# Web Usage Mining

# for E-Business Applications

**ECML/PKDD-2002 Tutorial, 19 August 2002**

Bettina Berendt, Bamshad Mobasher, Myra Spiliopoulou



## The Presenters

**Bettina Berendt**

- **Institute of Information Systems, Humboldt University Berlin, Berlin, Germany**

**Bamshad Mobasher**

- **Department of Computer Science, DePaul University, Chicago, USA**

**Myra Spiliopoulou**

- **Department of E-Business, Leipzig Graduate School of Management, Leipzig, Germany**

## Agenda

- Introduction
- Data Acquisition and Data Preparation
- Evaluation of Web Site Success
- Applications and KDD Techniques for them
- Privacy Concerns
- Research Issues and Future Directions



## What is so particular about Web Usage Mining?

**The generic Knowledge Discovery circle:**

- Problem specification
- Data collection
- Data preparation
- Data mining
- Presentation of the results
- Evaluation and Interpretation of the results
- Action upon the results

**holds for Web Usage Mining as well.**

## What is so particular about Web Usage Mining?

- Problem specification
- Data collection

Data in Web Usage Mining:
- Web server logs
- Site contents
- Data about the visitors, gathered from external channels
- Further application data

Not all these data are always available.
When they are, they must be integrated.



## What is so particular about Web Usage Mining?

- Problem specification
- Data collection
- Data preparation

The quality of Web server data varies considerably.

Their integration with data from other sources is difficult.

## What is so particular about Web Usage Mining?

- Problem specification
- Data collection
- Data preparation
- Data mining

The data being mined are records, sets of records and sequences of records.
There are conventional mining techniques that can process such data types.



## What is so particular about Web Usage Mining?

- Problem specification
- Data collection
- Data preparation
- Data mining

Some Web applications call for a particular analysis of the data. There are dedicated techniques to deal with them.
For many Web applications, general purpose techniques are sufficient.

## What is so particular about Web Usage Mining?

- Problem specification
- Data collection
- Data preparation
- Data mining
- Presentation of the results

There are conventional presentation tools, designed to display the results of general purpose mining techniques.

There are dedicated tools for some Web applications, e.g. for the visual inspection of site traffic and of pages accessed together.



## What is so particular about Web Usage Mining?

- Problem specification
- Data collection
- Data preparation
- Data mining
- Presentation of the results
- Evaluation and Interpretation of the results
- Action upon the results

Obviously, dependent on the Web application.

## What is so particular about Web Usage Mining? Web applications

There are three generic types of Web applications:

- Revolutionary applications: They have emerged with the Web and have no counterpart in the pre-Web era.
- Innovative applications: They have emerged with Information Technology. The capabilities and particularities of the Web have a major impact on them.
  - e-learning
- Web-empowered conventional applications: They were transferred in the Web context; the Web revolutionized the way of doing them.
  - marketing of products
  - literature search
  - imaging and public relations



## What is so particular about Web Usage Mining? Applications in the Web

Conventional applications have:

- well-known processes
- well-known evaluation methods

The impact of the Web:

- It offers new ways of performing an application
  - marketing: acquiring information about a product interactively
  - sales: recommending products

  without the involvement of a salesperson
- It demands new ways of evaluating how good an application is.
  - marketing impact of a Web site
  - effect of recommendations
- It increased the competition and, indirectly, the need for fast and effective evaluation.

## What is so particular about Web Usage Mining?

- Problem specification
- Data collection
- Data preparation
- Data mining
- Presentation of the results
- Evaluation and Interpretation of the results
- Action upon the results

Web applications



## Agenda

- Introduction
- Data Acquisition and Data Preparation
- Evaluation of Web Site Success
- Applications and KDD Techniques for them
- Privacy Concerns
- Research Issues and Future Directions

## Web Usage Mining

**Discovery of meaningful patterns from data generated by client-server transactions on one or more Web servers**

**Typical Sources of Data**

- **automatically generated data stored in server access logs, referrer logs, agent logs, and client-side cookies**
- **e-commerce and product-oriented user events (e.g., shopping cart changes, ad or product click-throughs, etc.)**
- **user profiles and/or user ratings**
- **meta-data, page attributes, page content, site structure**

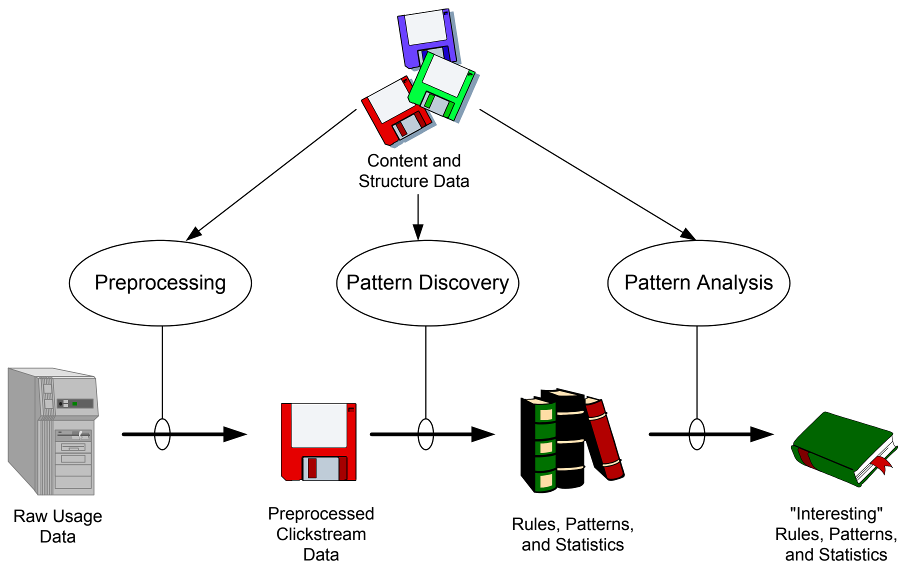


## What's in a Typical Server Log?

<ip_addr><base_url> - <date><method><file><protocol><code><bytes><referrer><user_agent>

