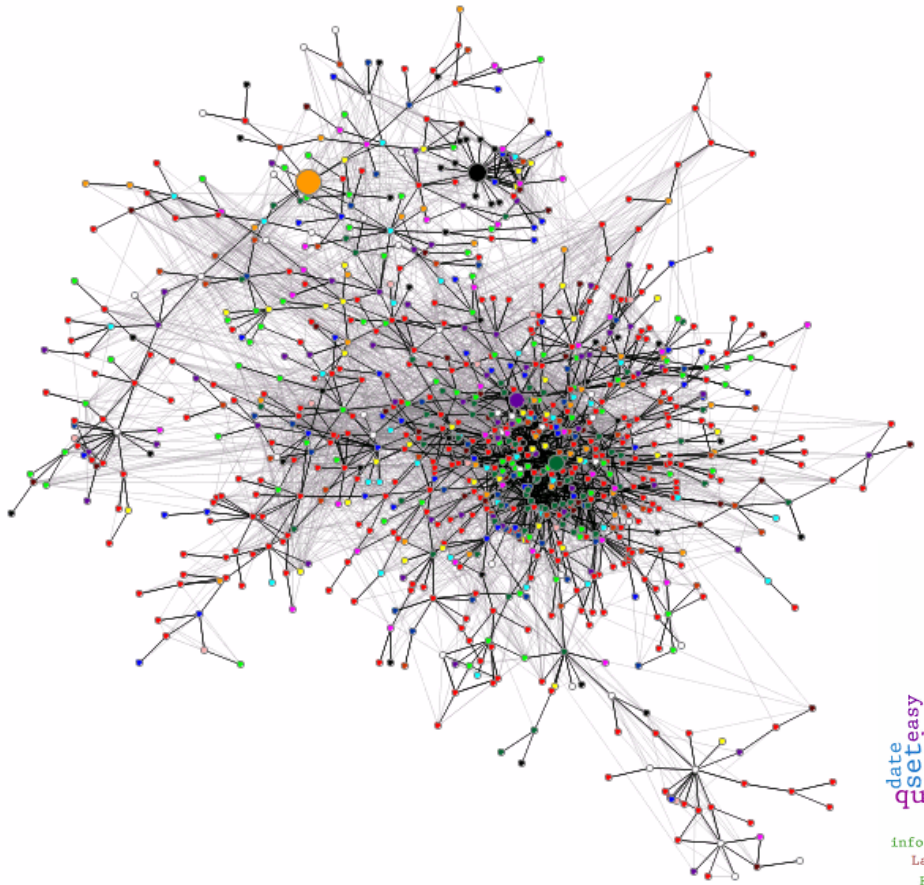
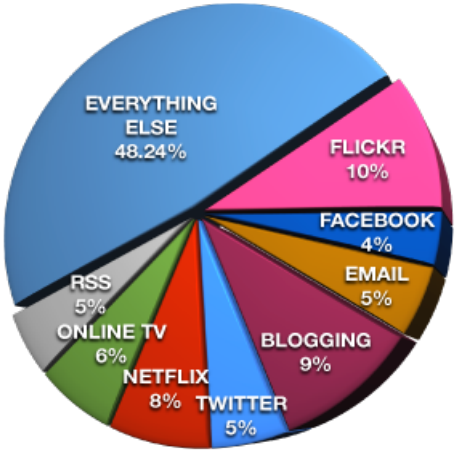
Web Data Mining





Oberseminar Data Mining SS 2010
Martin Vogel und Patrice Khuori

Agenda

1. Motivation und Einführung
2. Web Content Mining
3. Web Structure Mining
4. Web Usage Mining
   - Anwendungsgebiete
   - Datenquellen
   - KDD Prozess
   - Schnittstellen
   - Probleme
5. Tools
6. Ausblick

Literatur

1 Motivation

- **Datenquellen im heutigen Internet:**
  - große, stark wachsende Menge an Daten
  - heterogene Informationen
  - starke Verlinkung von Inhalten
  - Dynamik (Communities, Social Media, etc.)

- **digitales Datenaufkommen 2007:**

  2.25×10^{21} Bit (281 Exabyte oder 281 Milliarden Gigabyte !)

- **Anzahl Domains (.com, .net, .org, .biz., .info):**

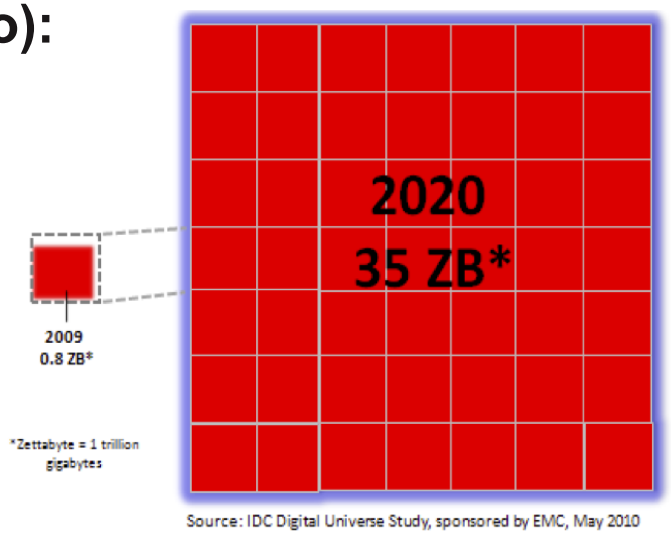
  119.003.182

- **Indexierte URLs in Google 2008**

  1.000.000.000.000

- **Internet User 2010**

  1.802.330.457



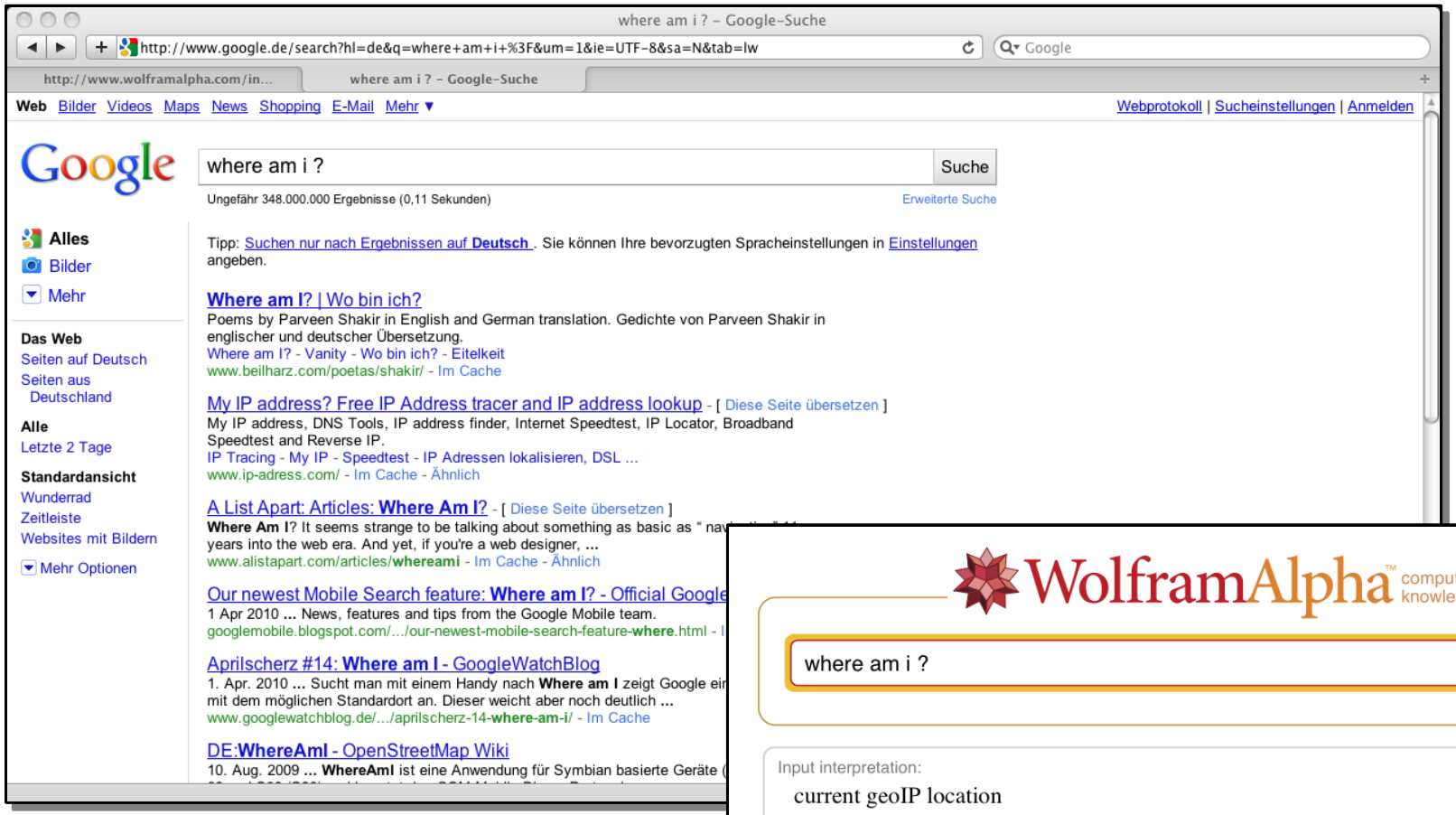
[IDC]

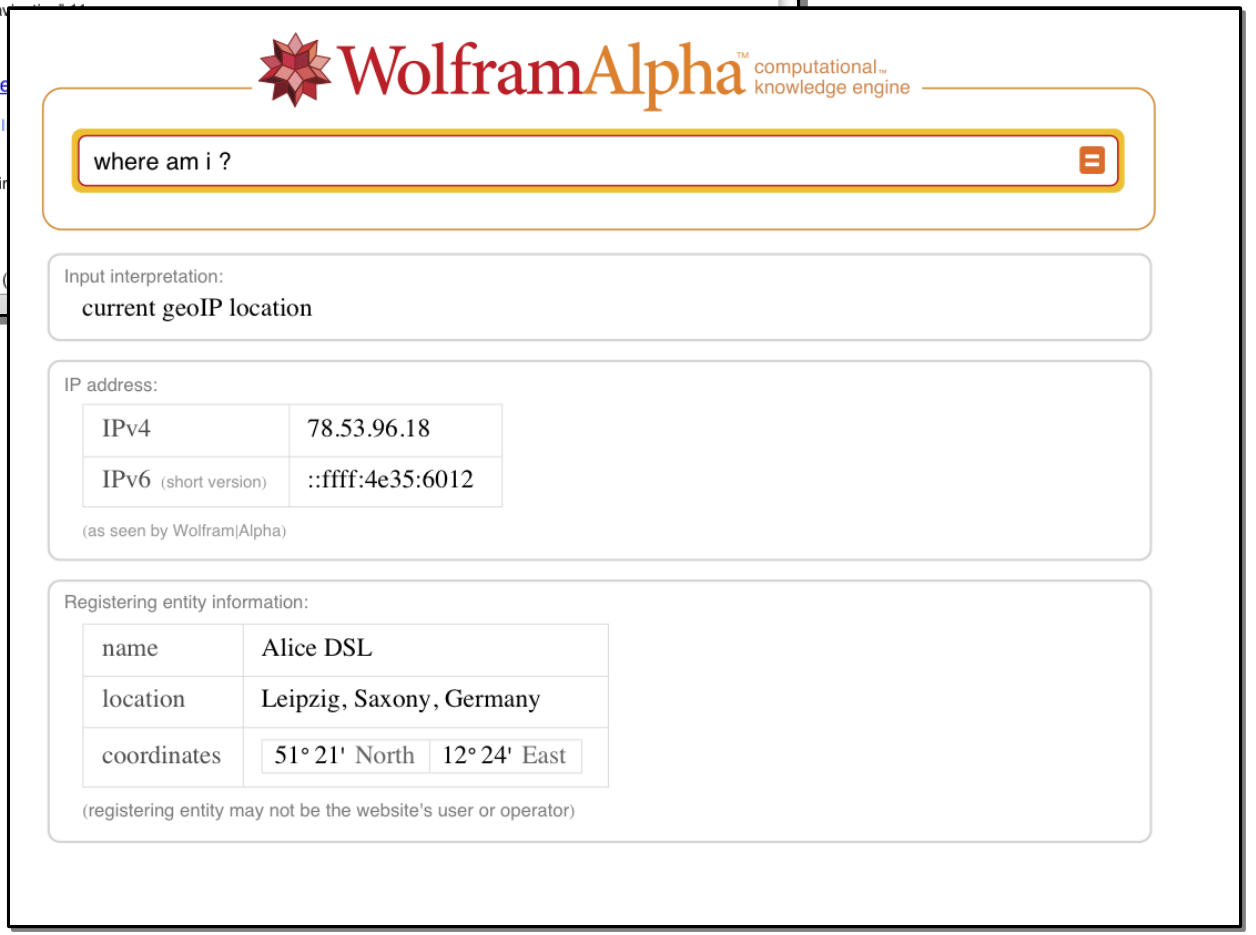
1 Motivation

- **Entscheidene Frage:**
  - Wie können die gesuchten Informationen gefunden werden ?
  - Wie nutzt die „Digitale Gesellschaft“ diese Informationen

- **Zugriff auf Informationen über Suchmaschinen**
  - Technik so alt wie das Web
  - Nutzung von Keywords
  - Problem: keine/kaum Semantik
  - (Wenig Zugriff auf das “Deep Web”))

- **Herausforderungen**
  - Relevante Seiten mit Informationen finden
  - Extraktion der relevanten Teilinformationen
  - Kontext Schlussfolgern – Querverweise herstellen





348.000.000 Ergebnisse

vs.

