitebetreibern
  - Wer besucht meine Website? Wer kauft dort ein (bei Online-Shops!)?
  - Welche Seiten werden in einer Session zusammen besucht?
  - Welche Seiten werden nacheinander besucht (Sequenz)?

Web Log Mining

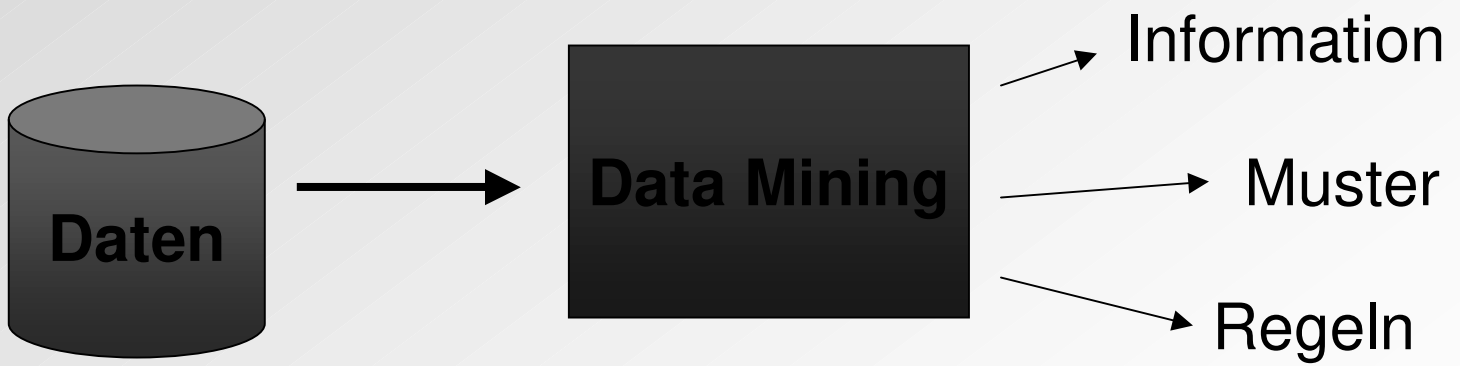
- oder auch folgende Fragen
  - Welche Werbemaßnahmen, welche Banner sollte ich einsetzen?
  - Wie unterscheiden sich Käufer von Nicht-Käufern (eCommerce)?
  - Verhalten sich registrierte Benutzer anders als anonyme?
  - Wie erhöhe ich die Anzahl meiner Besucher und Kunden?

Übersicht

1. Motivation
2. Begriffsabgrenzung DataMining
3. Domäne Web Log Mining
4. Prozess des Web Log Mining unterstützt durch Fallbeispiele
5. Ausblick

Data Mining

- Data Mining is
  - data and discovery driven
  - **not:** confirmation or verification driven



Data Mining

- Data Mining Software analysiert Daten und liefert Regeln und Muster
  - Mustererkennung
  - Hypothesenfindung
- Statistiksoftware bietet
  - Mittel zum Bestätigen von Hypothesen, die Sachverständige aufgestellt haben

Data Mining

- Einige Verfahren sind
  - Assoziationsanalyse
  - Sequenzanalyse
  - Clusteranalyse
  - Entscheidungsbäume
- Allgemein sind Methoden des „**Maschinellen Lernens**“ nutzbar (beispielsweise Neuronale Netze)

Data Mining

- Anwendungen von Data Mining haben spezifischen Kontext, der mit **Domänenwissen** bezeichnet wird
- Beachte: Data Mining und hier im speziellen Web Log Mining sind nicht standardisiert, sondern anwendungs- und fragenabhängig...

Übersicht

1. Motivation
2. Begriffsabgrenzung DataMining
3. Domäne Web Log Mining
4. Prozess des Web Log Mining unterstützt durch Fallbeispiele
5. Ausblick

Webserver Architektur

Anfragen

- Requestobjekt
- Client-Infos
- Cookies

Webserver

Protokoll
Logfiles

Dateien

Skripte

Datenbank

Daten

- Dateien
- Status
- Cookies

Protokollierung

Protokoll
Logfiles

- Verschiedene Webserver am Markt
  - Apache
  - Internet Information Server
  - Netscape iPlanet
  - NCSA
  - CERN
  - ...
- Aber: **Ein Format** für Logdateien

`http://www.netcraft.com`

Format Logfiles

Protokoll
Logfiles

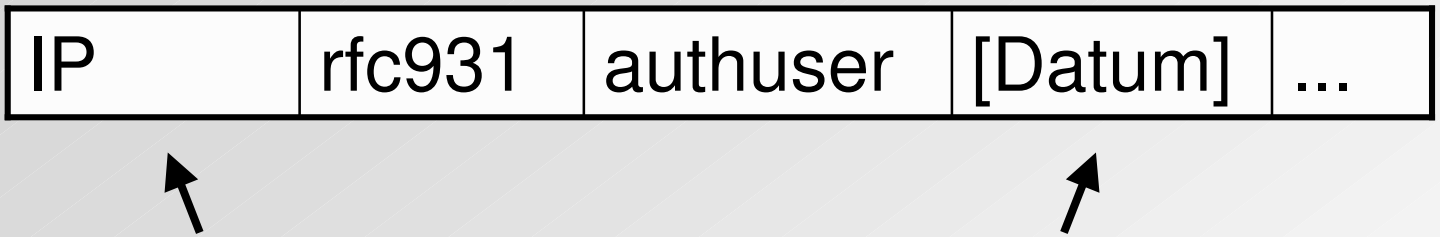
- **Common Log Format**
  - ascii-Datei
  - jede Zeile entspricht einer Anfrage an den Webserver
  - verschiedene Felder beschreiben die Anfrage

IP	rfc931	authuser	Datum	requeststring	...