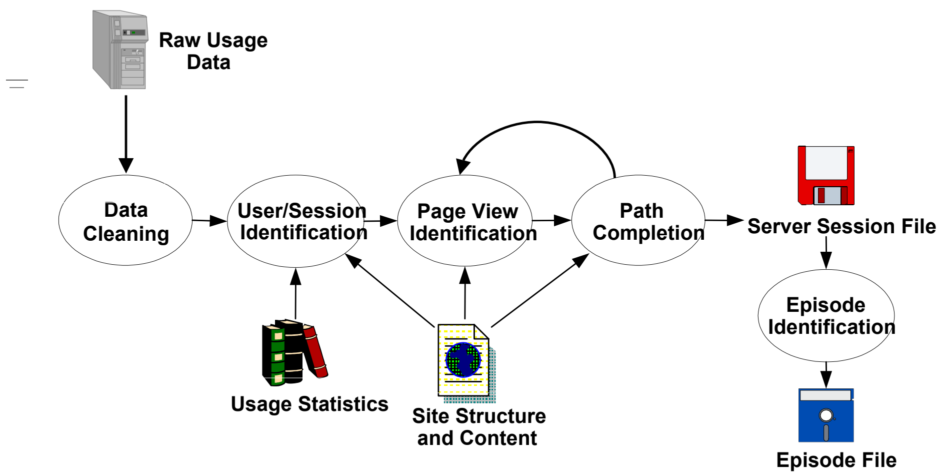
```
203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:21 -0600] "GET /Calls/OWOM.html
HTTP/1.0" 200 3942 "http://www.lycos.com/cgi-
bin/pursuit?query=advertising+psychology&maxhits=20&cat=dir" "Mozilla/4.5 [en] (Win98; I)"

203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:23 -0600] "GET
/Calls/Images/earthani.gif HTTP/1.0" 200 10689 "http://www.acr-news.org/Calls/OWOM.html"
"Mozilla/4.5 [en] (Win98; I)"

203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:24 -0600] "GET /Calls/Images/line.gif
HTTP/1.0" 200 190 "http://www.acr-news.org/Calls/OWOM.html" "Mozilla/4.5 [en] (Win98; I)"

203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:25 -0600] "GET /Calls/Images/red.gif
HTTP/1.0" 200 104 "http://www.acr-news.org/Calls/OWOM.html" "Mozilla/4.5 [en] (Win98; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:31 -0600] "GET / HTTP/1.0" 200 4980 ""
"Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:35 -0600] "GET /Images/line.gif
HTTP/1.0" 200 190 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:35 -0600] "GET /Images/red.gif
HTTP/1.0" 200 104 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:35 -0600] "GET /Images/earthani.gif
HTTP/1.0" 200 10689 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:33:11 -0600] "GET /CP.html HTTP/1.0" 200
3218 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"
```

## The Web Usage Mining Process






## Preprocessing of Web Usage Data [CMS99]



## Data Preprocessing (I)

Data cleaning

- remove irrelevant references and fields in server logs
- remove references due to spider navigation
- remove erroneous references
- add missing references due to caching (done after sessionization)

Data integration

- synchronize data from multiple server logs
- integrate e-commerce and application server data
- integrate meta-data (e.g., content labels)
- integrate demographic / registration data



## Data Preprocessing (II)

Data Transformation

- user identification
- sessionization / episode identification
- pageview identification
  - a pageview is a set of page files and associated objects that contribute to a single display in a Web Browser

Data Reduction

- sampling and dimensionality reduction (ignoring certain pageviews / items)

Identifying User Transactions (i.e., sets or sequences of pageviews possibly with associated weights)

## Why sessionize?

- Quality of the patterns discovered in KDD depends on the quality of the data on which mining is applied.
- In Web usage analysis, these data are the sessions of the site visitors: the activities performed by a user from the moment she enters the site until the moment she leaves it.
- Difficult to obtain reliable usage data due to proxy servers and anonymizers, dynamic IP addresses, missing references due to caching, and the inability of servers to distinguish among different visits.
- Cookies and embedded session IDs produce the most faithful approximation of users and their visits, but are not used in every site, and not accepted by every user.
- Therefore, *heuristics* are needed that can sessionize the available access data.



## Sessionization strategies: Cookies, session IDs, heuristics

**Session reconstruction =**
correct mapping of activities to different individuals +
correct separation of activities belonging to different visits of the same individual

| While users navigate the site: identify ... | | In the analysis of log files: identify … | | Resulting partitioning of the log file |
|---|---|---|---|---|
| users by | sessions by | users by | sessions by | |
| — | — | IP & Agent | sessionization heuristics | constructed sessions (“**u-ipa**”) |
| cookies | — | — | sessionization heuristics | constructed sessions (“**cookies**”) |
| cookies | embedded session IDs | — | — | real sessions |

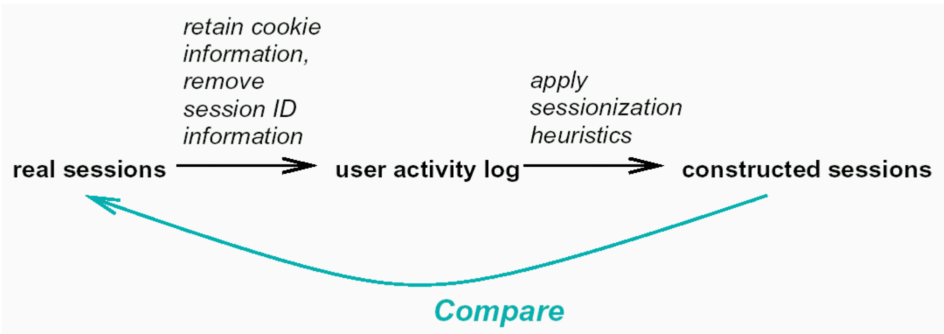
## Mechanisms for User Identification

| Method | Description | Privacy Concerns | Advantages | Disadvantages |
|---|---|---|---|---|
| IP Address + Agent | Assume each unique IP address/Agent pair is a unique user | Low | Always available. No additional technology required. | Not guaranteed to be unique. Defeated by rotating IPs. |
| Embedded Session Ids | Use dynamically generated pages to associate ID with every hyperlink | Low to medium | Always available. Independent of IP addresses. | Cannot capture repeat visitors. Additional overhead for dynamic pages. |
| Registration | User explicitly logs in to the site. | Medium | Can track individuals not just browsers | Many users won't register. Not available before registration. |
| Cookie | Save ID on the client machine. | Medium to high | Can track repeat visits from same browser. | Can be turned off by users. |
| Software Agents | Program loaded into browser and sends back usage data. | High | Accurate usage data for a single site. | Likely to be rejected by users. |

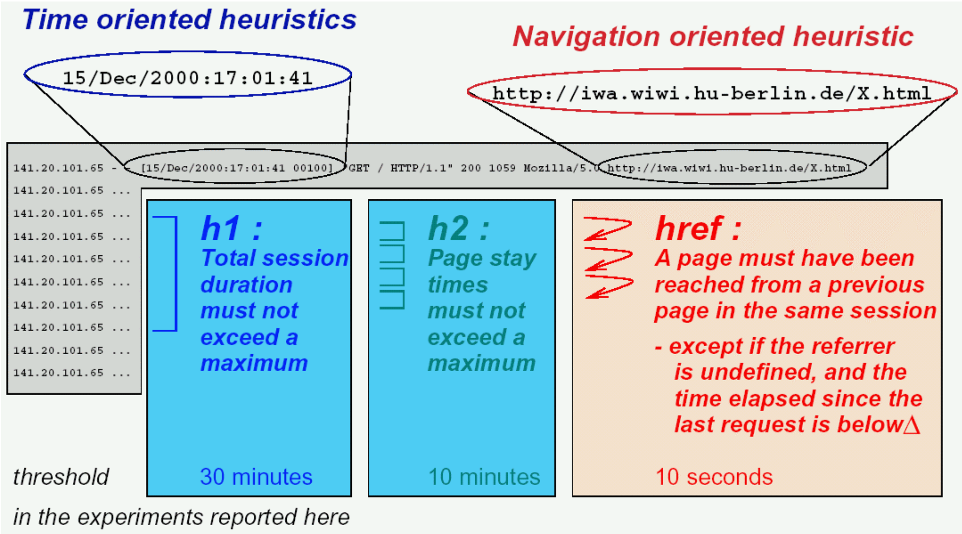


## Sessionization strategies: Real and constructed sessions

| identify ... users by | sessions by | identify ... users by | sessions by | resulting log partioning |
|---|---|---|---|---|
| cookies | — | — | sessionization heuristics | constructed sessions |
| cookies | session IDs | — | — | real sessions |