1 Ergebnis

WolframAlpha computational knowledge engine

german population 1900 vs. german population 2010

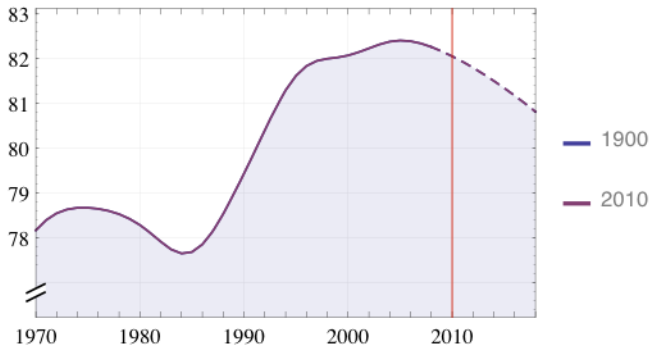
Input interpretation:

Germany	population	1900		Germany	population	2010

Results:

1900	56.4 million people
2010	82.1 million people (2008 projection)

Recent history:



(from 1970 to 2018) (in millions of people)

WolframAlpha computational knowledge engine

green + blue

Input interpretation:

color blend	green
	blue

Result:

color | red 0 | green 0.5 | blue 0.5

Color swatch:

Representations:

fractions	red 0 | green 0.5 | blue 0.5
	0 | green 128 | blue 128
	180° | saturation 100% | brightness 50%
	08080

WolframAlpha computational knowledge engine

whats the meaning of life ?

Input interpretation:

Answer to the Ultimate Question of Life, the Universe, and Everything

Result:

42

(according to Douglas Adams' humorous science- fiction novel *The Hitchhiker's Guide to the Galaxy*)

1.1 Forschungsgegenstand

Web Mining == Data Mining ?

Wikipedia:

"Web Mining [ist die] **Übertragung von Techniken des Data Mining zur** [...] **Extraktion von Informationen aus dem Internet** [...]. Web Mining übernimmt **Verfahren und Methoden aus** den Bereichen **Information-Retrieval, Maschinelles Lernen, Statistik, Mustererkennung und Data Mining.**"

Liu, Bing in [LIU, S.6]*:*

"Web mining aims to **discover useful information and knowledge from the Web** hyperlink structure, page contents, and usage data."

1.2 Web Mining

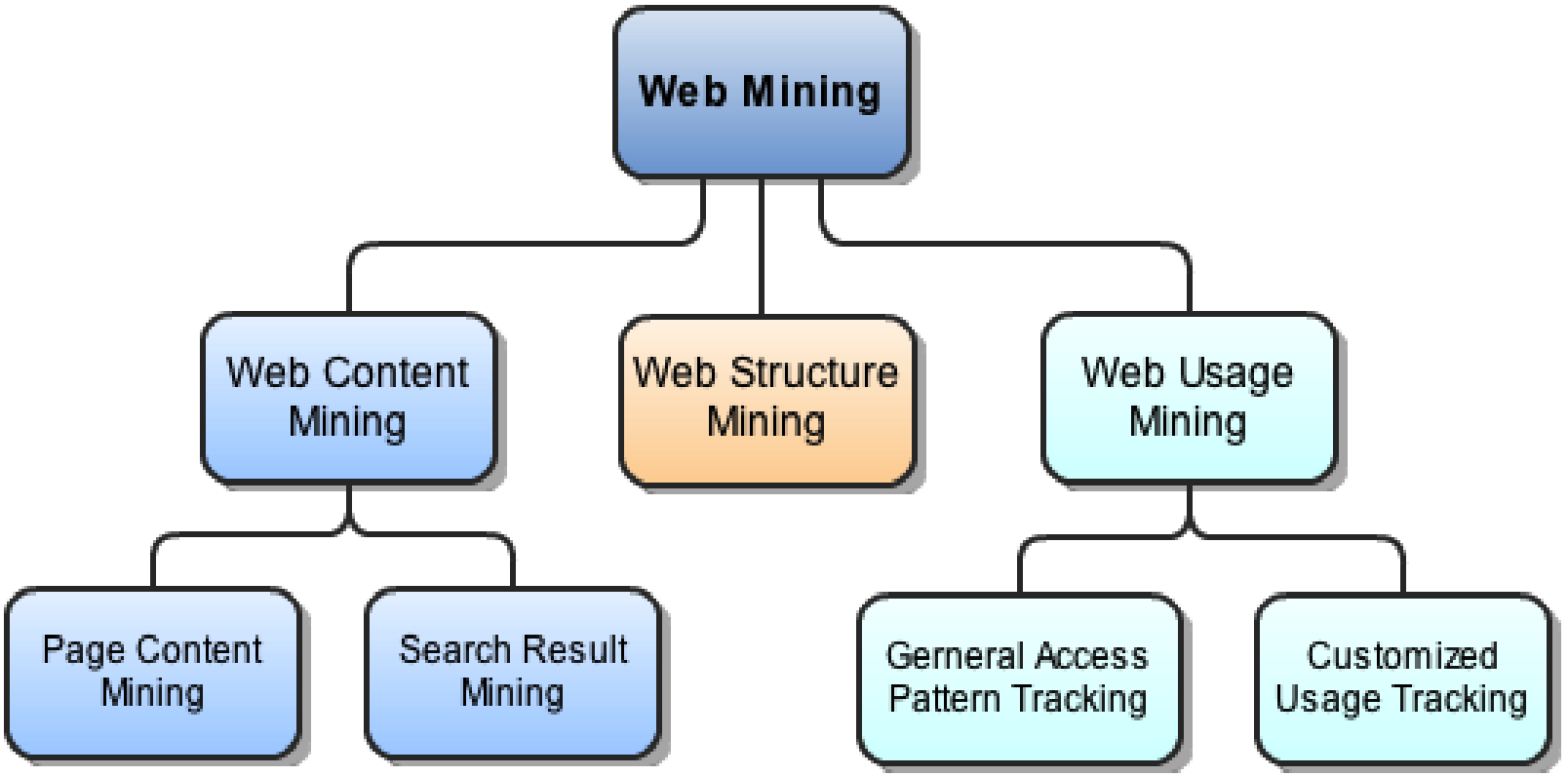
- Nutzung von Data Mining Methoden
- keine “rein klassische” Anwendung des Data Mining

Data Mining	**Web Mining**
strukturiert	Semi-strukturiert, unstrukturiert
relational	Links
definierte Tabellenstruktur	Spontane Änderung und Evolution

⟶ Entwicklung neuer Verfahren und Algorithmen

1.2 Web Mining

- **Gewinnung von Informationen aus:**
  - Inhalten von Websites
  - Struktur durch Hyperlinks
  - Nutzungsdaten



nach [WBT]

2 Web Content Mining

Extraktion von Wissen aus Website-Inhalten

Was ist Web Content Mining?

- Web Content Mining bezeichnet Untersuchungsansätze dem User eine strukturierte Übersicht bestehender Webseiten zur Verfügung zu stellen
- befasst sich mit der Analyse des Inhaltes von Webseiten
- Ziel ist es die Suche nach Informationen im Netz zu erleichtern
- Klassifizierung und Gruppierung von Online-Dokumenten oder Auffinden von Dokumenten nach bestimmten Suchbegriffen
- Text-Mining Verfahren kommen zum Einsatz
- Agent-basierter Ansatz
- Datenbank-basierter Ansatz

Agent-basierter Ansatz

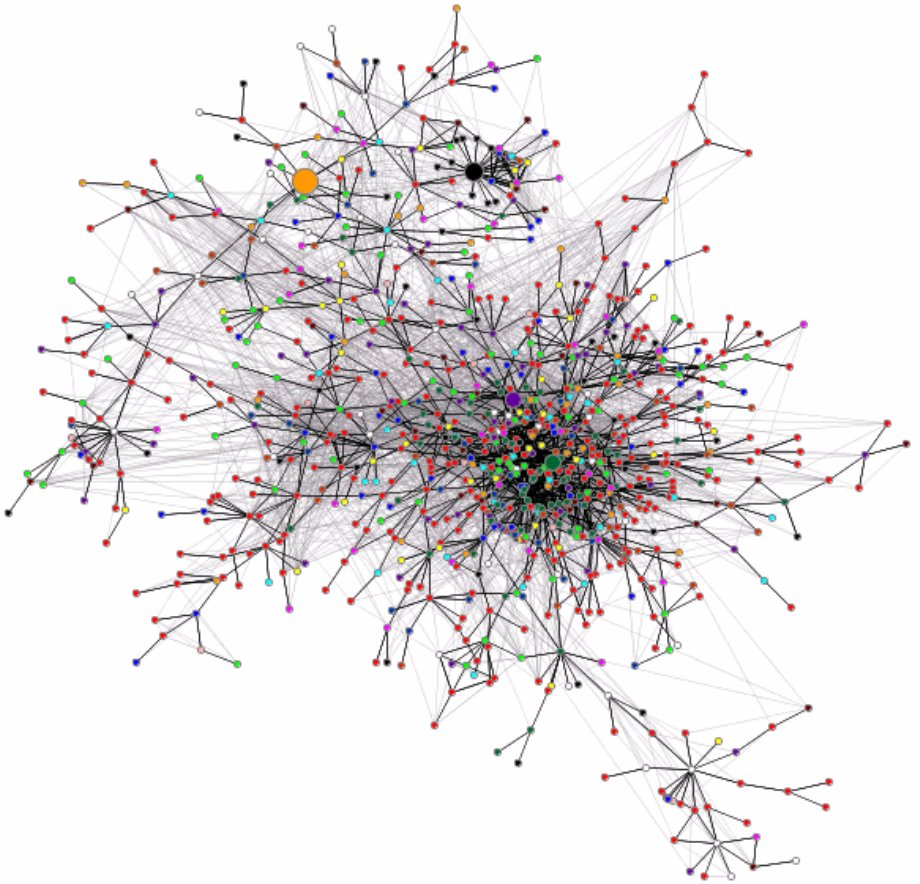
- **Agent-basierter Ansatz (Information Retrieval View)**
  - **Indexierung und Klassifizierung der Daten**
  - **Intelligente Such-Agenten**
    - durchsuchen das Internet nach relevanten Informationen
    - Verwendet ein bestimmtes User-Profil
    - Strukturiert und interpretiert gefundene Informationen
    - Bsp: ShopBots
  - **Informationsfilterung/ Kategorisierung**
    - Verwenden unterschiedliche Informationsabfragetechniken
    - Verwendung von Clusteringverfahren oder Linkstrukturen zur Klassifikation der Informationen
  - **Personalisierte Web-Agenten**
    - Lernen Präferenzen der Internetnutzer kennen und weisen auf Vergleiche mit Nutzern, die ähnliche Interessen haben hin (Collaborative Filtering)

Search Result Mining

- **Datenbank-basierter Ansatz (Database View)**
  - **„Webseite in eine Datenbank umwandeln“**
  - **Inhalt für anspruchsvolle Queries aufzubereiten**
  - **Multilevel-Datenbanken**
    - Datenbanken mit unterschiedlichen Niveaus
    - Im untersten Niveau sind halb-strukturierte Informationen enthalten
    - Datenbanken mit höherem Niveau beinhalten Meta-Daten oder Generalisierungen
    - Wurden aus niedrigeren Niveaus extrahiert
    - Daten liegen anschließend in strukturierter Form vor

3 Web Structure Mining

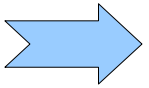
Extraktion von Wissen aus Hyperlinkstrukturen