Format Logfiles

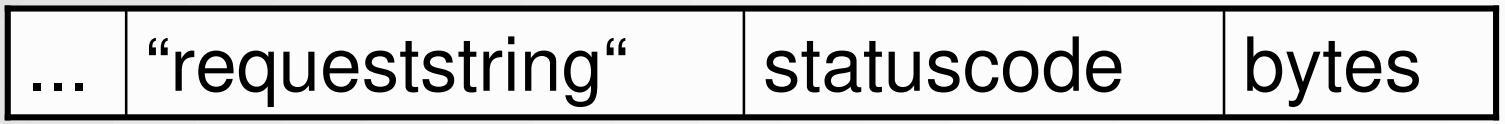
Protokoll
Logfiles

- **Felder sind...**



z. B. 141.2.114.129

z. B. [21/May/2001:17:02:22 +0100]



z. B. "GET /start.php3 HTTP/1.1"

z. B. 202, 304 oder 404...

Format Logfiles

- Im **Extended Log Format** erweitert um die Felder
  - **Referrer**
    - Die im Browser zuvor angezeigte URL, z. B. "http://www.stormzone.de/uni/future.html"
  - **user_agent**
    - Der Browserstring des Clients, z. B. "Mozilla/4.0 (compatible; MSIE 5.01; Windows 98)"

Transaktionen

- Jede Zeile im Logfile stellt eine Anfrage an den Webserver dar
- Eine Anfrage besteht dabei aus
  - Html-Datei / Skript
  - Grafikdateien, Stylesheets, Applets, ...
- Eine **Transaktion** ist der Abruf einer Folge von Seiten...

Session

- Menge von Anfragen eines Besuchers in einer begrenzten Zeitspanne wird als **Session** bezeichnet
- Charakterisiert durch
  - IP
  - Uhrzeit

$$Session_{BenutzerA,2001\text{-}05\text{-}21} = \{ index.html, seite1.html, seite2.html, ... \}$$

Session

- **Problem:** Datensätze nicht immer eindeutig, daher zur Unterscheidung:
  - Begrenzung der Session auf ca. 25 Minuten oder Inaktivitätsspanne
  - Identifizierung über Sessionkeys in URL oder Cookies, beispielsweise

https://ssl.moneyshelf.com/DE/de/functions/Cash/cashAllAccounts.jhtml;$sessionid$P4AB000FXLOPKCQCECCSFFIKAIQIIIV0

Weitere Daten

- Für weitere Analysen ist Anreicherung der Daten sinnvoll, beispielsweise
  - Benutzerdaten aus Formularen
  - Demographische Datenbanken