## Sessionization strategies: Sessionization heuristics




**(Heuristics used in, e.g., [CMS99, SF99], formalized in [BMSW01])**



## Sessionization strategies: The "undefined referrer" problem

**An undefined referrer ("–" in the log) may occur, for example,**

- **as the referrer of the start page, or after a brief excursion to another server,**
- **as the referrer of a typed-in or bookmarked URL,**
- **when a frameset page is reloaded in mid-session,**
- **for all these pages, when they are reached via the back button,**
- **in a frame-based site: as the referrer of the first frames that are loaded when the start page containing the top frameset is requested,**

**→ More occurrences of undefined referrers in frame-based sites.**
**→ Special treatment of undefined referrers only in heuristic href.**
**→ Frameset loading may also cause problems for temporal heuristics.**

**⇒ Expectation: The performance of the heuristics, in particular that of href, will differ between frame-based and frame-free sites.**

**[BMSW01, BMNS02]**

## Measuring reconstruction quality: Sessionization accuracy

**A heuristic *h* maps entries** in the log *L* into elements of constructed sessions, such that

- each entry of *L* is mapped to exactly one element of a constructed session,
- the mapping is order-preserving.

**Measures** quantify the successful mappings of real sessions to constructed sessions:

- a measure *M* evaluates a heuristic *h* based on the differences between the set of constructed sessions of this heuristic $C_h$, and the set of real sessions *R*,
- each measure assigns to *h* a value $M(h) \in [0; 1]$ such that the perfect heuristic would have $M(h^*) = 1$.

[BMSW01, BMNS02]



## Measuring reconstruction quality: Types of measures

**Categorical measures** are based on the number of real sessions that are completely reconstructed by the heuristic. A real session is *completely reconstructed* if all its elements are contained in the same constructed session, with no intervening foreign elements.

The base categorical measure $M_{cr}(h)$ is the ratio of the number of completely reconstructed real sessions in $C_h$ to the total number of real sessions $|R|$.

**Gradual measures** are based on the degree to which the real sessions are reconstructed by the heuristic.

These measures consider the number of elements in the intersection of a real and a constructed session, and they aggregate over all sessions.

## Measuring reconstruction quality: The measures used

**Derived categorical measures** consider the location of a real session within the (unique) constructed session that completely reconstructs it.

- complete reconstruction with correct entry page, or with correct exit page
- identical reconstruction (with correct entry *and* exit pages)

*Recall* and *precision* scores are obtained by dividing the number of "correct guesses" by $|R|$ or by $|C_h|$.

The **gradual measures** $M_o(h)$ and $M_s(h)$ are aggregates, over all sessions, of two measures of partial overlap:

- Degree of overlap between a real session *r* and a constructed session *c*: $|c \cap r| / |r|$.
- Degree of similarity: $|c \cap r| / |c \cup r|$.



## Measuring reconstruction quality: Which measures?

The choice of measures depends on the goals of usage analysis, for example:

**Categorical measures** are useful if sessions in their entirety are of interest, including sequential order
*Example*: analyses of navigation behavior

**Derived categorical** measures are useful if, in addition, entry points or exit points are of interest
*Example*: analysis for site redesign

**Gradual measures** are useful if entire sessions, and the order of access, is less important
*Examples*: page prefetching, market basket analysis, recommender systems

## Experimental evaluation

## Data and measures

**The test environment and data:**

- Logs of two versions of the same university site were investigated:
  - frame-based site: 174660 requests,
  - frame-free site: 115434 requests.
- Data preprocessing: removed robot accesses, accesses without cookies (1 - 2 %)

**The measures:**

1. Base statistics: number of sessions, session duration and length, page stay time
2. Measures of the accuracy of content reconstruction

[BMNS02]

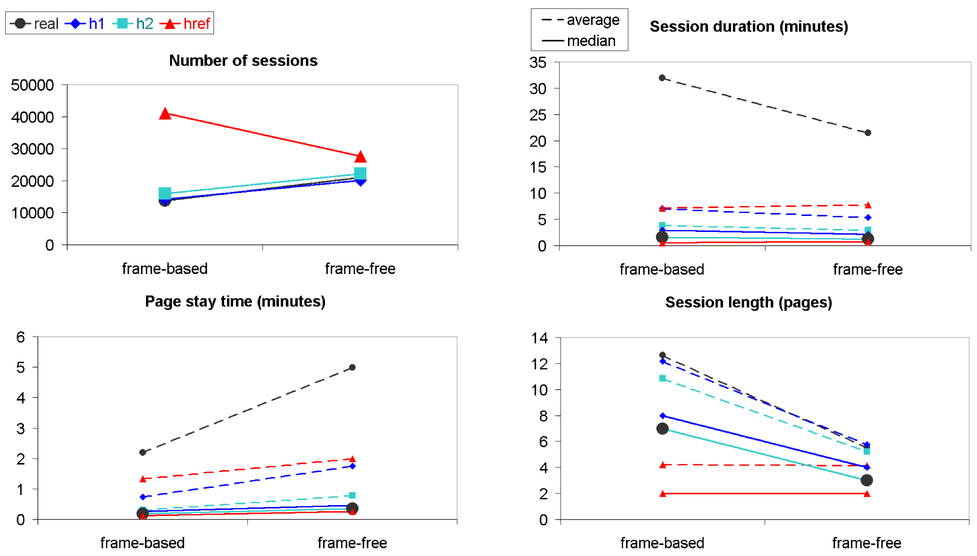
## The user environment

- The frame free site had 5446 cookies (= users), 6849 IPs, 8409 IP+agent
- 77.38% of users have unique IP, 96.49% of users have unique agent
- 75.98% of users have unique IP+agent
- < 5% of real sessions contained multiple IP+agent combinations
- ⇒ IP+agent could be quite effective for analysis at session level; problems may arise for analysis at user level
- 86.98% of IPs used by only one user, 92.02% of IP+agent used by only one user
- *simultaneous* access from different users with same IP+agent: ≃ 1% of sessions

⇒ Our logs present very good conditions for analysis; one IP+agent corresponds to one cookie in a large majority of cases.