- Geschichtlicher Abriss
- Das Web als Graph
- Ranking Algorithmen

3.1 Geschichtlicher Abriß

- **Situation vor 1996**
  - Abfragen der ersten Suchmaschinen über inhaltliche Übereinstimmung/Ähnlichkeit
  - Ist für **User Query** indexiertes **Keyword** vorhanden ?
  - Einsatz von Information Retrieval Algorithmen für Retrieval und Ranking
  - Probleme mit zunehmender Größe des Web:
    - zu **große Anzahl von Ergebnissen**
    - **Spamming** (Mißbrauch von Metatags, “versteckte” ”Keywords)
    - Verfälschung der Relevanzkriterien

Ausweg ?  Hyperlinks

3.1 Geschichtlicher Abriß

- **Situation ab 1996**
  - Forschung an Universitäten und bei Suchmaschinenbetreibern
  - Wie können die **Relationen zwischen Webseiten** genutzt werden ?
  - 2 Arten von Hyperlinks:
    - Intern
    - Extern ('out-going hyperlinks')
  - Externe Links **übertragen "Authorität"** auf die Seiten auf die sie verlinken
  - Seiten mit vielen 'incoming' Links von anderen Seiten besitzen best. **Qualität** --> sogn. Authorities
  - Einsatz als (zusätzliches) Ranking Kriterium

  - Verfahren bereits aus Co-Citation wissenschaftlicher Artikel bekannt

    → CiteSeer

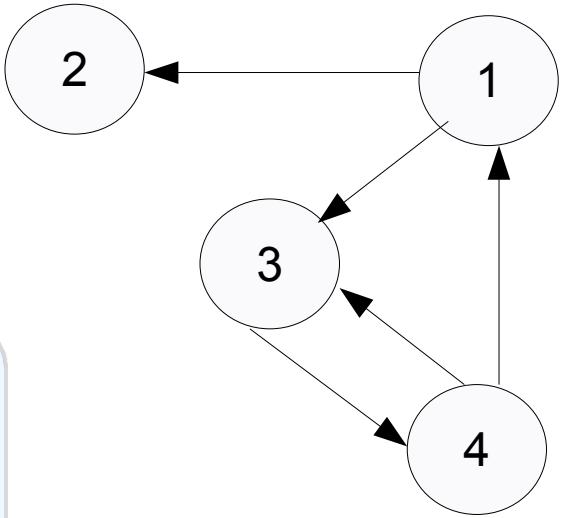
3.2 Graphenstruktur des Web

- Web kann als gerichteter Graph betrachtet werden
- Besitzt Ein- und Ausgangsgrad

$$G = (V, E)$$

V – Webseiten (Knoten)

E – Hyperlinks (Kanten)



Eingangsgrad $w(p)$ einer Seite $p \in V$ beschreibt die Menge aller eingehenden Links.

Ausgangsgrad $o(p)$ einer Seite $p \in V$ beschreibt die Menge aller ausgehenden Links.

Beispiel:

- V = {1,2,3,4}
- E = {(1,2), (1,3), (3,4), (4,3), (4,1)}
- *w*(3) = 2, *w*(2) = 1
- *o*(2) = 0, *o*(4) = 2

3.2 Graphenstruktur des Web

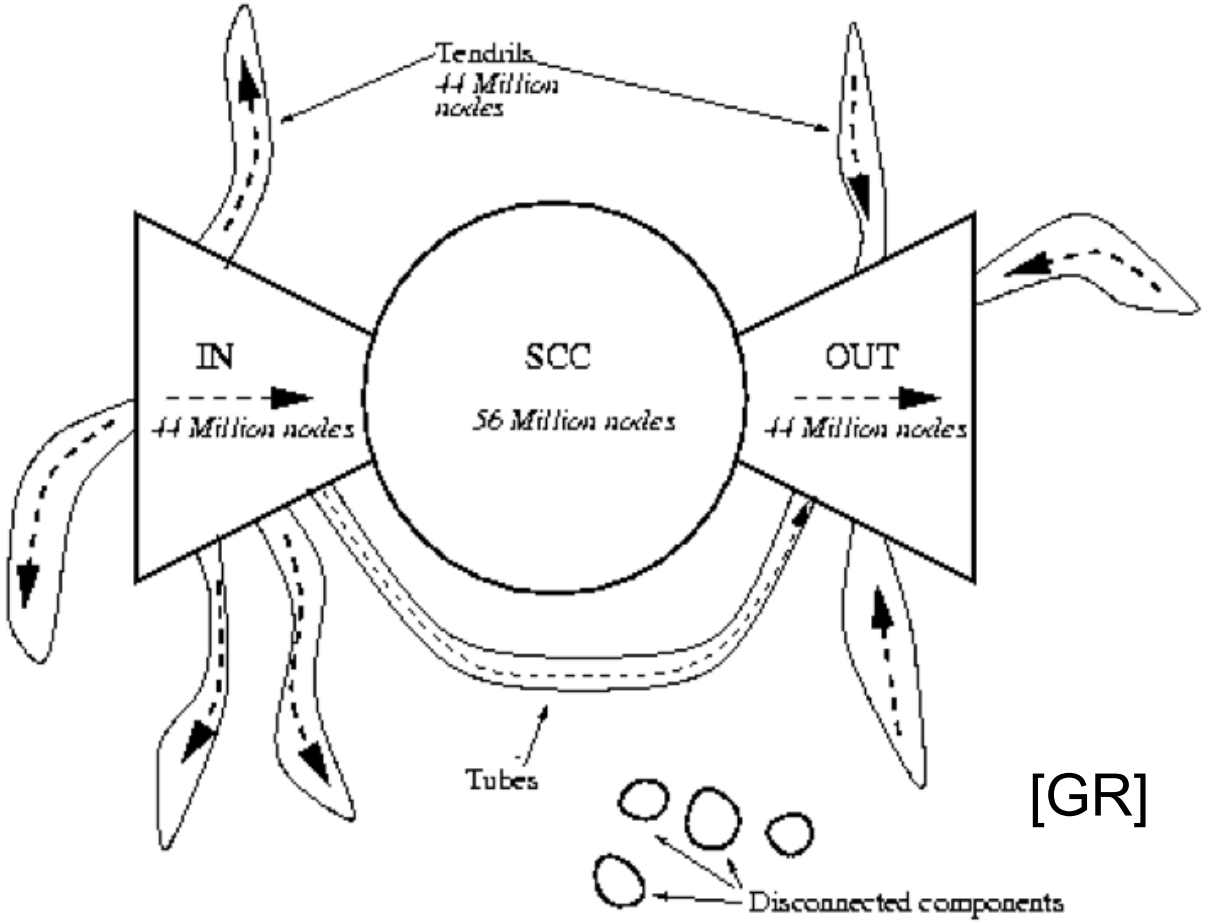
Erste umfangreiche Analyse der Grapheneigenschaften druch *Broder* et al. 2000:

- Altavista crawl (Mai 1999) mit 203 Millionen URLs und 1,4 Milliarden Links
- Größe: 9.5 GB
- Laufzeit BFS mit 100 Millionen Knoten: ca. 4 min.

Untersuchungen:

- Eingangs- / Ausgangsgrad der Knoten
- Größe de Web

3.2 Graphenstruktur des Web



[GR]

The “Giant Bow Tie”

- INSET
- OUTSET
- SCC
- Tendrils
- Tubes

- Wahrscheinlichkeit für Pfad zwischen 2 zufällig gewählten Seiten ist **0.24**
- Durchmesser in SCC : 27 “Hops”

3.2 Graphenstruktur des Web

Welche Erkenntnisse wurde gewonnen?

- Korrelation von Incoming Links und Popularität einer Seite
- “Incoming Degree” allein nicht ausreichend
- Wenig Aussagekraft, da nicht immer (konkrete) Inhalte verlinkt werden
- Notwendigkeit komplexerer Verfahren

Einflussreichste Hyperlinkbasierte Suchalgorithmen:

- **Page Rank**
- **HITS**

3.3 Page Rank

- Entwickelt von Lawrence “Larry” Page und Sergej Brin
- 1998 auf der 7. World Wide Web Conference vorgestellt
- **Grundlage für** die **Google** Suchmaschine
  - *“The Anatomy of a Large-Scale Hypertextual Web Search Engine”*
- Page Rank nutzt den demokratischen Charakter des Web
- **Hyperlink** von *x* auf *y* entspricht **Votum** von *x* an *y*
- Ranking erfolgt statisch
- SuchQueries haben keinen Einfluss
- Rekursivität → Gesamtes Web
- Implementierungsdetails geheim

...And that, in simple terms, is how you increase your ranking on search engines.”

3.3 Page Rank - Algorithmus

Page Rank der Seite i, ist die Summe von
Page Ranks aller auf sie verweisenden Backlinks

$$P(i) = (1-d) + d \sum_{(i,j) \in E} \frac{P(j)}{Oj}$$

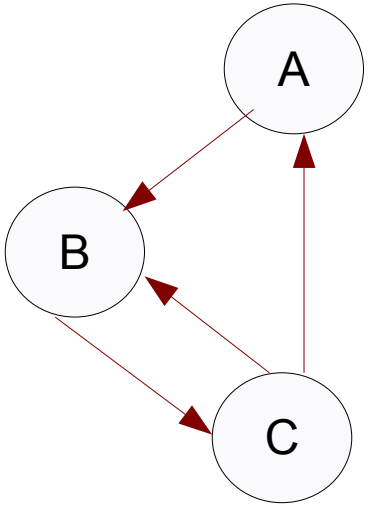
- P(i) – zu berechnender Rank der Seite i
- P(j) – Page Rank von *j*
- Oj – Anzahl der Outlinks von *j*
- D – Dämpfungsfaktor

- PageRank der Seiten *j* fließt nicht gleichmäßig in den PageRank von Seite *i*
- Gewichtung durch Anzahl der Links -> Oj
- je mehr ausgehende Links Seite j hat, umso weniger Page Rank geht an Seite i
- Dämpfung –> Modell zur Abbildung von Benutzer-Verhalten
- “Random Surfer Model”

3.3 Page Rank

- System aus *n* linearen Gleichungen mit *n* Unbekannten
- $P=(PR(1), PR(2), PR(3), ..., PR(n))$
- $P= A^T \, x \, P$

- **PR(A) = 1-0,5 + 0,5 (PR(C) / 1)**
- **PR(B) = 1-0,5 + 0,5 (PR(A) / 1 + PR(C) / 2)**
- **PR(C) = 1-0,5 + 0,5 (PR(B) / 1)**



d = 0.5

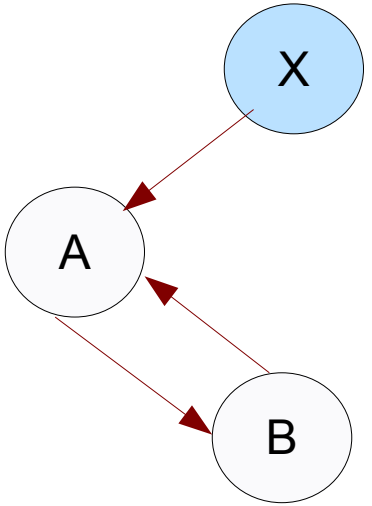
3.3 Page Rank

Beeinflussung durch *Incoming Links*

- X hat einen Page Rank von 8

- **PR(A) = 1-0,5 + 0,5 (8 + PR(B) / 1)**
- **PR(B) = 1-0,5 + 0,5 (PR(A) / 1)**

- **Summe = N**



d = 0.5

3.3 Page Rank

Schwachstellen:

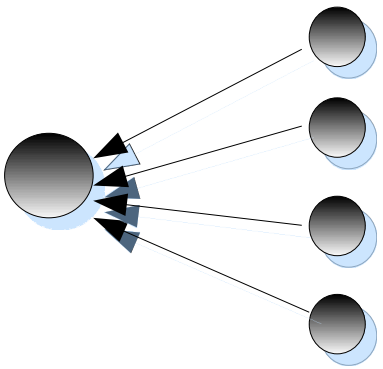
- Dangling Links
- Query-Unabhängiges Ranking

Verbesserungen und Alternativen:

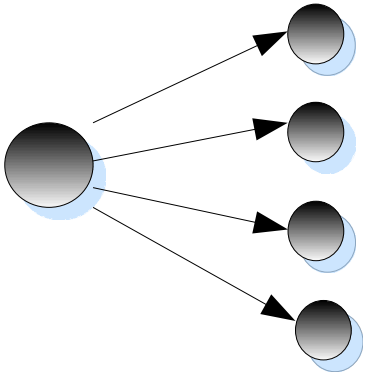
- Timed PageRank
- Topic Sensitive PageRank