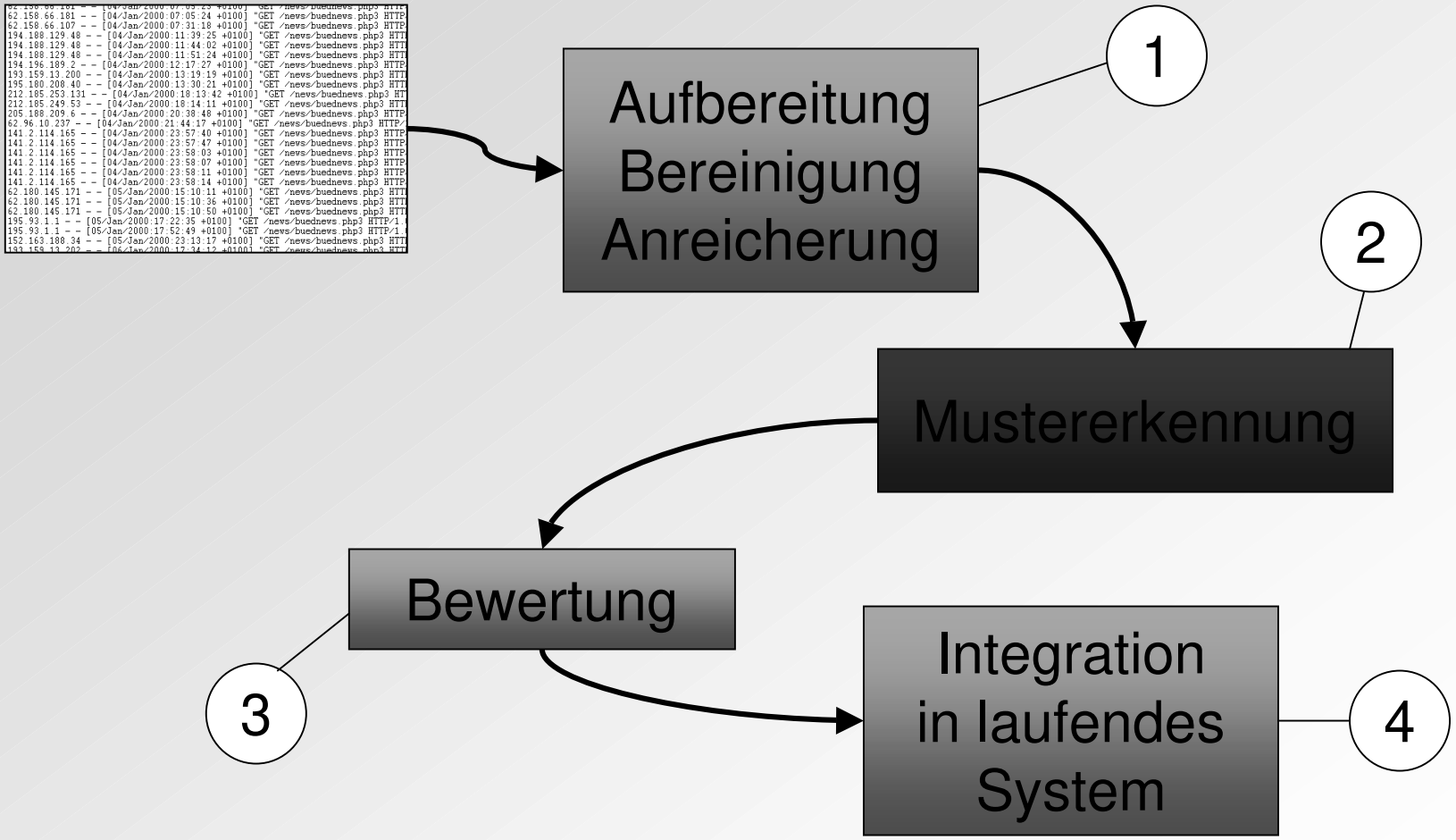
Übersicht

1. Motivation
2. Begriffsabgrenzung DataMining
3. Domäne Web Log Mining
4. Prozess des Web Log Mining unterstützt durch Fallbeispiele
5. Ausblick

Prozess Web Log Mining

- **Aufgabenstellung:**
  - Gegeben sind Logdaten einer Website
  - Ziel: Gewinnen neuer Informationen
- **Hier:** Einige Ergebnisse aus einem Web Log Mining Projekt
- **Allgemein:** Je nach Fragestellungen variieren die verwendeten Analysen

Prozess Web Log Mining



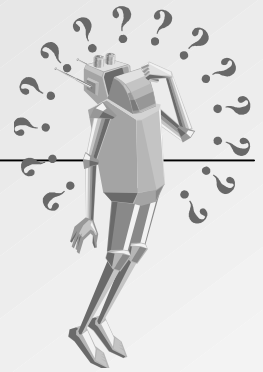
Datenaufbereitung

- Übertragen der Daten aus Logfiles in Format, das für die eingesetzte Data Mining Software lesbar ist
  - Auswahl der relevanten Felder im Logfile
  - Speicherung in Datenbank
  - Extraktion von Transaktionen und Sessions

Datenaufbereitung

- Entfernen von **redundanten** und **überflüssigen** Daten falls möglich
- Entdecken und Entfernen von **Ausreißern** - falls möglich
- Anreichern der Logdaten mit Daten aus **Personendatenbank** – falls vorhanden und den Fragestellungen entsprechend

Mustererkennung



- Zu Beginn wurden einige **Frage-stellungen** vorgestellt. Schauen wir uns einige näher an
  1. Welche Seiten wurden in einer Session zusammen besucht?
  2. Clickstreams der Besucher
  3. Gruppen von Besuchern ähnlichen Verhaltens?
  4. Unterschiede zwischen Käufern und Nichtkäufern?

Assoziationsanalyse

- Welche Seiten wurden in einer Session zusammen besucht?
  - **Assoziationsanalye**
  - Erzeugt Regeln der Form

Prämisse
„wenn A.html“

⇨

Konsequenz
„dann B.html“

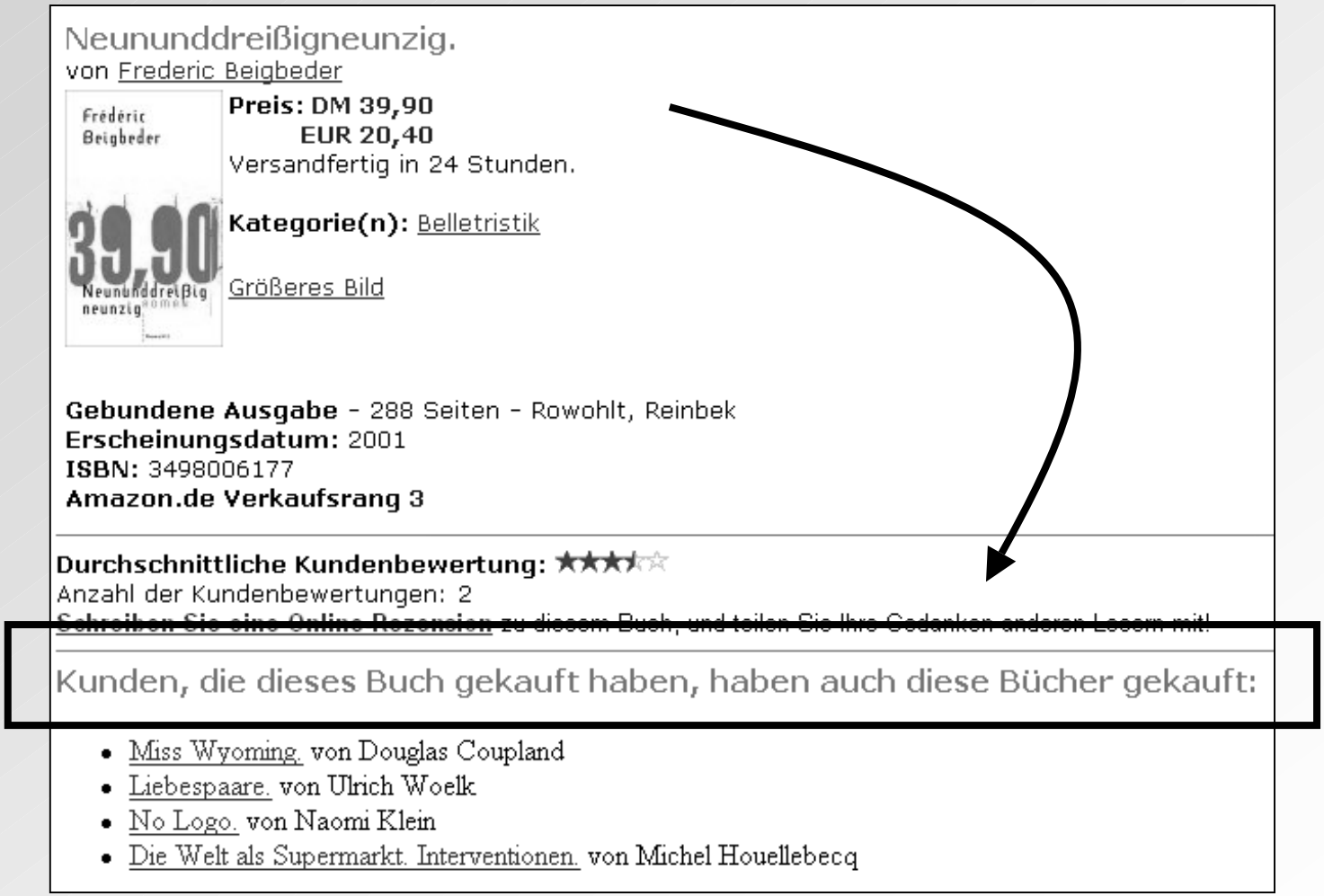
Mit Konfidenzfaktor `conf`
und Supportfaktor `sup`