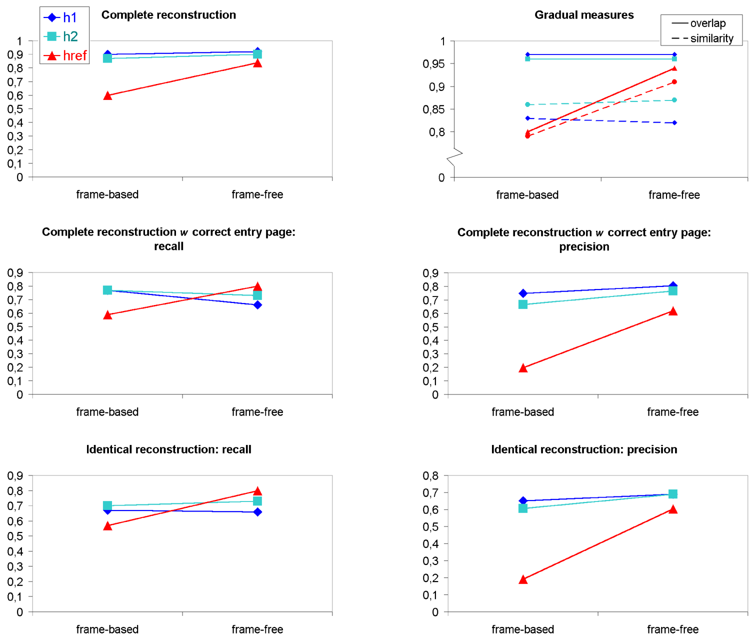


## The impact of site structure on base statistics




⇒ In both site versions: many short sessions, a few very long sessions; medians are approximated quite well; **href** generates *many* short sessions

## The impact of site structure on sessionization accuracy





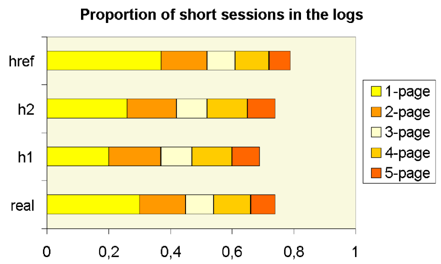

## The impact of session length on sessionization accuracy

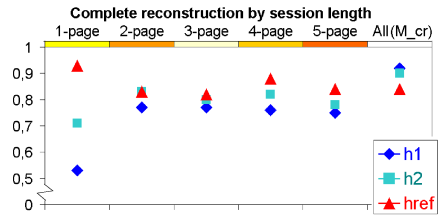
**Most of the real sessions are shorter than the average.**

⇒ **How effective is each heuristic in reconstructing the short sessions?**

**In the frame-free site, we found**

- **that the proportions of short sessions are high,**
- **that href is particularly successful in their reconstruction.**



## Which sessionization heuristic? Measure-based answers

**The choice of sessionization heuristic depends on the characteristics of the data and the goals of analysis:**

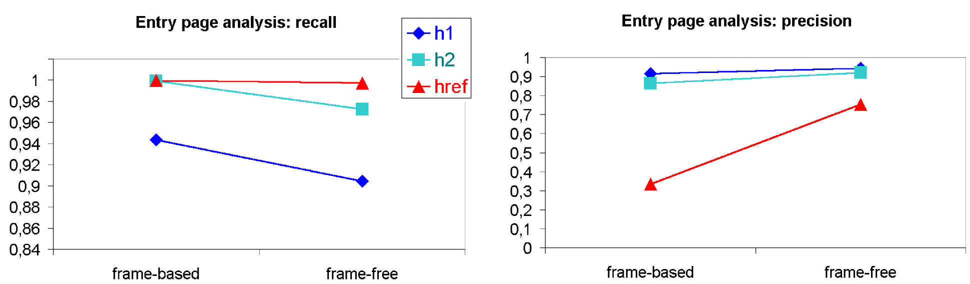
- **href good for reconstructing many short sessions, but overall, h1 and h2 are more robust.**
- **If individual users visit the site in short but temporally dense sessions, h2 may perform better than h1.**
- **When timestamps are not reliable (e.g., using integrated data across many log files), href may be the best choice.**
- **Referrer-based heuristics tend to perform worse in frame-based sites.**
- **Results of experiments that varied the heuristics' parameters indicate that a combination heuristic of href and h2 may be desirable.**



## Impact on mining applications: Entry/exit page analysis

**Important application:**
**Which pages are seen first (and determine whether user will visit further pages)?**
**Where do people leave the site – if unintended abandoning, need site redesign**

**Recall *(Precision)*: Number of pages correctly classified as entry pages / Number of all entry pages *(Number of all pages classified as entry pages)***




**(Results were virtually identical for exit pages.)**

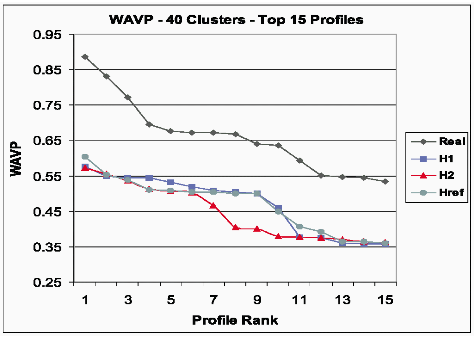
**Again, results depend on details of analysis question: much better precision scores, in particular for href, for top 10 (20) entry/exit pages.**

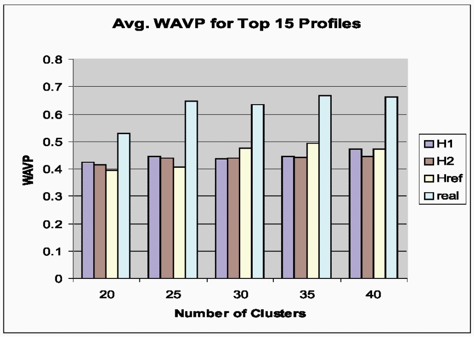
## Impact on mining applications: Page prediction / recommendation – procedure

- **Application example personalization: recommend pages based on co-occurrences in previous visitors' sessions**
- **In the frame-free site, we determined co-occurrences using PACT (Profile Aggregations based on Clustering Transactions):**
  - **transactions (sessions) expressed as vectors of $\langle p, weght \rangle$ pairs, here: *weight* = 1 if page was visited in session**
  - **cluster using *k*-means, threshold: use only pages visited in at least 70% of sessions in the cluster → cluster profile**
- **We measured predictive power by WAVP (weighted average visit percentage): likelihood that a user who visits any page in a given profile will visit the rest of the pages in that profile during the same session [MDLN02]**
- **To test prediction quality of reconstructed sessions, we compared with baseline defined by profiles based on real sessions.**



## Impact on mining applications: Page prediction / recommendation – results




⇒ **Results indicate that for prediction, href and h1 perform rather well.**

**General observation:**
**Application-based answers to the question "Which sessionization heuristic?" are similar to measures-based answers.**