3.4 Hyptertext Induced Topic Search

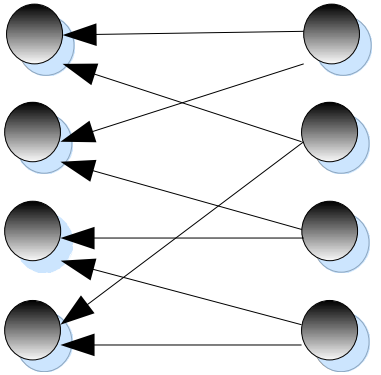
- 1998 von J. Kleinberg vorgestellt
- *“Authoritative sources in a hyperlinked environment*.”
- Im Gegensatz zu Page Rank dynamisch
- Einbeziehung der Suchanfrage
- Nutzung von Hubs & Authorities
- “Mutual Reinforcement Relationship”



Authority



Hub



Set von Auth. & Hubs

3.4 HITS Algorithmus

Gegeben: Suchanfrage *q*

(1) Senden der Suchanfrage an Suchmaschine

(2) Auswahl von *t* Seiten, der am höchsten gerankten (RootSet *W*)

(3) Hinzufügen von Seiten die auf Seiten in W verweisen
oder auf die von W aus verwiesen werden (BaseSet *S*)

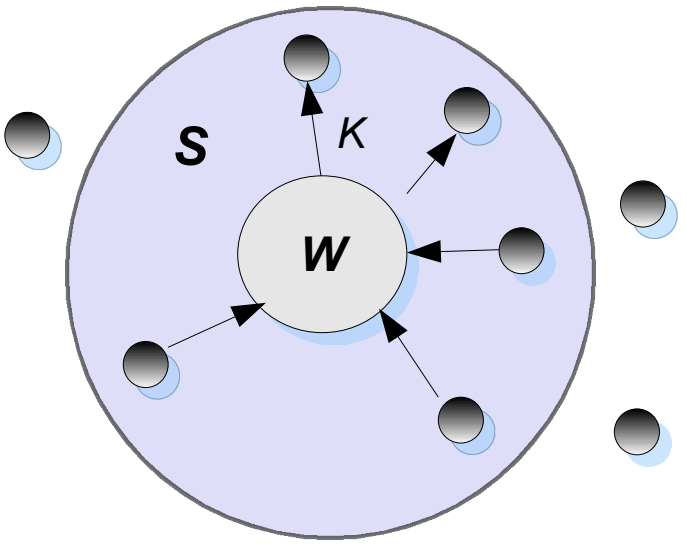
(4) Zuweisung von Hub und Authority Scores in S
1 für Kante zwischen i,j, sonst 0

Authority-Score

$$a(i) = \sum_{(j,i) \in E} h(j)$$

Hub-Score

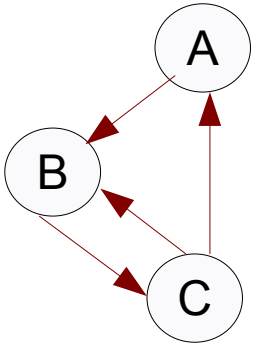
$$h(i) = \sum_{(i,j) \in E} a(j)$$



k = max. Outlinks in W

3.4 HITS Algorithmus

(5) Adjazenzmatrix *L* des Graphen *G*



$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

(6) k-te Iteration

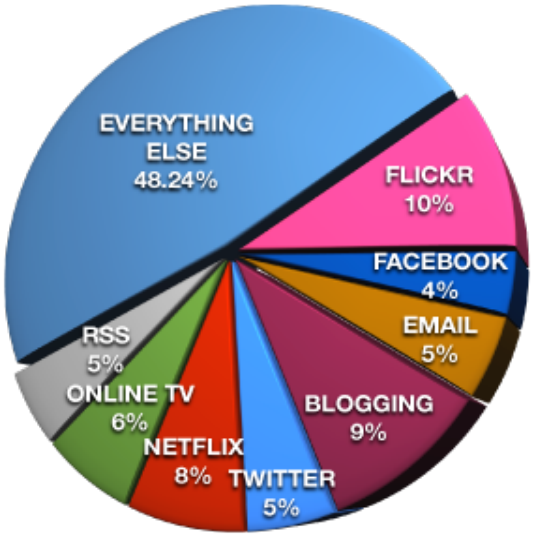
$$a_k = L^T L a_{k-1}$$

$$h_k = L L^T h_{k-1}$$

(7) Auswahl der Höchsten Hub-/Authority Ergebnisse, Rückgabe als “relevante” Suchtreffer

4 Web Usage Mining

Extraktion von Wissen aus dem Nutzerverhalten



4 Motivation

- Anzahl **Jahre um 50 Millionen** Nutzer **zu erreichen**
  - Radio - 38 Jahre
  - TV - 13 Jahre
  - Internet 4 Jahre
  - Facebook - 100 Mio. in 9 Monaten
- Anzahl der **Blogs: 200.000.000**
- **70 %** der Konsumenten **Vertrauen Empfehungen**
- Mehr als **1,5 Mio. Contentobjekte** werden **täglich auf Facebook** getauscht (web links, news stories, blog posts, notizen, photos, etc.)
- **50 % der Weltbevölkerung** ist **unter 30** Jahren
- **96 %** der nach 1980 geborenen sind **Mitglied eines Sozialen Netzwerkes**

[SOC]

4 Motivation

- hohe Anzahl an eCommerce, WebServices
- Generieren wertvolle Nutzer-/ Nutzungsdaten
- Usage Mining automatisiert Entdeckung und Analyse von Mustern in diesen Daten
- Abbildung der Interaktion mit Webressourcen

> *"*[...] *discover usage patterns from Web data, in order to understand and better serve the needs of web-based applications."* [SRIV]

> *"*[...] *automatic discovery and analysis of patterns in clickstream and* [...] *data as a result of user interactions with Web resources* [...]*."* [LIU]

Erkennung, Modellierung, Analyse von Verhaltensmustern- und Profilen bei der Nutzerinteraktion mit Webseiten

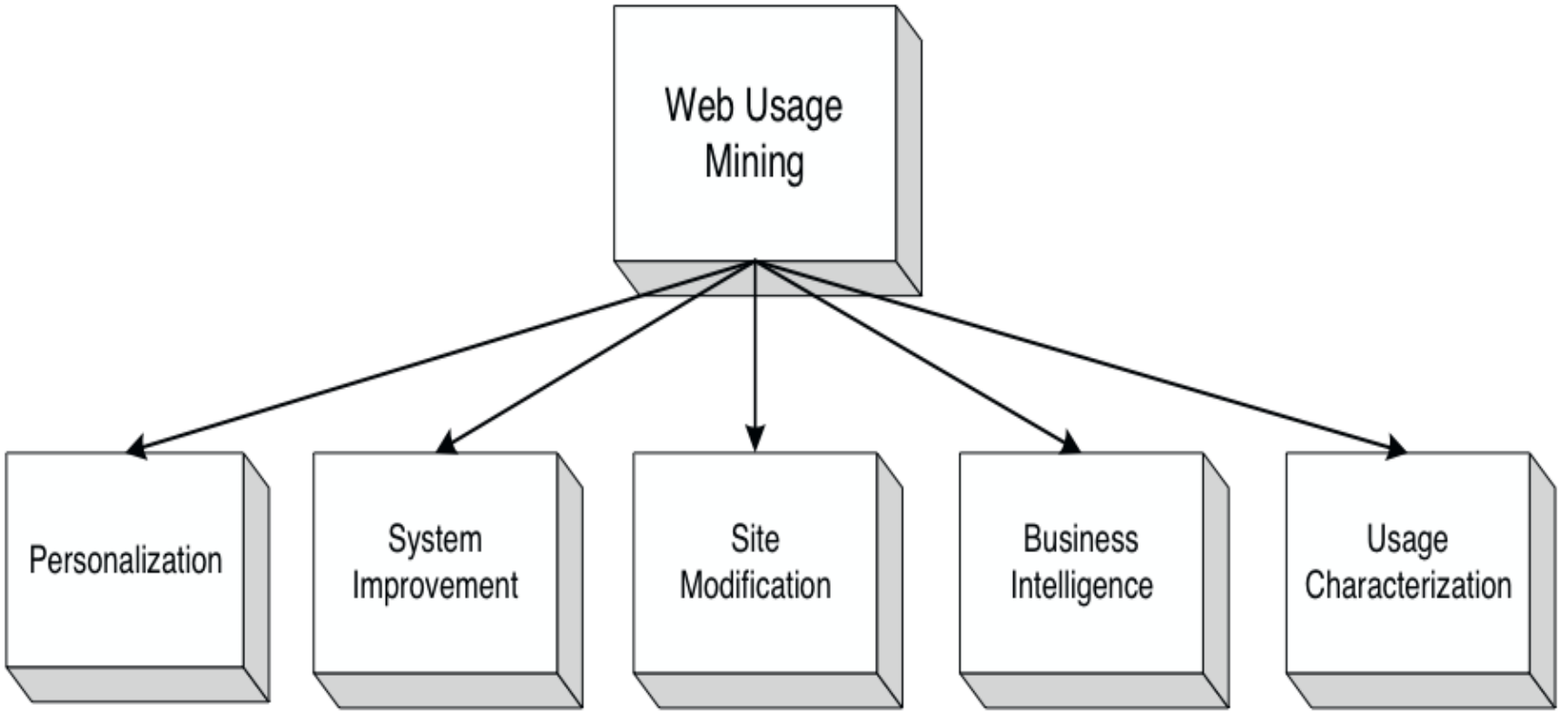
4 Motivation

- **Usage Mining als Grundlage für erfolgreiches CRM**
- Wertvolles Wissen über Wünsche und Bedürfnisse des Kunden
- Informationsbedarfe:
  - Zusammensetzung der Besucher
  - Wirkung von Werbung
  - Kaufverhalten
  - Bewertung von Seiteninhalten

- **Zustandserfassung und Optimierung von Web Auftritten**
  - E-Commerce-Unternehmen
  - Online - Buchhändler / Versandhandel

  [HWB, S.4]

4.1 Anwendungsgebiete



[SRIV]

4.2 Realisierungsansätze

- **Keine Datenbank (Dateiebene)**
  - Nutzung von Logfiles
  - Beschränkungen unterworfen
- **Einsatz einfacher Datenbank**
  - Speicherung und Auswertung von Logfiles
  - Große Datenmengen
  - Einsatz von Data Mining Methoden
- **Data Warehouse**
  - Einbeziehung multipler Datenquellen
  - OLAP-Auswertung

4.3 Web Daten

- **Inhaltsdaten (Content)**
  - überwiegend Text und Bilder
  - Semantische Zusatzinformationen
  - Meta-Daten, RDF

- **Strukturdaten**
  - „Inter /- Intrapage Struktur“
  - Beschreibt Organisation des Inhalts

- **Nutzungsdaten (Usage)**
  - Nutzungsmuster von Webseiten
  - IP-Adressen, Zugriffszeit, Pageviews, Cookies

- **Nutzer-Profil Daten (User Profile)**
  - Demographische Daten
  - Personengebunden

4.3 Web Daten

- **Inhaltsdaten (Content)**
  - überwiegend Text und Bilder
  - Semantische Zusatzinformationen
  - Meta-Daten, RDF

- **Strukturdaten**
  - „Inter /- Intrapage Struktur“
  - Beschreibt Organisation des Inhalts

- **Nutzungsdaten (Usage)**
  - Nutzungsmuster von Webseiten → ***Clickstreams / “Episoden“***
  - IP-Adressen, Zugriffszeit, Pageviews, Cookies

Nutzer-Profil Daten (User Profile)
- Demographische Daten
- Personengebunden

4.4 Datenquellen – Usage Data

- **Server Level**
  - Wichtige Quelle für Usage Mining
  - Verhalten einzelner Nutzer
  - Paralleler Zugriff mehrerer Nutzer
  - **Common Logfiles / Extended Logfile Format**
  - **Paket Sniffer**

Client Level

- Nutzung von Remote Agents → **JavaScript, Java Applets**
- **Cookies**
- Verbesserung von Problemen beim Caching und Session Identitifikation
- Nur einzelne Nutzerinteraktion “messbar“

- **Proxy Server Level**
  - Caching und Vorhersage von Requests
  - Identifikation von Nutzergruppen möglich

4.5 Common Log File Format

- **Logging innerhalb von http-deamons**

AccessLog
ProxyAccessLog
CacheAccessLog
ErrorLog
LogFileDateExt names
LogFormat
LogTime
Common logfile format

[W3C]