A.html ⇨ B.html$^{0,9;\ 0,2}$

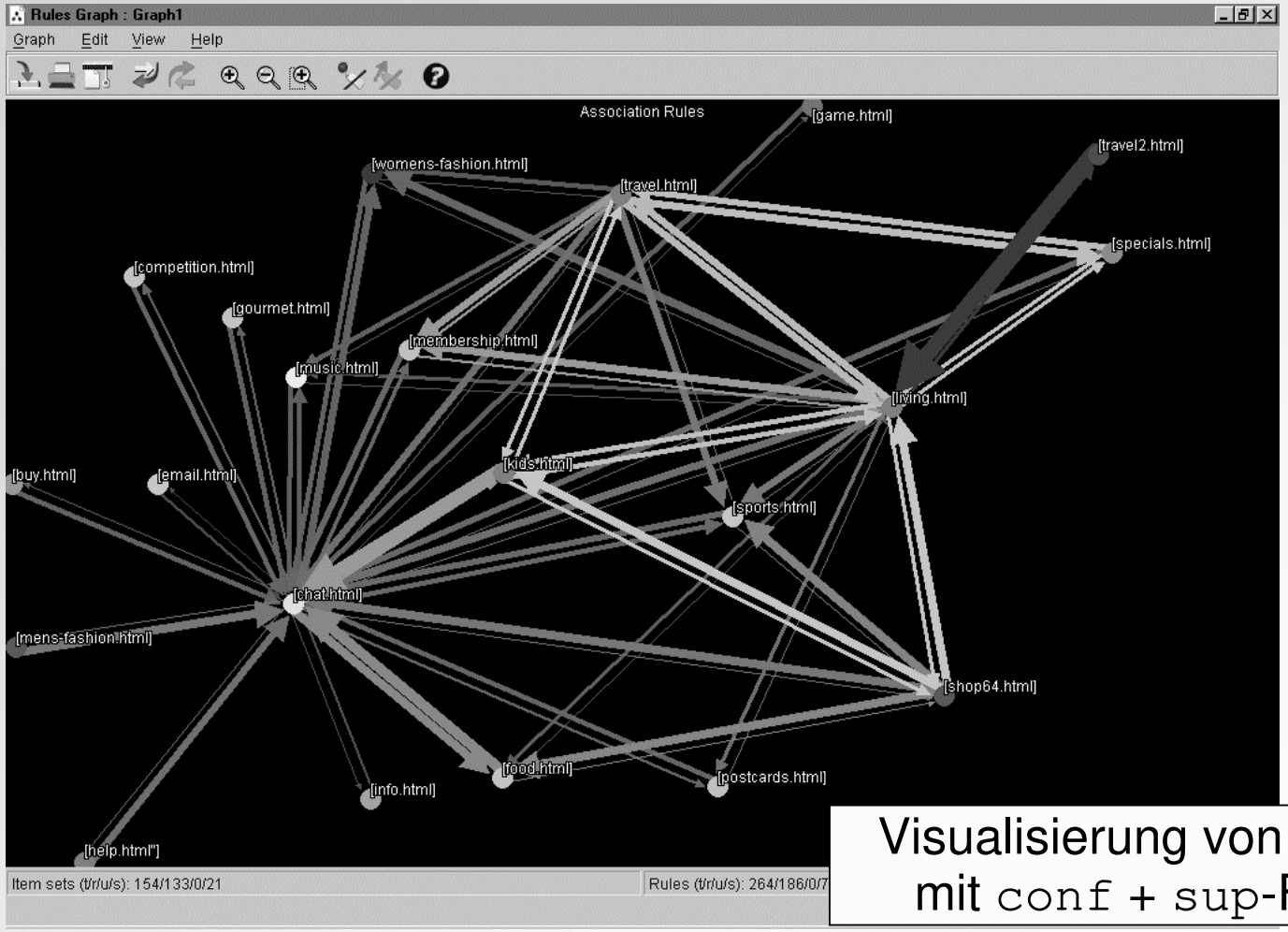
Assoziationsanalyse

- Die gefundenen Regeln geben Aufschluß darüber, wie sich die jeweiligen Seiten ergänzen
- Umgestaltung der Website denkbar, so dass von Seite A auf Seite B und umgekehrt verwiesen wird...
- ...je deutlicher Support- und Konfidenzfaktor für diese Regel