## Sessionization strategies revisited

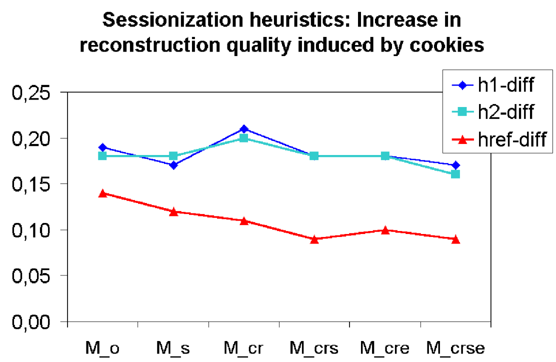
**Session reconstruction =**
**correct mapping of activities to different individuals +**
**correct separation of activities belonging to different visits of the same individual**

| While users navigate the site: identify ... | | In the analysis of log files: identify … | | Resulting partitioning of the log file |
|---|---|---|---|---|
| users by | sessions by | users by | sessions by | |
| — | — | IP & Agent | sessionization heuristics | constructed sessions (“**u-ipa**”) |
| cookies | — | — | sessionization heuristics | constructed sessions (“**cookies**”) |
| cookies | embedded session IDs | — | — | real sessions |



## The impact of cookies on sessionization accuracy

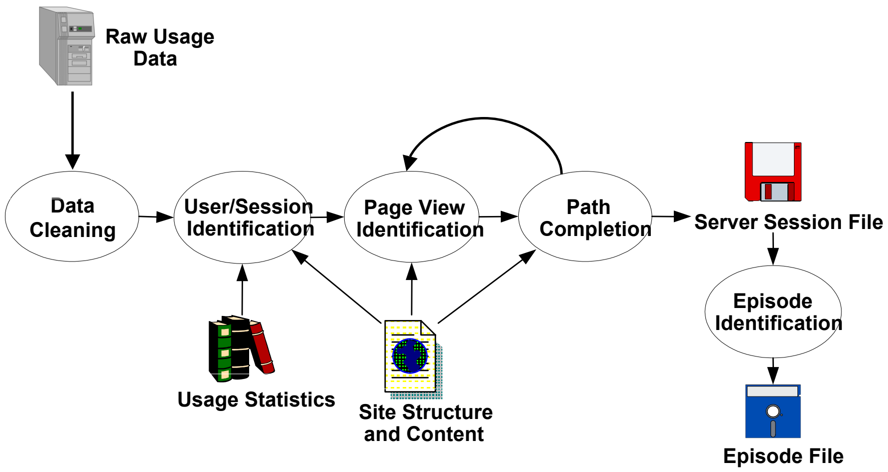
We compared constructed sessions based on logs with cookie information (but stripped of session ID information) with constructed sessions based on logs that were stripped of session ID and of cookie information (“**u-ipa**”). The difference between heuristic performance in the **cookie** setting and the **u-ipa** setting allowed us to measure the gain in reconstruction quality induced by cookies.




[SMBN03]

(From left to right: overlap, similarity, complete reconstruction, recall values for (a) complete reconstruction with complete entry page, (b) exit page, (c) identical reconstruction)

## Preprocessing of Web Usage Data [CMS99]






## Path Completion

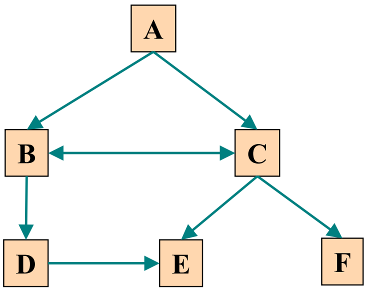
Refers to the problem of inferring missing user references due to caching.

Effective path completion requires extensive knowledge of the link structure within the site

Referrer information in server logs can also be used in disambiguating the inferred paths.

Problem gets much more complicated in frame-based sites.

## Sessionization Example




| 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 | - | IE4;Win98 |
| 0:12 | 2.3.4.5 | B | C | IE4;Win98 |
| 0:15 | 2.3.4.5 | E | C | IE4;Win98 |
| 0:19 | 1.2.3.4 | C | A | IE5;Win2k |
| 0:22 | 2.3.4.5 | D | B | IE4;Win98 |
| 0:22 | 1.2.3.4 | A | - | IE4;Win98 |
| 0:25 | 1.2.3.4 | E | C | IE5;Win2k |
| 0:25 | 1.2.3.4 | C | A | IE4;Win98 |
| 0:33 | 1.2.3.4 | B | C | IE4;Win98 |
| 0:58 | 1.2.3.4 | D | B | IE4;Win98 |
| 1:10 | 1.2.3.4 | E | D | IE4;Win98 |
| 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 | IE4;Win98 |
| 1:25 | 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 |



## Sessionization Example

1. Sort users (based on IP+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 | - | IE4;Win98 |
| 0:12 | 2.3.4.5 | B | C | IE4;Win98 |
| 0:15 | 2.3.4.5 | E | C | IE4;Win98 |
| 0:19 | 1.2.3.4 | C | A | IE5;Win2k |
| 0:22 | 2.3.4.5 | D | B | IE4;Win98 |
| 0:22 | 1.2.3.4 | A | - | IE4;Win98 |
| 0:25 | 1.2.3.4 | E | C | IE5;Win2k |
| 0:25 | 1.2.3.4 | C | A | IE4;Win98 |
| 0:33 | 1.2.3.4 | B | C | IE4;Win98 |
| 0:58 | 1.2.3.4 | D | B | IE4;Win98 |
| 1:10 | 1.2.3.4 | E | D | IE4;Win98 |
| 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 | IE4;Win98 |
| 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 |

| | | | | |
|---|---|---|---|---|
| 0:01 | 1.2.3.4 | A | - | IE5;Win2k |
| 0:09 | 1.2.3.4 | B | A | IE5;Win2k |
| 0:19 | 1.2.3.4 | C | A | IE5;Win2k |
| 0:25 | 1.2.3.4 | E | C | IE5;Win2k |
| 1:15 | 1.2.3.4 | A | - | IE5;Win2k |
| 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 |

| | | | | |
|---|---|---|---|---|
| 0:10 | 2.3.4.5 | C | - | IE4;Win98 |
| 0:12 | 2.3.4.5 | B | C | IE4;Win98 |
| 0:15 | 2.3.4.5 | E | C | IE4;Win98 |
| 0:22 | 2.3.4.5 | D | B | IE4;Win98 |

| | | | | |
|---|---|---|---|---|
| 0:22 | 1.2.3.4 | A | - | IE4;Win98 |
| 0:25 | 1.2.3.4 | C | A | IE4;Win98 |
| 0:33 | 1.2.3.4 | B | C | IE4;Win98 |
| 0:58 | 1.2.3.4 | D | B | IE4;Win98 |
| 1:10 | 1.2.3.4 | E | D | IE4;Win98 |
| 1:17 | 1.2.3.4 | F | C | IE4;Win98 |