- **Common logfile format**

```
remotehost - authuser [date] "request" status bytes

127.0.0.1 - bob [01/May/2010:10:11:12 -0700] "GET
/webmining.pdf HTTP/1.0" 200 2326
```

4.5 Extended Log File Format

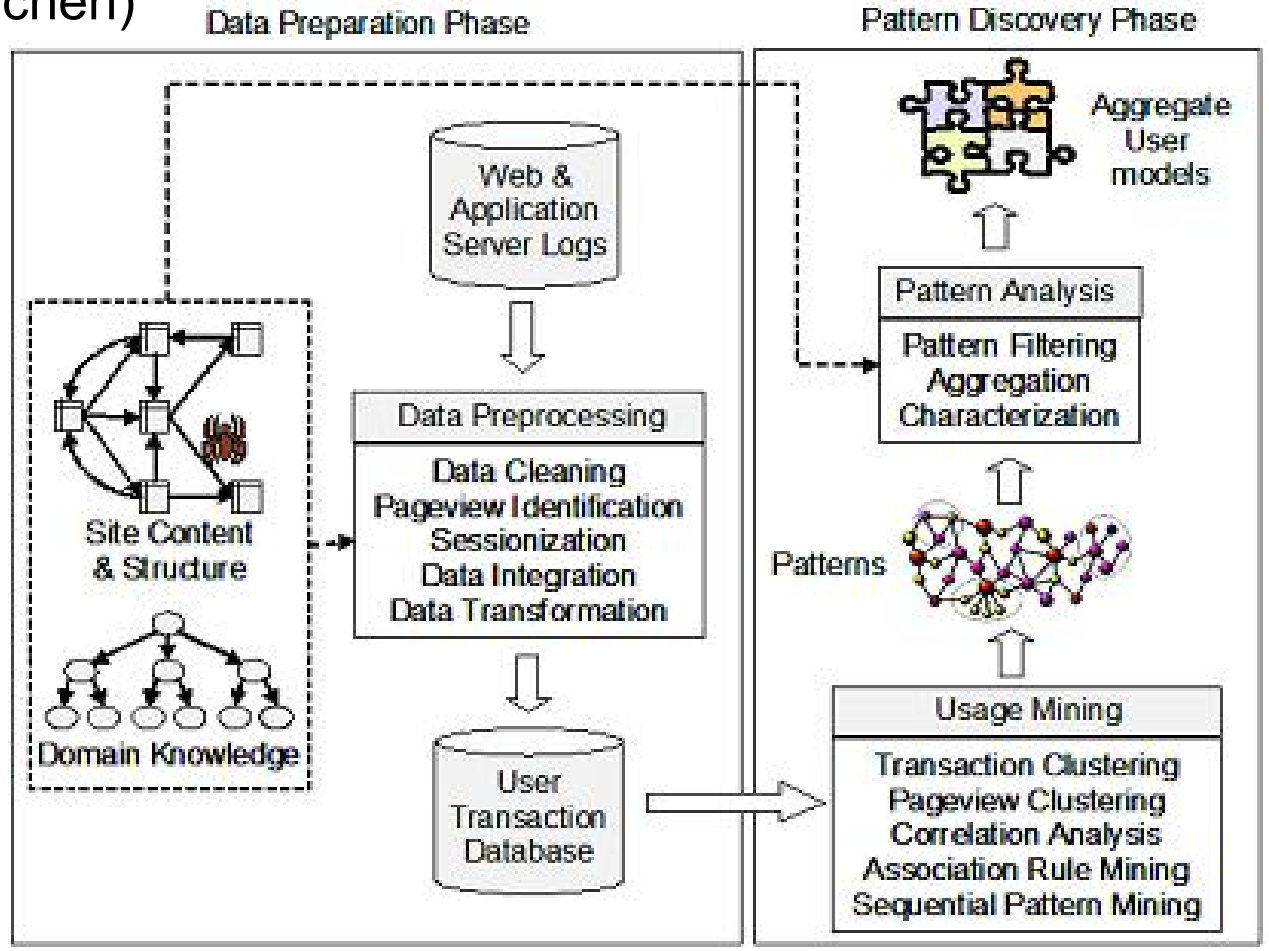
- Erweiterungen um Demographische Daten
- Möglichkeit der Erkennung von Proxy-Servern

```
80.202.8.93 - - [02/May/2010:22:43:28 -0600]  "GET
/foo/images/foobar.gif HTTP/1.1" 200 5006
"http://foo.foo.bar/doc/index.html" "Mozilla/4.0
(compatible; MSIE 5.0; Windows 2000) Opera 6.01“
```

[W3C]

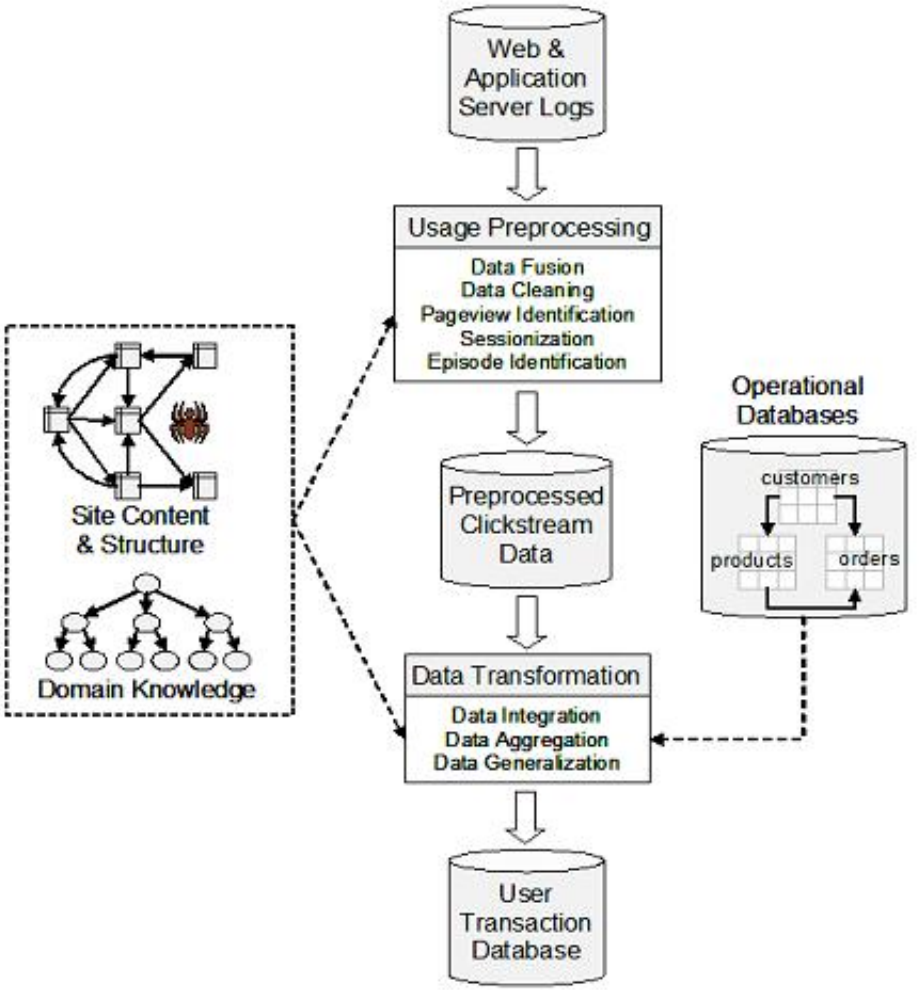
4.6 Der KDD-Prozess

- Ähnlich wie beim standart data mining prozess, kann der Web Usage Mining Prozess in 3 Teile geteilt werden
  - Data collection & pre-processing / Vorverarbeitung
    - Clickstream Daten werden gesäubert und in Benutzer Transaktions-Gruppen aufgeteilt (Repräsentieren die Aktivität jedes Users während den verschiedenen Besuchen)
  - Pattern discovery
  - Pattern analysis



4.6.1 Data collection & pre-processing

- Wichtiger Prozess für die erfolgreiche Extraktion von nützlichen Mustern in den Daten
- Meistens zeitaufwändig
- Gesamter Vorbereitungsprozess (auch data preparation)
  - Datenfusion und -säuberung
  - User und Session identification
  - pageview identification



4.6.1 Web Usage Data pre-processing

- Datenfusion (data fusion)
  - In umfangreichen Webseiten befinden sich die Inhalte auf verschiedenen Servern
  - Datenfusion -> Logfiles der verschiedenen Server fusionieren
  - Da es keine “shared embedded” session ids gibt, werden heuristische Methoden, basierend auf dem “referrer”-Feld des Serverlogs, zusammen mit verschiedenen sessionization und Useridentifikations methoden verwendet
- Datensäuberung (data cleaning)
  - Für die Analyse unwichtige Logeinträge entfernen:
    - Irrelevante Verweise zu eingebetteten Objekten
    - Verweise zu Styledateien, Graphiken, Sounddateien, ...
  - Aber auch einige Felder der Logdatei werden entfernt:
    - Menge der gesendeten Daten, HTTP Protocol version, ...
  - Crawlereinträge

183.121.143.32 - - [18/Mar/2003:08:05:03 +0200] "GET /images/bild.png HTTP/1.1" 200 805 "http://www.google.org/"

4.6.1 Web Usage Data pre-processing

- Pageview Indentification
  - Identifikation von Seitenzugriffe
  - Mehrere Attribute müssen berücksichtigt werden:
    - Pageview id (URL)
    - Pageview typ (Informationsseite, Indexseite, Produktseite,...)
    - Weitere Metadaten (keywords, Produktattribute,...)
- Benutzeridentifikation (User Identification)
  - Besucher unterscheiden
  - Sequenz die zu einem Besucher gehört wird auch user activity record genannt
  - Identifikation über
    - Cookies (cookie-id)
    - Kombination aus IP und weiteren Informationen (user agent, referrer,...)
    - Registrierung

4.6.1 Web Usage Data pre-processing

User Identifikation mittels Ip und Agent

Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE6;WinXP;SP1
0:12	2.3.4.5	B	C	IE6;WinXP;SP1
0:15	2.3.4.5	E	C	IE6;WinXP;SP1
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE6;WinXP;SP1
0:22	1.2.3.4	A	-	IE6;WinXP;SP2
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE6;WinXP;SP2
0:33	1.2.3.4	B	C	IE6;WinXP;SP2
0:58	1.2.3.4	D	B	IE6;WinXP;SP2
1:10	1.2.3.4	E	D	IE6;WinXP;SP2
1:15	1.2.3.4	A	-	IE5;Win2k
1:16	1.2.3.4	C	A	IE5;Win2k
1:17	1.2.3.4	F	C	IE6;WinXP;SP2
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

User 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

User 2

0:10	2.3.4.5	C	-
0:12	2.3.4.5	B	C
0:15	2.3.4.5	E	C
0:22	2.3.4.5	D	B

User 3

0:22	1.2.3.4	A	-
0:25	1.2.3.4	C	A
0:33	1.2.3.4	B	C
0:58	1.2.3.4	D	B
1:10	1.2.3.4	E	D
1:17	1.2.3.4	F	C

4.6.1 Web Usage Data pre-processing

- Sessionization
  - Aufteilen der user activity records in sessions
  - Webseiten ohne weitere Authentifizierungsinformationen der User oder eingebettete sessiond-ids müssen auf heuristische Methoden zurückgreifen
  - 2 Kategorien:
    - Zeit-orientierte Heuristik
      - Timeout wird verwendet um zwischen nachfolgenden Sessions zu unterscheiden
    - Struktur-orientierte (h-ref) Heuristik
      - Verfolgt die Linkstruktur aus dem “referrer” Feld

4.6.1 Web Usage Data pre-processing

Zeit-orientierte Heuristik

User 1

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Session 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C

Session 2

1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

h-ref Heuristik

User 1

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Session 1

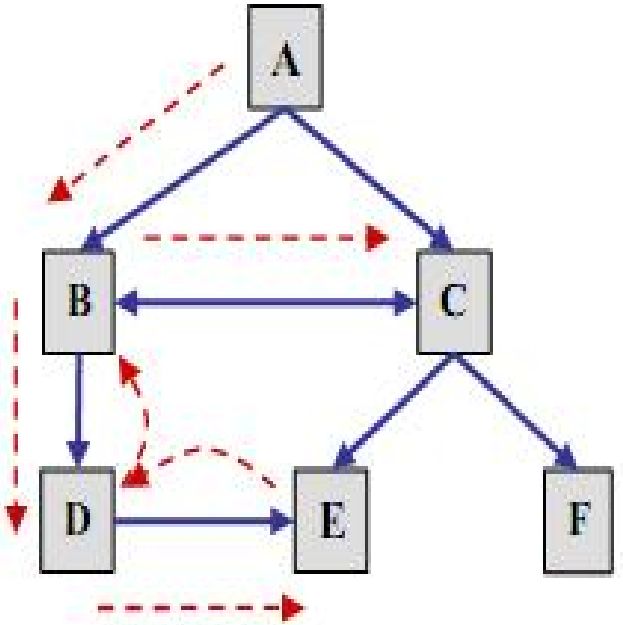
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:26	1.2.3.4	F	C