Assoziationsanalyse



Assoziationsanalyse



Visualisierung von Regeln mit `conf` + `sup`-Faktor

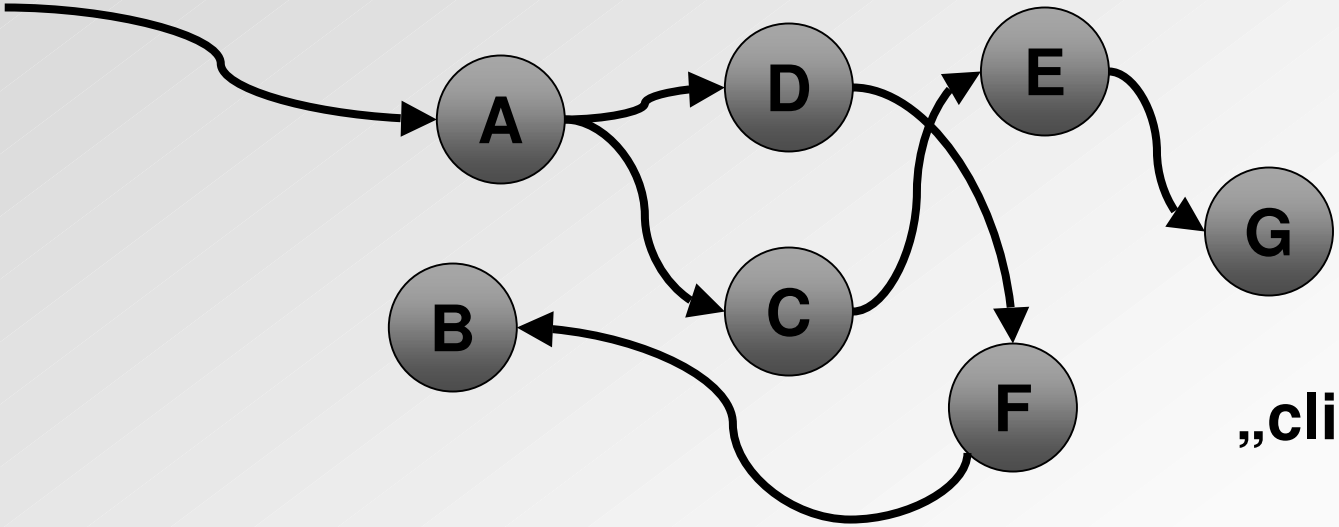
Sequenzanalyse

- **Clickstreams** der Besucher
  - Die Assoziationsanalye sagt nichts über die Reihenfolge der Seitenabrufe aus, dazu die:
  - **Sequenzanalyse** (hier Pfadanalyse)
  - Finden von Regeln für nacheinander besuchte Seiten mit Häufigkeit