## Sessionization Example

2. Sessionize using heuristics (*h*1 with 30 min)

| | | | | |
|---|---|---|---|---|
| 0:01 | 1.2.3.4 | A | - | IE5;Win2k |
| 0:09 | 1.2.3.4 | B | A | IE5;Win2k |
| 0:19 | 1.2.3.4 | C | A | IE5;Win2k |
| 0:25 | 1.2.3.4 | E | C | IE5;Win2k |
| 1:15 | 1.2.3.4 | A | - | IE5;Win2k |
| 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 |

| | | | | |
|---|---|---|---|---|
| 0:01 | 1.2.3.4 | A | - | IE5;Win2k |
| 0:09 | 1.2.3.4 | B | A | IE5;Win2k |
| 0:19 | 1.2.3.4 | C | A | IE5;Win2k |
| 0:25 | 1.2.3.4 | E | C | IE5;Win2k |

| | | | | |
|---|---|---|---|---|
| 1:15 | 1.2.3.4 | A | - | IE5;Win2k |
| 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 |

**The *h*1 heuristic (with timeout variable of 30 minutes) will result in the two sessions given above.**



## Sessionization Example

2. Sessionize using heuristics (another example with *href*)

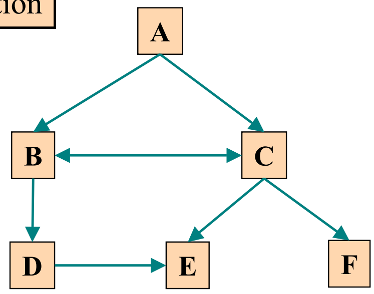
| | | | | |
|---|---|---|---|---|
| 0:22 | 1.2.3.4 | A | - | IE4;Win98 |
| 0:25 | 1.2.3.4 | C | A | IE4;Win98 |
| 0:33 | 1.2.3.4 | B | C | IE4;Win98 |
| 0:58 | 1.2.3.4 | D | B | IE4;Win98 |
| 1:10 | 1.2.3.4 | E | D | IE4;Win98 |
| 1:17 | 1.2.3.4 | F | C | IE4;Win98 |

**In this case, the referrer-based heuristics will result in a single session, while the *h*1 heuristic (with timeout = 30 minutes) will result in two different sessions.**

## Sessionization Example

3. Perform Path Completion

| | | | | |
|---|---|---|---|---|
| 0:22 | 1.2.3.4 | A | - | IE4;Win98 |
| 0:25 | 1.2.3.4 | C | A | IE4;Win98 |
| 0:33 | 1.2.3.4 | B | C | IE4;Win98 |
| 0:58 | 1.2.3.4 | D | B | IE4;Win98 |
| 1:10 | 1.2.3.4 | E | D | IE4;Win98 |
| 1:17 | 1.2.3.4 | F | C | IE4;Win98 |




A=>C , C=>B , B=>D , D=>E , C=>F

**Need to look for the shortest backwards path from E to C based on the site topology. Note, however, that the elements of the path need to have occurred in the user trail previously.**

E=>D, D=>B, B=>C



## Integrating E-Commerce Events

**Either product oriented or visit oriented**

**Not necessarily a one-to-one correspondence with user actions**

**Used to track and analyze conversion of browsers to buyers**

**Major difficulty for E-commerce events is defining and implementing the events for a site**

- **however, in contrast to clickstream data, getting reliable preprocessed data is not a problem**

**Another major challenge is the successful integration with clickstream data**

## Product-Oriented Events

Product View

- Occurs every time a product is displayed on a page view
- Typical Types: Image, Link, Text

Product Click-through

- Occurs every time a user "clicks" on a product to get more information
  - Category click-through
  - Product detail or extra detail (e.g. large image) click-through
  - Advertisement click-through



## Product-Oriented Events

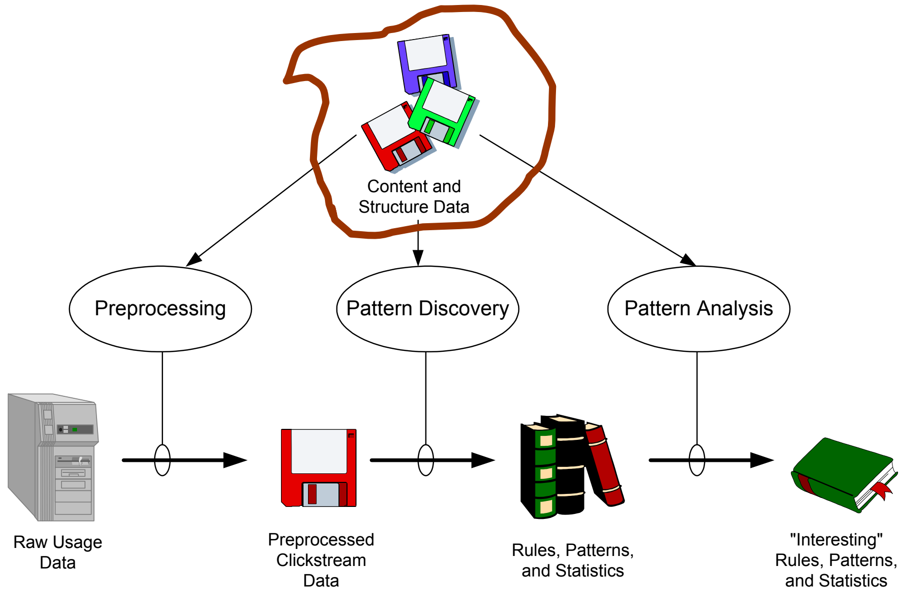
Shopping Cart Changes

- Shopping Cart Add or Remove
- Shopping Cart Change - quantity or other feature (e.g. size) is changed

Product Buy or Bid

- Separate buy event occurs for each product in the shopping cart
- Auction sites can track bid events in addition to the product purchases

## The Web Usage Mining Process






## Why integrate content? (I)

**Basic idea**: associate each requested page with one or more domain concepts, to better understand the process of navigation

*Example: a travel planning site*

From ...

```
p3ee24304.dip.t-dialin.net - - [19/Mar/2002:12:03:51 +0100]
     "GET /search.html?l=ostsee%20strand&syn=023785&ord=asc HTTP/1.0" 200 1759
p3ee24304.dip.t-dialin.net - - [19/Mar/2002:12:05:06 +0100]
     "GET /search.html?l=ostsee%20strand&p=low&syn=023785&ord=desc HTTP/1.0" 200 8450
p3ee24304.dip.t-dialin.net - - [19/Mar/2002:12:06:41 +0100]
     "GET /mlesen.html?Item=3456&syn=023785 HTTP/1.0" 200 3478
```

To ...