Session 2

1:15	1.2.3.4	A	-
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

4.6.1 Web Usage Data pre-processing

- Pfadvervollständigung (Path Completion)
  - Versuch, fehlende Zugriffe (enstanden durch Verwendung des Cache, back-button), auf direkt verlinkte Verbindungsseiten im Logfile, zu ergänzen
  - Bsp:



User's actual navigation path:

A → B → D → E → D → B → C

What the server log shows:

URL	Referrer
A	--
B	A
D	B
E	D
C	B

4.6.1 Web Usage Data pre-processing

- Data Integration
  - Zuvorige Vorverarbeitungsschritte resultieren in sog. user session, die jeweils einer bestimmten pageview sequenz entsprechen
  - Um bestmögliches Gerüst für die Mustererkennung zu bieten, müssen u.U weitere Quellen integriert werden (Kundendatenbank, Onlineumfragen,...)
  - Wichtig für e-commerce
    - Durch Registrierung(session_id) kann ein user eindeutig identifiziert werden
    - Durch Verwendung der Log-Files und zusätzlicher Quellen z.B. Kundendatenbank können zusätzliche kundenindividuelle Informationen zur Logfile-Datenbasis hinzugefügt werden
    - Unterschiedliche Datenquellen können mit einem eindeutigen Kundenschlüssel versehen werden und in einem Datawarehouse ihr Zusammenhang weiter analysiert werden

4.6.1 Web Usage Data pre-processing

- Data Transformation
  - Vorbereiten der gesammelten Daten für die Analyse
  - Diese müssen als Datenmatrix strukturiert sein
  - user-pageview matrix (Transaction matrix)
    - Binäre Darstellung -> pageview existiert (nicht)
    - Dauer in Sekunden
  - Content-enhenced transaction matrix

Pageviews

Sessions / users

	A	B	C	D	E	F
user0	15	5	0	0	0	185
user1	0	0	32	4	0	0
user2	12	0	0	56	236	0
user3	9	47	0	0	0	134
user4	0	0	23	15	0	0
user5	17	0	0	157	69	0
user6	24	89	0	0	0	354
user7	0	0	78	27	0	0
user8	7	0	45	20	127	0
user9	0	38	57	0	0	15

Transaction matrix

4.6.1 Web Usage Data pre-processing

- Content-enhenced transaction matrix

	A.html	B.html	C.html	D.html	E.html
user1	1	0	1	0	1
user2	1	1	0	0	1
user3	0	1	1	1	0
user4	1	0	1	1	1
user5	1	1	0	0	1
user6	1	0	1	1	1

User-pageview matrix

	A.html	B.html	C.html	D.html	E.html
web	0	0	1	1	1
data	0	1	1	1	0
mining	0	1	1	1	0
business	1	1	0	0	0
intelligence	1	1	0	0	1
marketing	1	1	0	0	1
ecommerce	0	1	1	0	0
search	1	0	1	0	0
information	1	0	1	1	1
retrieval	1	0	1	1	1

Term-pageview matrix

	web	data	mining	business	intelligence	marketing	ecommerce	search	information	retrieval
user1	2	1	1	1	2	2	1	2	3	3
user2	1	1	1	2	3	3	1	1	2	2
user3	2	3	3	1	1	1	2	1	2	2
user4	3	2	2	1	2	2	1	2	4	4
user5	1	1	1	2	3	3	1	1	2	2
user6	3	2	2	1	2	2	1	2	4	4

Content-enhenced transaction matrix aus den 2 vorherigen Matrizen

4.6.2 Mustererkennung & Analyse

- Mustererkennung
  - Schlüsselkomponente des Webmining
  - Verwendung von Algorithmen und Techniken aus dem data mining
  - Statistische Analyse
  - Clustering
  - Assoziationsanalyse
  - Klassifikation
  - Sequentielle Muster und Navigationsmuster
  - Klassifikation und Prognose
- Analyse
  - Letzter Schritt im KDD-Prozess
  - Unwichtige Regeln oder Muster entfernen
  - Interessante Regeln oder Muster extrahieren

4.6.2 Mustererkennung & Analyse

- Statistische Analyse
  - Session- und Besucheranalyse
  - Vorverarbeitete Daten werden nach bestimmten Einheiten zusammengesetzt (Tage, Sessions, Besucher, Domains)
  - Statistische Techniken werden angewendet um Wissen über das Nutzerverhalten zu erlangen
    - Meistbesuchte Seiten
    - Durchschnittsverweilzeit auf einer Seite
    - Durchschnittslänge des Besuchspfades
    - Eintritts-/Austrittspunkte
  - Dieses Wissen kann u.U nützlich sein für Marketingentscheidungen

4.6.2 Mustererkennung & Analyse

- Clustering
  - Daten die gleiche Eigenschaften haben werden gruppiert
  - Page clusters und User clusters
  - Page clusters:
    - Seiten oder Produkte zum selben Thema/Kategorie gruppieren
    - Objekte, die zusammen gekauft werden, werden automatisch gruppiert
    - Erstellen von Links im Zusammenhang mit zuvor besuchten Seiten / gekauften Produkten(Querverweise)
  - User clusters:
    - Gruppierung von Benutzern die ein gleiches Surfverhalten aufzeigen
    - Mittels standart clustering Algorithmen(z.B. k-means) kann User transaction Tabelle in mehrere Transaktionen aufgeteilt werden. → transaction clusters
    - Ziel ist es die Möglichkeit zu bieten jedes Segment analysieren zu können im Sinne der Business Intelligence oder Personalisierung

4.6.2 Mustererkennung & Analyse

Bsp Clusters :

		A	B	C	D	E	F
Cluster 0	user 1	0	0	1	1	0	0
	user 4	0	0	1	1	0	0
	user 7	0	0	1	1	0	0
Cluster 1	user 0	1	1	0	0	0	1
	user 3	1	1	0	0	0	1
	user 6	1	1	0	0	0	1
	user 9	0	1	1	0	0	1
Cluster 2	user 2	1	0	0	1	1	0
	user 5	1	0	0	1	1	0
	user 8	1	0	1	1	1	0

Aggregate Profile for Cluster 1	
Weight	Pageview
1.00	B
1.00	F
0.75	A
0.25	C

4.6.2 Mustererkennung & Analyse

- Assoziationsanalyse
  - Mit Assoziationsregeln können Gruppen von Objekten /Seiten erkannt werden die in der Regel zusammen abgerufen oder gekauft werden
  - Bietet Webseiten die Möglichkeit ihren Inhalt effizienter zu organisieren oder auch cross-sale Produkte auf der gleichen Seite anzubieten
  - Eine der Hauptaufgaben der Assoziationsregeln sind Empfehlungen

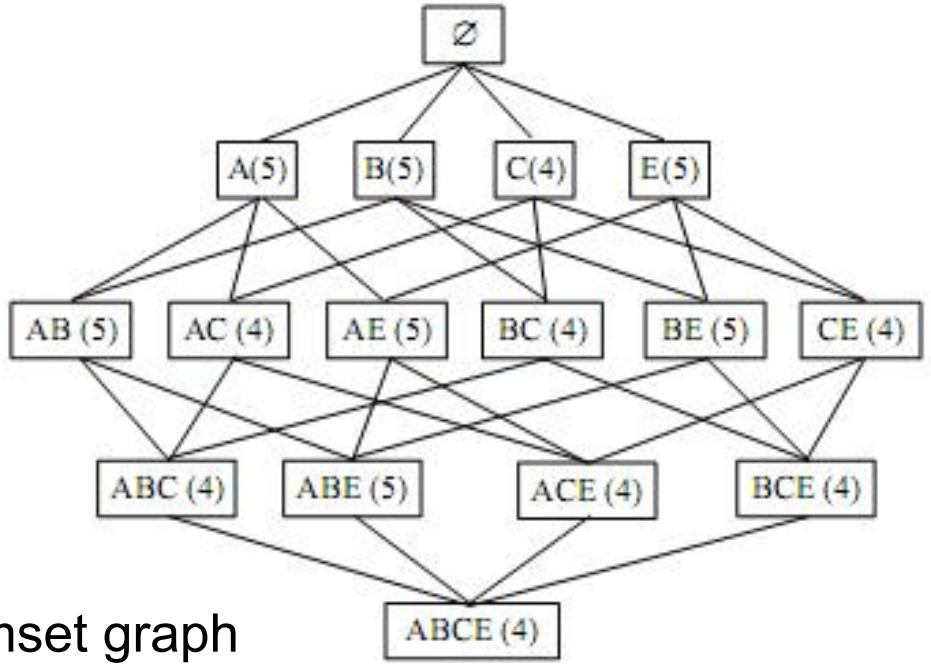
    bzw. “collaborative filtering”
  - Verwendung des Apriori Algorithmus
    - Findet Objektgruppen (pageviews auf dem vorverarbeiteten Log) die häufig zusammen in vielen Transaktionen auftreten
    - Häufige itemsets werden in einem azyklischen Graphen gespeichert → frequent itemset graph
  - z.B. Special-offers/,/products/software/ ->shopping-cart/
  - Warenkorbanalyse

4.6.2 Mustererkennung & Analyse

Web transactions und frequent itemsets

Transactions
A, B, D, E
A, B, E, C, D
A, B, E, C
B, E, B, A, C
D, A, B, E, C

Size 1		Size 2		Size 3		Size 4	
Itemset	Supp.	Itemset	Supp.	Itemset	Supp.	Itemset	Supp.
A	5	A,B	5	A,B,C	4	A,B,C,E	4
B	5	A,C	4	A,B,E	5		
C	4	A,E	5	A,C,E	4		
E	5	B,C	4	B,C,E	4		
		B,E	5				
		C,E	4				



frequent itemset graph

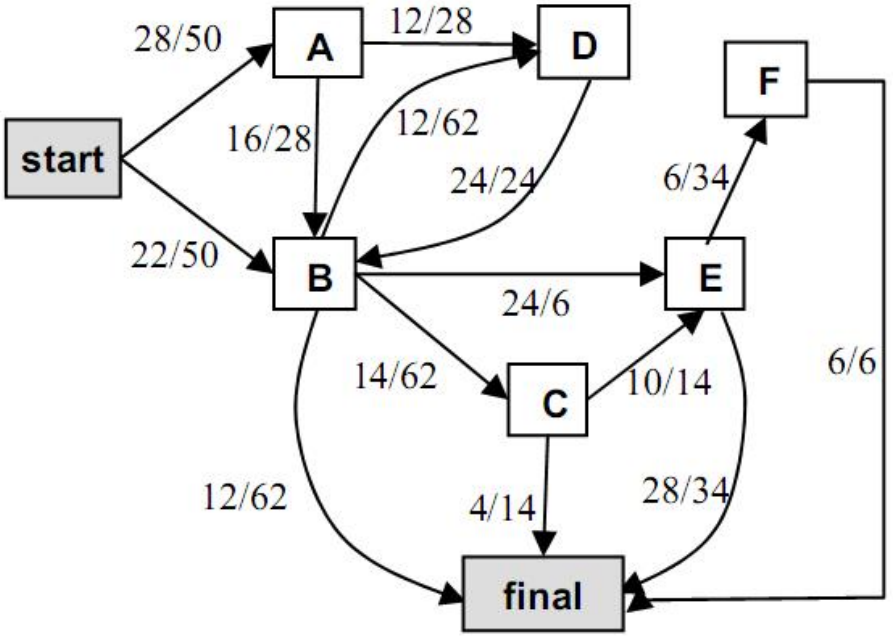
4.6.2 Mustererkennung & Analyse

- Sequentielle Muster und Navigationsmuster
  - Muster innerhalb einer Session erkennen, z.B auf eine Objektgruppe folgt ein anderes Objekt in einer bestimmten zeitlichen Reihenfolge
  - Erkennen von oft verwendeten Pfaden
  - Hiermit können Vorhersagen getroffen werden bzgl. der besuchten Seiten machen.
  - Sinnvoll um z.B. Zielgruppen basierte Werbung zu platzieren
  - Navigationsaktivität kann als sog. Markov model dargestellt werden:
    - Jeder pageview kann als Zustand dargestellt werden und die Übergangswahrscheinlichkeit zwischen 2 Zuständen stellt die Wahrscheinlichkeit dar, dass ein User von einer Seite zu einer anderen Navigieren wird