A.html ⇨ D.html⇨ F.html⇨ B.html^{0,05}

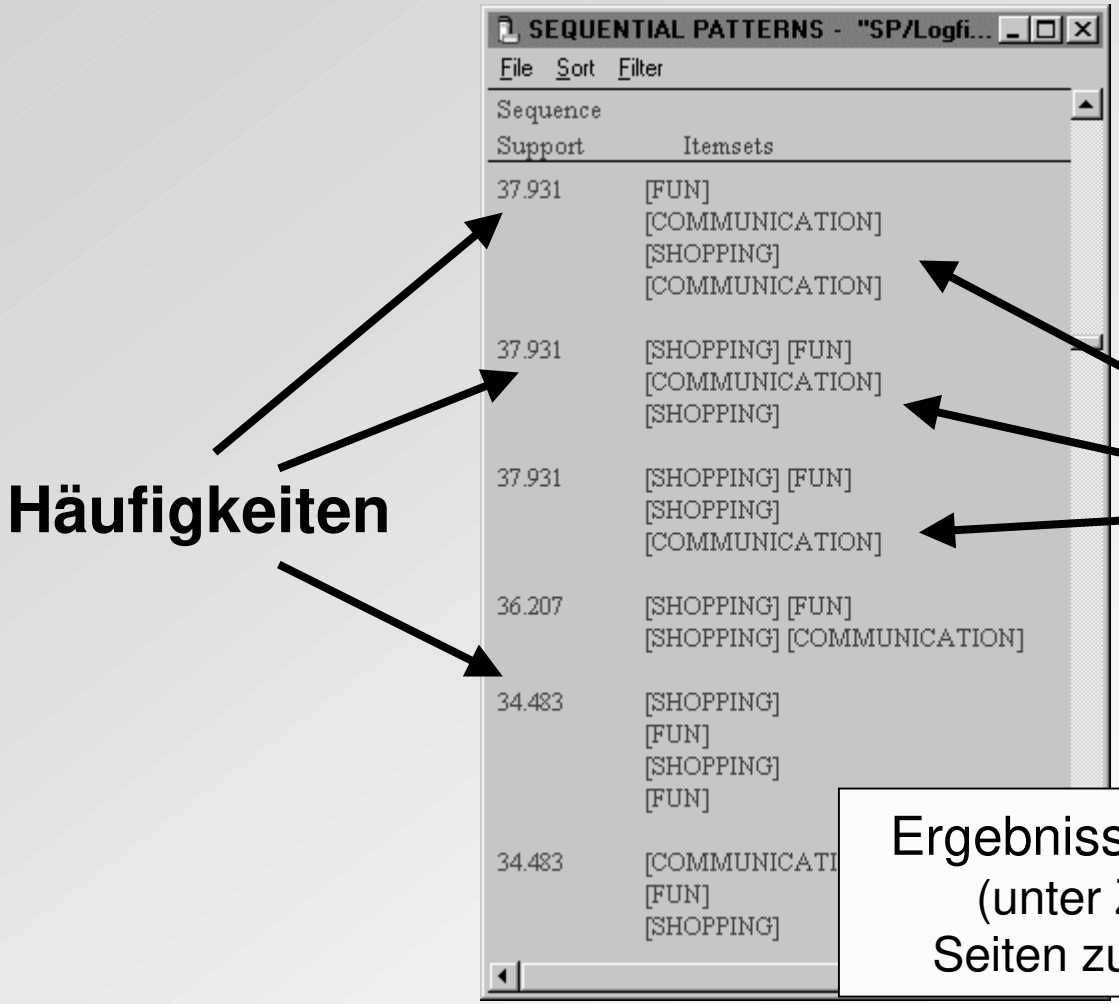
Sequenzanalyse

- Gewinnen von Erkenntnissen über die Reihenfolge der Informationssuche von Besuchern (⇨ Siteoptimierung!)



„clickstreams“

Sequenzanalyse



Häufigkeiten

„clickstreams“

Ergebnisse einer Sequenzanalyse
(unter Zusammenfassung von
Seiten zu semantischen Gruppen)

Clusteranalyse

- Suche nach Gruppen von Besuchern mit ähnlichen Verhalten
  - **Clusteranalyse**
  - Bilden von Clustern mit
    - homogenen Merkmalen im Cluster
    - hoher Heterogenität zwischen den Clustern
  - Dabei werden die Merkmale aller Objekte miteinander über ein Proximitätsmaß verglichen (z. B. Euklidische Norm)

Clusteranalyse

- Reine Logdaten bieten nur Pfade, Verweildauer und technische Details (wie Browsertyp) als Merkmale an
- Sinnvoll daher bei **Logdaten von registrierten Benutzern** mit Personeninformationen
- Im folgenden ein Beispiel mit Anreicherung durch Personendaten