**Search by location** → *Refine search* → **Search by location+price** → *Choose item* → **Look at individual hotel**

## Why integrate content? (II)

- This abstraction form details is also needed to describe the behavior of groups of users, I.e., the patterns sought for in mining
- An example is

  "People who buy jackets tend to buy shoes."

  – a pattern independent of the individual jackets and shoes
- *New item problem* in recommendation systems:
  - A newly added item cannot be part of groups of previously identified co-occurring items. However, it may be semantically related to existing items, and should therefore be recommended analogously.



## Why integrate structure?

**Page type defined by hyperlink structure bears information on function, or the designer's view of how pages will be used [from Cool00]:**

| Page Type | Expected Physical Characteristics | Expected Usage Characteristics |
|---|---|---|
| Head | • In-links from most site pages<br>• Root of site file structure | • First page in user sessions |
| Media | • Large text/graphic to link ratio | • Long average reference length |
| Navigation | • Small text/graphic to link ratio | • Short average reference length<br>• Not a maximal forward reference |
| Look-up | • Large number of in-links<br>• Few or no out-links<br>• Very little content | • Short average reference length<br>• Maximal forward reference |
| Data Entry | • "FORM" tag is present | • Followed by a POST request |

- can be assigned manually by the site designer,
- or automatically by using classification algorithms
- a classification tag can be added to each page (e.g., using XML tags).

## Content and structure: Preprocessing tasks

- Processing content and structure of the site are often essential for successful page analysis
- Two primary tasks:
  - determine what constitutes a pageview
  - represent content and structure of pages in a quantifiable form



## Content and structure: Basic elements of preprocessing

- Creation of a site map:
  - captures linkage and frame structure of the site
  - also needs to identify script templates for dynamically generated pages
- Extraction of important content elements in pages: Meta-information, keywords, internal and external links, etc.
- Identification and classification of pages based on their content and structural characteristics

## Quantifying content and structure: Static pages

- All information contained in the HTML files
- Parse each file to get a list of links, frames, images, text
- Obtain files through file system, or from spiders issuing HTTP requests



## Quantifying content and structure: Dynamic pages

- Pages do not exist until created due to a specific request
- Information from various sources: templates, databases, scripts, HTML, ...
- This information may be available in various forms:
  1. A domain model exists; pages are generated from it
  2. A domain model can be compiled from internal sources (e.g., database schemas)
  3. Semantic information can be automatically extracted by analyzing URLs (from the log or from a spider) and/or page content

## Content and structure: Information from available domain models

Explicit domain models can be available in several forms, including

- A Content Management System generates Web pages from a product catalog

  ⇒ map server objects to application objects as described in the product hierarchy

  Examples of using retailing product hierarchies for mining: KDD Cup 2000: http://www.ecn.purdue.edu/KDDCUP/

- Pages are generated from an ontology and an inference engine

  ⇒ map server objects to concepts and relations as described in the product hierarchy

  Example: Knowledge Annotation Initiative of the Knowledge Acquisition Community (http://ka2portal.aifb.uni-karlsruhe.de, [Obe00])



## Content and structure: Information compiled from internal sources

In the absence of an explicit domain model, an understanding of the database schemas, query options, and transaction models can help the analyst

1. construct a classification or taxonomy of pages (manual step)
2. map URLs into the concepts of this domain model (semi-automatic step)

## Content and structure: Identifying concepts

Regardless of whether concepts are formulated and structured for an explicit model before page generation, or a model compiled during analysis, concepts can be structured according to the questions of the analysis, e.g.:

- Content-based taxonomies: based on database tables and attributes examples: product groups (cf. above); entity classes in an information site [BS00]

  This focuses on the objects of users request actions.

- Service-based / activity-based taxonomies:

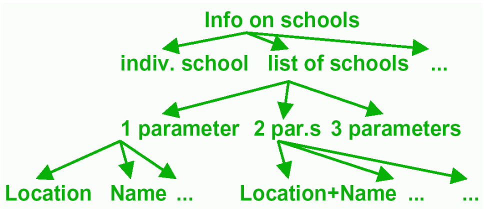
  These focus on the services users requested and/or the activities they performed, rather than on the objects of their actions.



## Content and structure: Example I of concept structure

**Search options:** [BS00] propose service-based taxonomies of the search and display options used; e.g., search by location, name, or both; short or detailed listing

In this domain model:



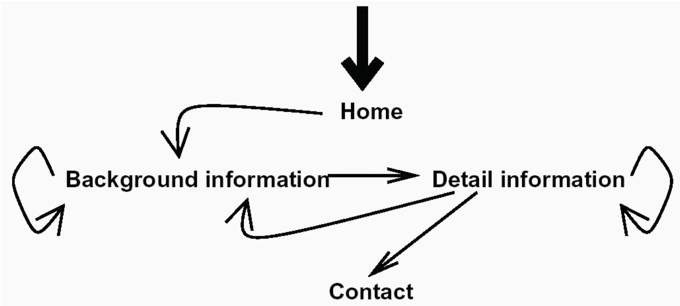

an optimal "three-click search" looks like this (from the transformed log):

[Home, List-of-Schools, Indiv-School]

## Content and structure: Example II of concept structure

**Activity models:** [SPT02] model the customer buying cycle known from marketing. In this cycle, the following activities can be distinguished. Their use gives rise to a customer typology.

For example, "knowledge builders" are users that follow a path through the pages of the site of the following form




[Ber02a, Ber02b] uses a similar activity model, enriched by a canonical event sequence for visualizing and comparing the intended and the actual usage of a site.



## Content and structure: Automatic information extraction

- Basic idea: Keywords are extracted from (visited) pages for content description.
- Based on the *vector space model* of document collections:
  - Each unique word in a corpus of Web pages = one dimension
  - Each page(view) is a vector with non-zero weight for each word in that page(view), zero weight for other words
- Words are also called "features".

## Data Preparation Tasks for Mining Content Data

Feature representation for pageviews

- each pageview p is represented as a *k*-dimensional feature vector, where *k* is the total number of extracted features from the site in a global dictionary
- feature vectors obtained are organized into an inverted file structure containing a dictionary of all extracted features and posting files for pageviews

Conceptually, the inverted file structure represents a document-feature matrix, where each row is the feature vector for a page and each column is a feature



## Basic Automatic Text Processing

Parse documents to recognize structure

- e.g. title, date, other fields

Scan for word tokens

- lexical analysis using finite state automata
- numbers, special characters, hyphenation, capitalization, etc.
- record positional information for proximity operators

Stopword removal

- based on short list of common words such as “the”, “and”, “or”

## Basic Automatic Text Processing

**Stem words**

- morphological processing to group word variants such as plurals
- better than string matching (e.g. comput*)
- can make mistakes but generally preferred

**Weight words**

- using frequency in documents and database
- frequency data is independent of retrieval model