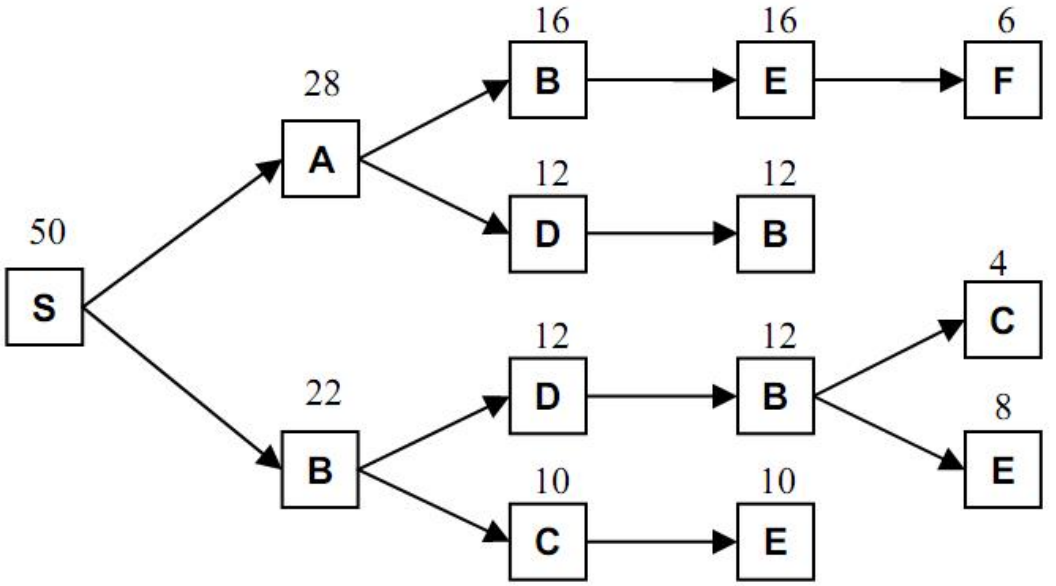
4.6.2 Mustererkennung & Analyse

Navigationspfad als Markov Kette

Transaction	Frequency
A, B, E	10
B, D, B, C	4
B, C, E	10
A, B, E, F	6
A, D, B	12
B, D, B, E	8



Navigationspfad als Aggregatbaum



4.6.2 Mustererkennung & Analyse

- Klassifikation und Prognose
  - Information in eine von vordefinierten Klassen einzuordnen
  - Profil von Benutzern erstellen die zu einer gemeinsamen Klasse gehören
  - Überwachte Lernalgorithmen werden hier verwendet:
    - Entscheidungsbäume
    - Naive Bayesian Classifier
    - K-nearest neighbor classifiers
  - Bsp:
    - Bestimmte Anzahl von user transactions -> Summe der Einkäufe jeden Users innerhalb einer bestimmten Periode kann berechnet werden
    - Erstellen eines Klassifikationsmodells um User einzuteilen in solche die eine hohe Tendenz zu Kaufen haben und solche die es nicht haben

4.7 Schnittstellen

- Weiter gehende Lösungsansätze zum WUM sind datenbankbasiert
- erforderlich, um eine effiziente und skalierbare Verwaltung der riesigen Datenmengen sowie flexible und interaktive Auswertungen zu ermöglichen
- Kopplung geschieht am besten durch Integration der Daten im Rahmen eines Data Warehouse, auf dem dann die Auswertungen erfolgen
- Bietet die Möglichkeit der Verbindung von WUM → CRM
- E-commerce Daten enthalten produktorientierte Events
  - Warenkorbänderungen
  - Bestellinformationen
  - Impressions (User besucht Seite die für ihn wichtiges Objekt enthält)
  - Click-troughs (User klickt auf dieses Objekt)
  - ...
- Gewünschte clickstream-Daten werden zusammengefügt und bestimmten Events zugeordnet ,sog. „event models“
- Diese Daten werden dann in einem data warehouse, sog. „e-commerce Data Mart“ gespeichert

4.7 Schnittstellen

- Dabei handelt es sich um multidimensionale Datenbanken, die Daten aus verschiedenen Quellen und verschiedenen Aggregationszuständen enthält
- Sie werden als Primärquelle für OLAP (Online Analytical Processing) verwendet, für die visualisierung von Daten und verschiedene Data Mining Aufgaben
- Anwendungsbeispiele:
  - Wert von Einkäufen,
  - Durschnittsgröße der Einkaufswagen,
  - Anzahl verschiedener gekauften Objekte,
  - Anzahl verschiedener Kategorien aus denen gekauft wurde
  - Reaktionen auf Empfehlungen
  - ....

4.8 Probleme

- Während des Usage Mining Prozesses können Probleme auftreten, die verschiedene Ursachen haben
  - Session-Identifikation
  - CGI-Daten
  - Caching
  - Dynamische Seiten
  - Robots Erkennung und Filterung

4.8 Probleme

- Session-identifikation
  - Problem:
    - Proxy Server → Eine Ip Adresse, mehrere Benutzer
    - Anonymisierungstool → mehrere Ip Adressen / Eine Session
  - Lösung:
    - Cookies
    - Registrierung/ Login
- CGI Daten
  - Problem:
    - Versteckte Werte: mittels “hidden” option werden beim POST request Name/Wert- Paar aus der URI entfernt
  - Lösung
    - HTTP traffic überwachen
    - Access log erstellen

4.8 Probleme

- **Caching**
  - Problem
    - Vor-/zurück-Button → gecachte Seite wird angezeigt und nicht nochmal vom Server angefordert
  - Lösung
    - Pfadvervollständigung
- **Robots**
  - Problem
    - Logfiles bestehen manchmal bis zu 50% aus Enträgen von Besuchen von Crawlern
    - Identifikation
  - Lösung
    - Bekannte Suchmachinencrawler können meist Identifiziert und entfernt werden
    - “wohlerzogene” Crawler versuchen erst die “robots.txt” aufzurufen
    - Nicht “wohlerzogene” Crawler werden mittels heuristischen Methoden erkannt (typisches Crawler verhalten ↔ typisches User verhalten)

5 Tools

- Web Analyics Tools dienen der Sammlung und Auswertung des Verhaltens von Besuchern auf Websites.

- Google Analytics
- Webalizer
- Piwik
- AWStats
- Weitere:
  - etracker Web Analytics
  - Analog

5 Tools

- Google Analytics
  - Das mit Abstand meistverwendete Web Analytics Werkzeug
  - Bietet bekannten Funktionen wie Herkunft der Besucher, Verweildauer und Suchbegriffe in Suchmaschinen
  - Erlaubt eine Integration in die Benutzeroberfläche von Google AdWords → bessere Erfolgskontrolle von AdWords-Kampagnen
  - Identifizierung durch Einbettung eines JavaScript Codes
  - Der Zugang ist zurzeit auf die Analyse von 50 Webseiten pro Nutzer beschränkt.
  - Probleme mit dem Datenschutz

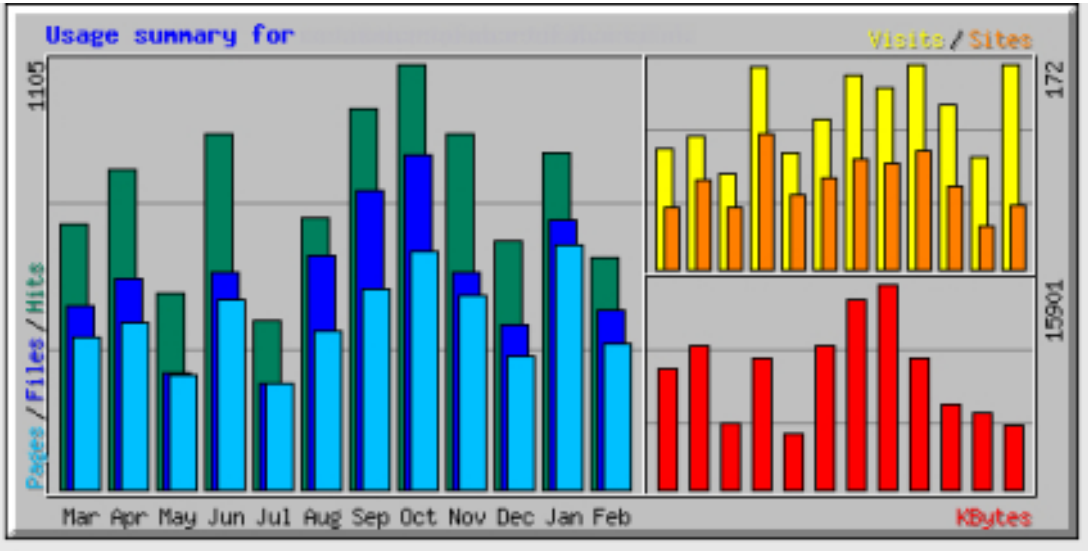
5 Tools


Google Analytics
My Account | Help | Contact Us | Sign Out
Dashboard
Saved Reports
Visitors
Traffic Sources
Content
Goals
Settings
Email
Help Resources
About this Report
Conversion University
Common Questions
Report Finder
Beta Feedback
Dashboard
Export
Email
Apr 1, 2007 - Apr 30, 2007
Avg. Time on Site
00:03:20
April 3, 2007
Avg. Time on Site: 00:02:27
00:01:40
April 2, 2007
April 9, 2007
April 16, 2007
April 23, 2007
Site Usage
16,635 Visits
28,827 Pageviews
1.73 Pages/Visit
00:02:05 Avg. Time on Site
69.56% Bounce Rate
65.52% % New Visits
Visitors Overview
900
450
11,916 Visitors
view report
Traffic Sources Overview
Referring Sites
6,785 (40.79%)
Search Engines
6,276 (37.73%)
Direct
3,567 (21.44%)
Other
7 (0.04%)
view report
Goals Overview
60
30
709 Goal Conversions
view report
Map Overlay
view report

5 Tools

- Webalizer
  - Arbeitet mit Logdateianalyse
  - Anfragen, Besuche, Verweise, Länder der Besucher und Menge der ausgelieferten Daten.
  - grafische als auch textuelle Betrachtung möglich und wird auf unterschiedlichen Zeitskalen (Stunden, Tage, Monate, Jahre) dargestellt.
  - Erstellt Jahresüberblick und eine detaillierte Monatsauswertung
  - Kritik: kein Unterschied zwischen User und Crawler
  - OpenSource
  - Aktuelle Version 2.21-02