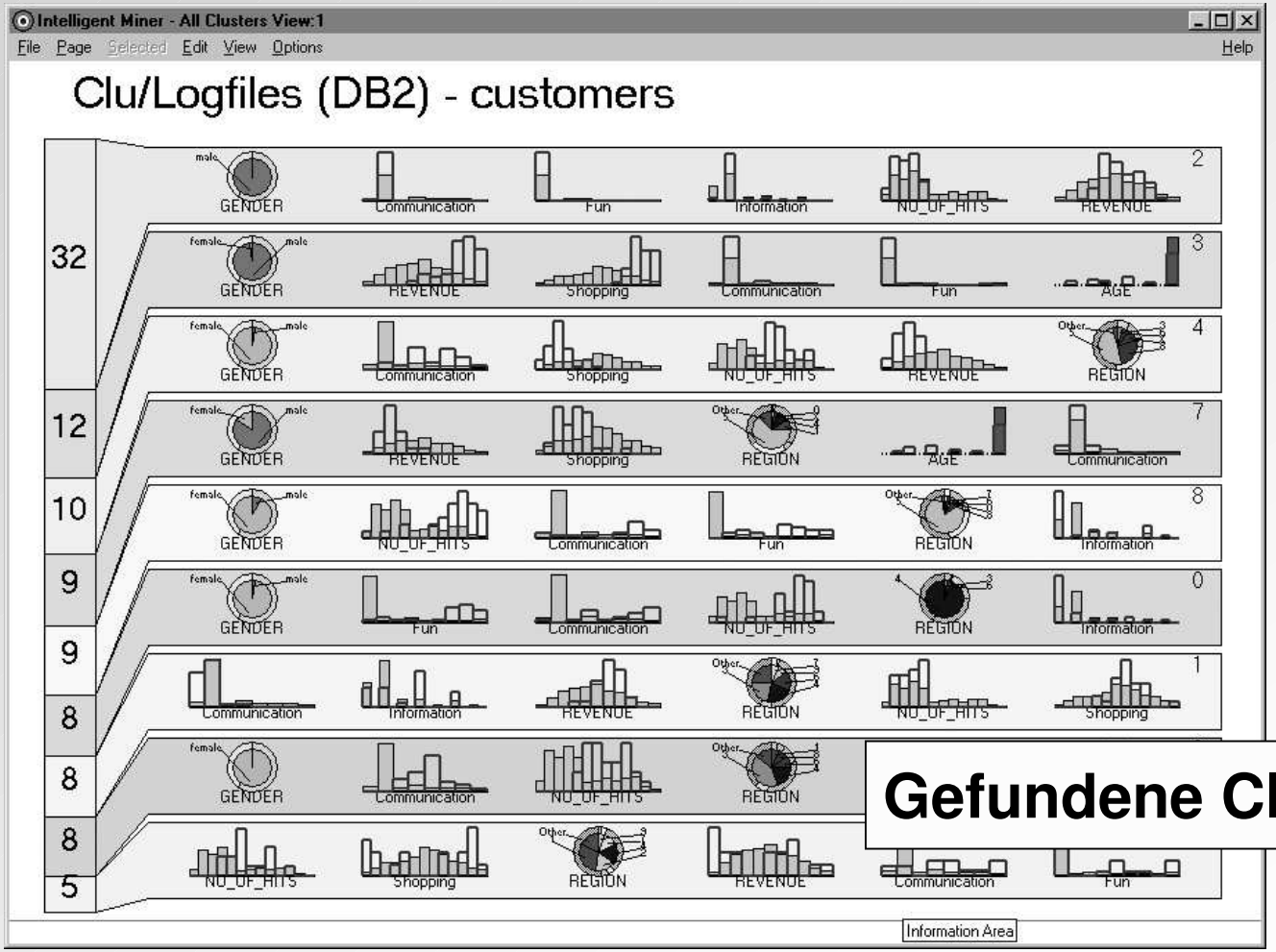
Clusteranalyse

Sample Contents - WEBM_CUSTOMER

)_OF_INDICA...	NO_OF_ARTIC...	NO_OF_HITS	GENDER	REVENUE	REGION	AGE
	26	258	male	455	4	38
	15	140	female	488	3	30
	20	111	male	1008	3	32
	14	82	male	934	1	26
	24	144	male	576	2	35
	26	25	male	2680	4	43
	13	192	female	200	6	27
	18	198	female	800	6	26
	12	211	female	200	5	22
	25	205	female	296	5	42
	12	50	male	506	5	20
	13	92	male	1640	4	27
	19	257	female	734	X	27
	27	171	female	1660	3	39
	30	81	male	2012	3	51
	43	32	male	2882	3	69
	35	24	male	1664	4	58
	21	107	female	598	5	29