5 Tools

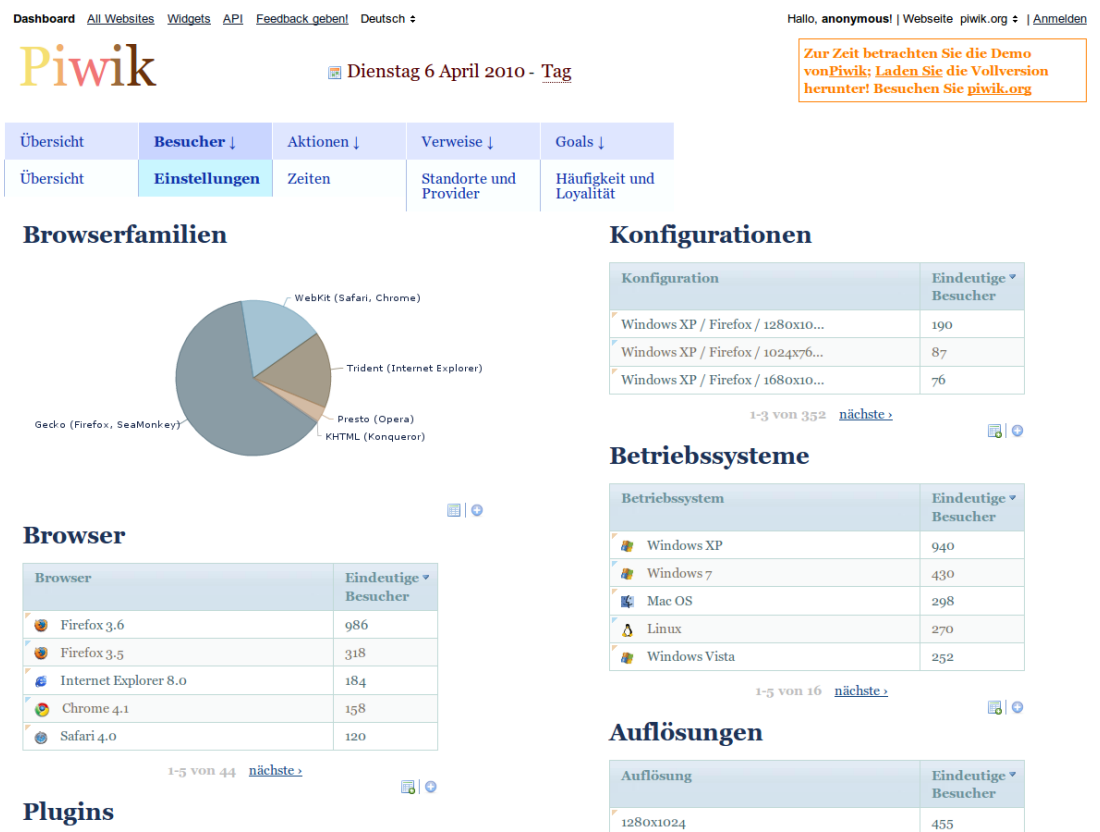


Summary by Month

Month	Daily Avg				Monthly Totals					
	Hits	Files	Pages	Visits	Sites	KBytes	Visits	Pages	Files	Hits
Feb 2007	21	16	13	6	54	4982	172	378	466	601
Jan 2007	28	22	20	3	36	5937	94	632	698	875
Dec 2006	20	13	11	4	70	6633	137	345	429	646
Nov 2006	30	18	16	5	100	10209	171	502	561	925
Oct 2006	35	27	19	4	89	15901	152	617	866	1105
Sep 2006	32	25	17	5	92	14733	163	518	773	988
Aug 2006	22	19	13	4	77	11075	125	410	608	705
Jul 2006	14	8	8	3	62	4356	98	276	277	438
Jun 2006	30	18	16	5	114	10142	169	493	565	920
May 2006	16	9	9	2	52	5074	80	297	300	511
Apr 2006	27	18	14	3	75	11049	112	433	545	829
Mar 2006	22	15	12	3	52	9304	101	396	474	688
Totals						**109395**	**1574**	**5297**	**6562**	**9231**

5 Tools

- Piwik
  - Installation auf dem eigenen Server
  - Benötigt PHP und MySQL Datenbank
  - Plugin basierend → Erweitung möglich (bereitgestellte /eigene Plugins)
  - Informationsanzeige in Echtzeit
  - „openSource Alternative zu GoogleAnalytics“
  - Akt Version:
    0.6.2 (28. Mai 2010)



5 Tools

- AWStats (Version 6.95 (25. Oktober 2009))
  - generiert aus den Logdateien eine grafische und textbasierte Statistik
  - die Grafiken werden durch HTML-Tabellen und CSS-Konstrukte simuliert
  - Muss Aufgerufen werden (z.B. Cronjob)
  - Anschliessende Analyse der Logfiles, neue Informationen werden dem Datenbestand hinzugefügt
  - In Pearl geschrieben
  - OpenSource
  - Kann als Hintergrundprogramm
    - nach einmaligem Aufruf werden statische HTML-Dateien erzeugt
  - oder als CGI Programm laufen
    - nach einmaligem Aufruf werden statische HTML-Dateien erzeugt
    - individuelle Anfragen von Besuchern zulassen (z.B. Übersicht der verwendeten Browser zwischen März 2007 und April 2009)
    - Analysedaten werden dann unmittelbar bei Anforderung generiert
    - führt zwangsläufig zu einer höheren Serverlast

5 Tools

Statistik für:
destailleur.fr

Zusammenfassung
Wann:
Monatliche Historie
Tage im Monat
Wochentage
Stunden (Serverzeit)
Wer:
Länder
- Gesamte Liste
- Regions
- Cities

Rechner
- Gesamte Liste
- Letzter Zugriff
- Unaufgelöste IP Adressen

beglaubigte Benutzer
- Gesamte Liste
- Letzter Zugriff

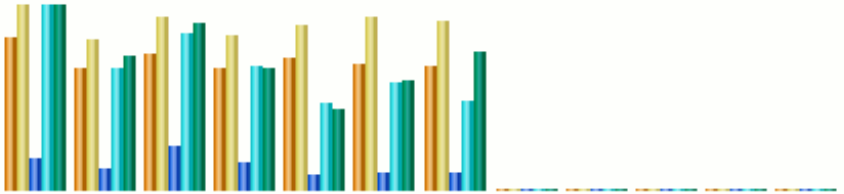
Robots/Spiders (Suchmaschinen)
- Gesamte Liste
- Letzter Zugriff

Navigation:
Aufenthaltsdauer
Datei-Typen
Zugriffe
- Gesamte Liste
- Einstiegsseiten
- Exit Seiten

Betriebssysteme
- Versionen
- Unbekannt

* Nicht gesehener Traffic ist Traffic, welcher von Robots, Würmern oder Antworten mit speziellem HTTP-Statuscode

Monatliche Historie



Monat	Unterschiedliche Besucher	Anzahl der Besuche	Seiten	Zugriffe	Bytes
Jan 2007	1782	2157	19156	108283	2.02 GB
Feb 2007	1424	1760	12684	71225	1.47 GB
März 2007	1597	2036	25299	91878	1.84 GB
Apr 2007	1433	1817	16327	72609	1.33 GB
Mai 2007	1550	1923	8780	50753	904.93 MB
Juni 2007	1468	2030	10411	63760	1.21 GB
Juli 2007	1458	1970	9666	52765	1.51 GB
Aug 2007	0	0	0	0	0
Sep 2007	0	0	0	0	0
Okt 2007	0	0	0	0	0
Nov 2007	0	0	0	0	0
Dez 2007	0	0	0	0	0
Total	10712	13693	102323	511273	10.25 GB

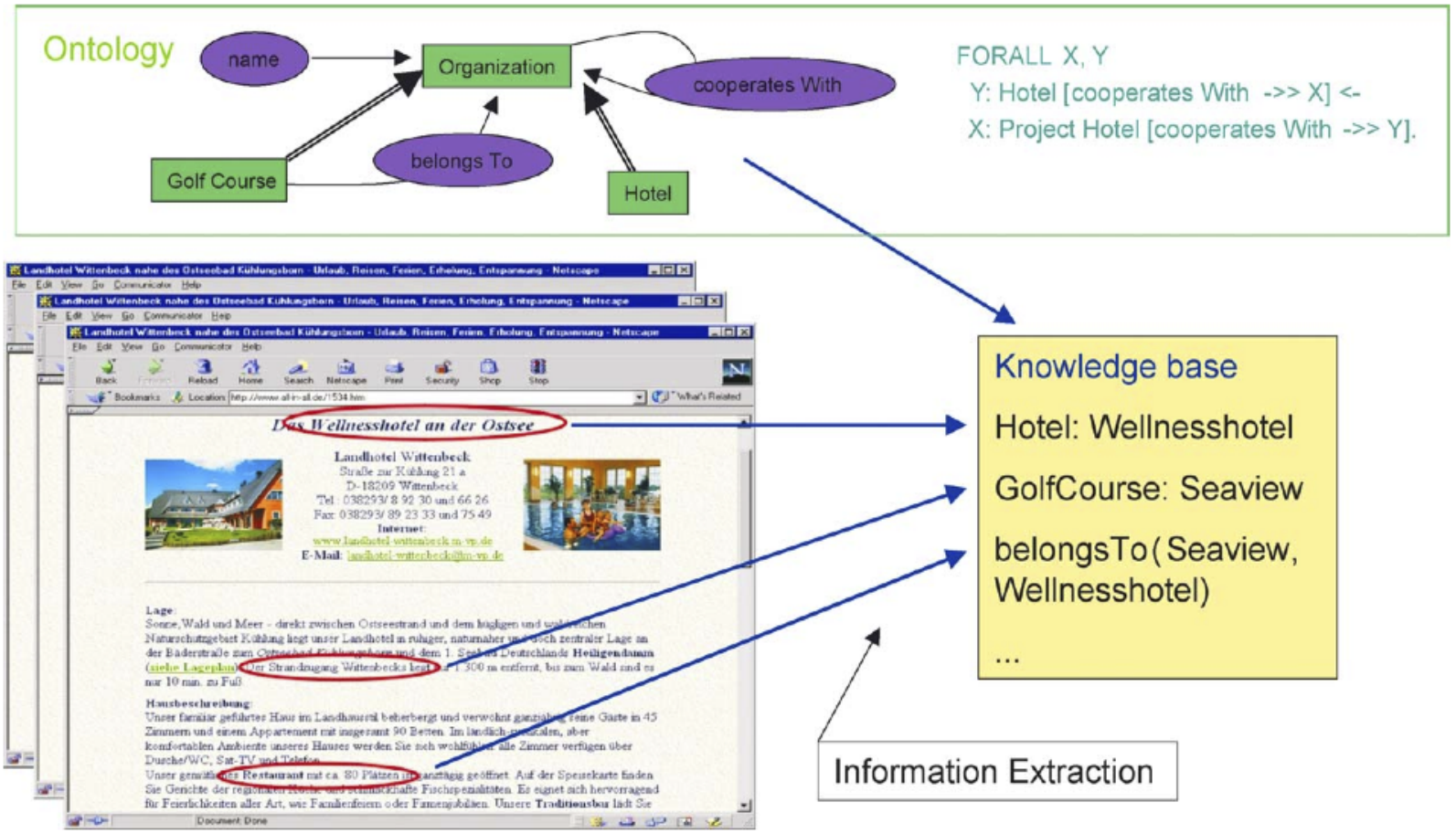
6 Ausblick

[FUT]

- Erweiterung der “3 Achsen“ um Ontologien
- → Semantic Web Mining
- Web of Knowledge → Web of Meaning
- Abbildung von Ontologien
  - Topic Maps
  - RDF/RDF(S)
  - OWL
- Kombination von Web Mining Techniken und Ontologien um Semantisches “Wissen“ zu extrahieren
- Erste Ansätze in Bing, Wolfram Alpha ....
- Ansätze zur Ontologieextraktion meist semi-automatisch
- Web Mining könnte helfen den Prozess zu verbessern
- Nutzung bestehender Konzeptualisierungen als Ontologien und Annotation von Webdaten

[SW]

6 Ausblick



[SW]

6 Ausblick

- **Weitere Forschungsgebiete:**
  - Temporale Entwicklung des Web → Archive.org
  - Web Metriken → Methoden zur Bewertung von Webseiten(Inhalten)
  - Optimierung von Web Services
  - Bessere Kombination des Content und Structure Mining

Danke

Vielen Dank für die Aufmerksamkeit !
Fragen ?

Quellen

[LIU] **Bing Liu. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data**
(Data-Centric Systems and Applications). Springer, 2 ed. 2008

[RAHM] **Web & Datenbanken. Konzepte, Architekturen, Anwendungen**
E Rahm, G Vossen - Verlag dpunkt, Heidelberg, 2003

[HWM] **Handbuch Web Mining im Marketing: Konzepte, Systeme, Fallstudien**
Hajo Hippner, Melanie Merzenich, Klaus Wilde, Vieweg - 2002

[SRIV] **Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data**
Jaideep Srivastava et. al

[SW] **Semantic Web Mining State of the art and future directions**
Stumme, Gerd et. al

[IDC] **The Digital Universe Decade - Are You Ready?**
John Gantz and David Reinsel, Mai 2010

Internet

[W3C] http://www.w3.org/Daemon/User/Config/Logging.html, Zugegriffen am 26.05.2010

[FUT] http://blog.northstarmanifesto.com/wp-content/uploads/2008/12/the-future.jpg, Zugegriffen am 30.05.2010

[SOC] http://socialnomics.net/2009/08/11/statistics-show-social-media-is-bigger-than-you-think/, Zugegriffen am 12.05.2010

[WIKI] Wikipedia, http://www.wikipedia.de

[GR] http://www9.org/w9cdrom/160/160.html

[WBT] http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter9/sld009.htm, Zugegriffen am 10.04.2010

[FKR] Web Mining – Data Mining im Internet,Johannes Fürnkranz
http://www.ke.tu-darmstadt.de/lehre/archiv/ss06/web-mining/wm-intro.pdf

[KIETZ] Data Mining zur Wissensgewinnung aus Datenbanken, Dr. Jörg-Uwe Kietz
http://www.kietz.ch/DataMining/Vorlesung/folien/13-WEB.pdf

[MBG] Electronic Retailing - Marketinginstrumente und Marktforschung in Internet, M.Madlberger
http://books.google.de/books?id=OJN3GtG3nhYC&pg=PA228&dq= web+content+mining &cd=4#v=onepage&q=web%20content%20mining&f=true