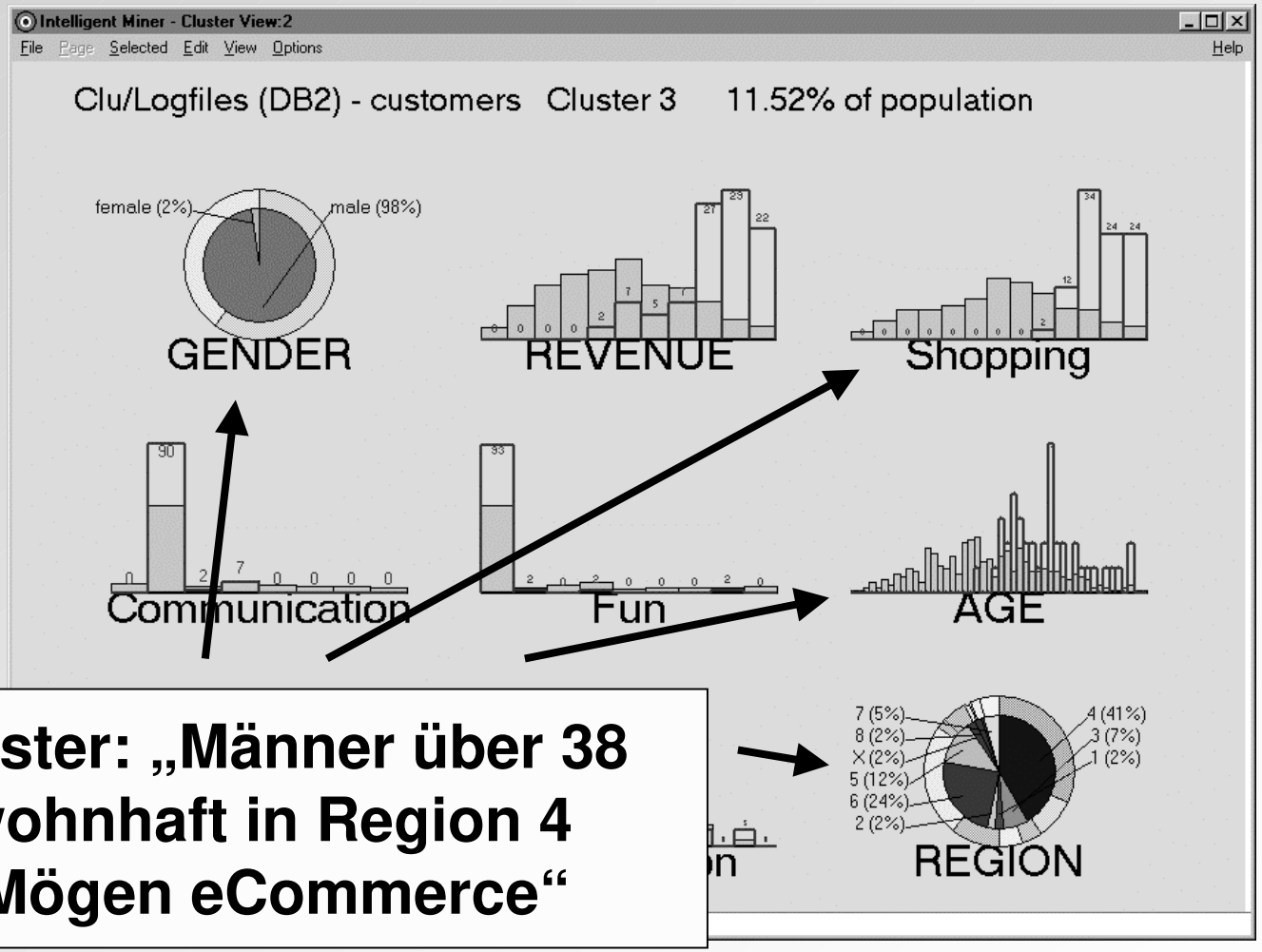
Datenbank mit Personendaten

Clusteranalyse



Gefundene Cluster

Clusteranalyse



Cluster: „Männer über 38 wohnhaft in Region 4 Mögen eCommerce“

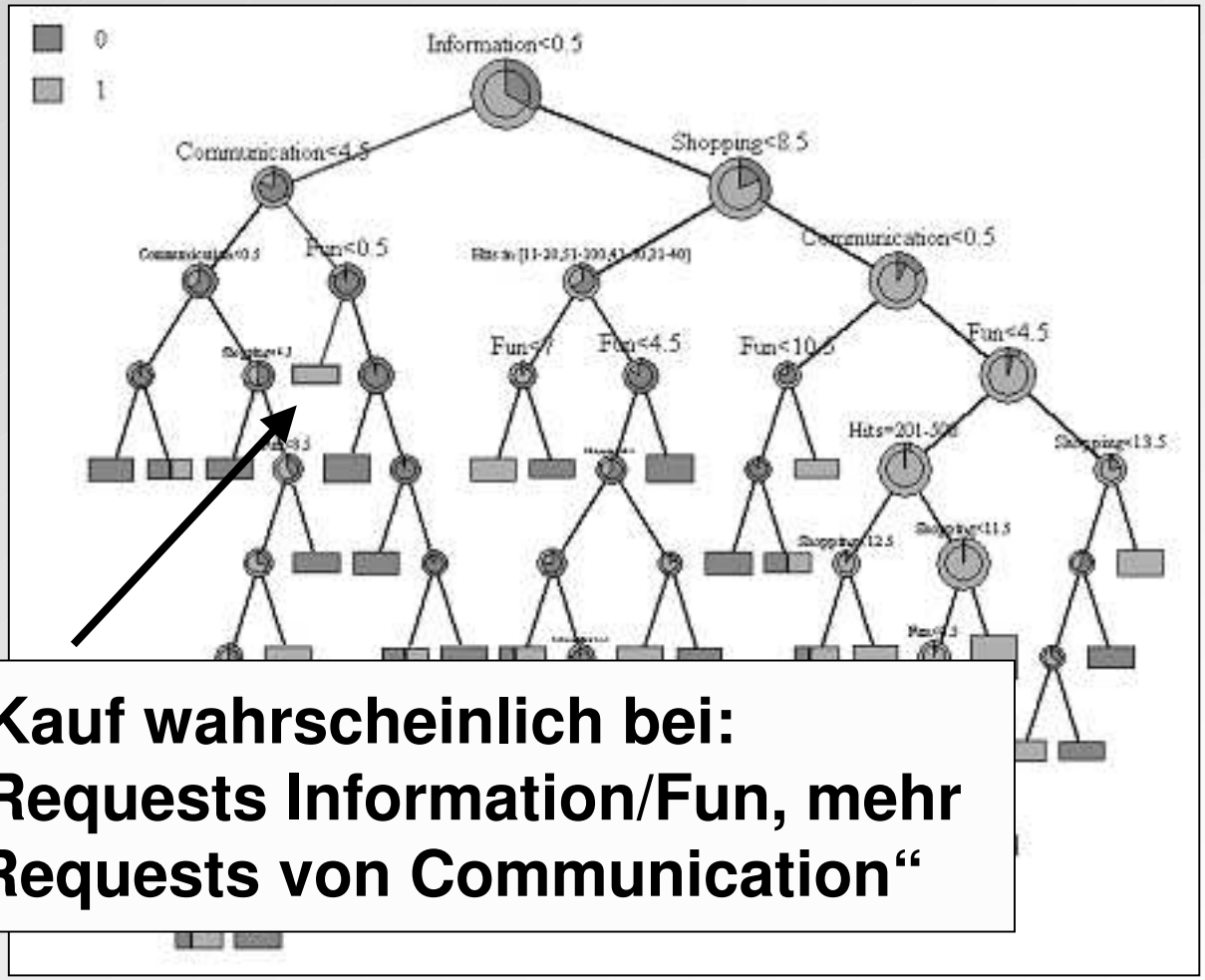
Clusteranalyse

- Mit Hilfe der Clusteranalyse und angereicherten Logdaten lassen sich also Angaben über die Art der Besucher der Website machen
- **Denkbar:** „Automatische Personalisierung“, um Besuchern des gleichen Clusters ähnliche Informationen anzubieten!

Entscheidungsbäume

- **Online-Shop:** Unterscheidung zwischen Käufern und Nichtkäufern?
  - **Entscheidungsbäume**
  - Hier Segmentierung der um Personen-daten angereicherten Logdaten, um Käufer von Nicht-Käufern zu unterscheiden
  - Ableiten von Regeln zur Vorhersage des Kaufverhaltens von Besuchern

Entscheidungsbäume



Kauf wahrscheinlich bei:
„Keine Requests Information/Fun, mehr als 5 Requests von Communication“

Übersicht

1. Motivation
2. Begriffsabgrenzung DataMining
3. Domäne Web Log Mining
4. Prozess des Web Log Mining unterstützt durch Fallbeispiele
5. Ausblick

Ausblick

- Beobachtungen und Vermutungen
  - Für große Sites ist Web Log Mining **unerläßlich**, um
    - **Bannerwerbung** optimal zu platzieren
    - **Personalisierung** zu ermöglichen und so vermutlich den Umsatz zu steigern (Kundenbindung contra Kundenaquise!)
    - Streuverluste durch **one2one-Marketing** zu verringern

Ausblick

- Interessante Möglichkeiten
  - **Standardisierung** von Web Log Mining Verfahren
  - **Integration** in Serversoftware
    - siehe Microsoft Commerce Server

Microsoft Commerce Server

Ausblick

- sowie zu überlegen...
  - Verbindung von Web Content und Web Log Mining?
  - Optimierung und Weiterentwicklung von **Analyseverfahren**
  - Aber auch: Datenschutzrechtliche Begrenzungen!

Fazit

„Web Log Mining“ ist kein Hypethema sondern ein aus wirtschaftlichen Erfordernissen enstandenes Verfahren zur Analyse von Verbindungsdaten, das Anwendung findet.

Kontakt

- Kontakt:
  - martin@klossek3000.de
- Folien und Ausarbeitung in verschiedenen Formaten unter
  - http://www.stormzone.de/uni/Hauptstudium/seminare/wwwdb/list.php3

Das war's