**Optional**

- phrase indexing, concept indexing, thesaurus classes

**Store in inverted index**



## Document Representation as Vectors

**Starting point is the raw term frequency as term weights**

**Other weighting schemes can generally be obtained by applying various transformations to the document vectors**

Document Ids

Features

| | nova | galaxy | heat | actor | film | role | diet |
|---|---|---|---|---|---|---|---|
| A | 1.0 | 0.5 | 0.3 | | | | |
| B | 0.5 | 1.0 | | | | | |
| C | 0.4 | 1.0 | 0.8 | | 0.7 | | |
| D | | | | 0.9 | 1.0 | 0.5 | |
| E | 0.5 | 0.7 | | | 0.9 | | |
| F | | | 0.6 | 1.0 | 0.3 | 0.2 | 0.8 |

a document vector

## Computing Similarity Among Documents

**Advantage of representing documents as vectors is that it facilitates computation of document similarities**

**Example (Vector Space Model)**

- **the dot product of two vectors measures their similarity**
- **the normalization can be achieved by dividing the dot product by the product of the norms of the two vectors**
- **given vectors** $X = \langle x_1, x_2, \cdots, x_n \rangle$ $Y = \langle y_1, y_2, \cdots, y_n \rangle$
- **the similarity of vectors X and Y is:**

$$sim(X,Y) = \frac{\sum_i (x_i \times y_i)}{\sqrt{\sum_i x_i^2 \times \sum_i y_i^2}}$$

**Note: this measures the cosine of the angle between two vectors**



## Inverted Indexes

**An Inverted File is essentially a vector file "inverted" so that rows become columns and columns become rows**

| *docs* | *t1* | *t2* | *t3* |
|---|---|---|---|
| D1 | 1 | 0 | 1 |
| D2 | 1 | 0 | 0 |
| D3 | 0 | 1 | 1 |
| D4 | 1 | 0 | 0 |
| D5 | 1 | 1 | 1 |
| D6 | 1 | 1 | 0 |
| D7 | 0 | 1 | 0 |
| D8 | 0 | 1 | 0 |
| D9 | 0 | 0 | 1 |
| D10 | 0 | 1 | 1 |

| *Terms* | D1 | D2 | D3 | D4 | D5 | D6 | D7 | … |
|---|---|---|---|---|---|---|---|---|
| *t1* | 1 | 1 | 0 | 1 | 1 | 1 | 0 | |
| *t2* | 0 | 0 | 1 | 0 | 1 | 1 | 1 | |
| *t3* | 1 | 0 | 1 | 0 | 1 | 0 | 0 | |

**Term weights can be:**

**- Binary**

**- Raw Frequency in document (Text Freqency)**

**- Normalized Frequency**

**- TF x IDF**

## Assigning Weights

**tf x idf measure:**

**term frequency (tf) x inverse document frequency (idf)**

- **Want to weight terms highly if they are frequent in relevant documents … BUT infrequent in the collection as a whole**

**Goal: assign a tf x idf weight to each term in each document**

$$\boxed{w_{ik} = tf_{ik} * \log(N / n_k)}$$

$T_k$ = term $k$ in document $D_i$

$tf_{ik}$ = frequency of term $T_k$ in document $D_i$

$idf_k$ = inverse document frequency of term $T_k$ in $C$

$N$ = total number of documents in the collection $C$

$n_k$ = the number of documents in $C$ that contain $T_k$

$$idf_k = \log\left(\frac{N}{n_k}\right)$$

$$\log\left(\frac{10000}{10000}\right) = 0$$

$$\log\left(\frac{10000}{5000}\right) = 0.301$$

$$\log\left(\frac{10000}{20}\right) = 2.698$$

$$\log\left(\frac{10000}{1}\right) = 4$$



## How are content and structure used in subsequent mining?

**The structures shown so far are used in different ways:**

- **Mining is performed on the transformed structure, e.g.,**
  - **On the sessions transformed into points in feature space [MDL+00]**
  - **On the sessions transformed into sequences of content/activity units at a given level of description [Ber02a,BS00,SPT02].**
- **A pattern identified by mining is transformed and then processed further in an interactive way, e.g.,**

  **A frequent sequence is represented as a sequence of keyword sets; the analyst can interpret and name this as a search for a specific goal [CPP01].**
- **During mining, the most specific level of relationships is identified [SA95,DM02].**

## Content and structure: Example of keyword-based analysis (1)

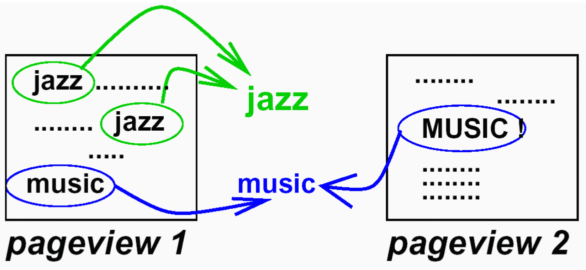
[MDL+00] present a common representation of usage, content, and structure
Goal: combine recommendations based on semantic relatedness, co-occurrence in user sessions, and structural characteristics

1. Sessions are represented as pageview-weight vectors

   $[\langle \text{pageview 1}, 0.3\rangle, \langle \text{pageview 2}, 0.2\rangle, \ldots]$,

   with the weights according to the relative importance (frequency) of that pageview in the session.
2. The content of pages is represented as a set of pageview-weight vectors:

   1. Keyword extraction → each document is a point in feature space






## Content and structure: Example of keyword-based analysis (2)

2. Feature-pageview matrix → each feature is a point in pageview space

| | music | jazz | artist | ... |
|---|---|---|---|---|
| pv1 | 1.00 | 0.80 | 0.05 | |
| pv2 | 1.00 | 0.00 | 0.70 | |
| ... | | | | |

⇒ jazz = [<pv1, 0.80>, <pv2, 0.00>, ...]

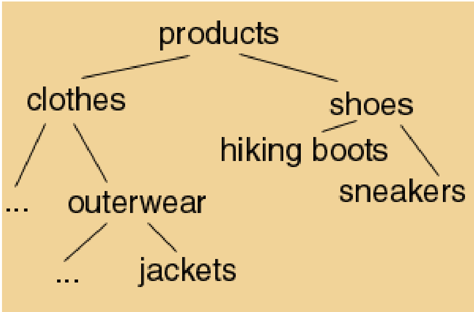
3. The structure of pages is represented as a set of pageview-weight vectors:

   Ex.: $[\langle \text{pv1}, 0\rangle, \langle \text{pv2}, 1\rangle, \langle \text{pv3}, 0\rangle]$: Only pv2 is a content page and therefore has weight 1.

⇒ Example application:
recommend content pages on jazz visited by users with similar navigation

## An example of the identification of the most specific relationship

Search for *associations* in the following taxonomy [cf. SA95]:




May obtain rules like:

"People who buy jackets tend to buy shoes."

"People who buy outerwear tend to buy hiking boots."

Here, taxonomy is given => clear how to generalize concepts.

[DM02] present a scheme for aggregating towards more general concepts when an explicit taxonomy may be missing. *Clustering* is applied to sets of sessions; it identifies related concepts at different levels of abstraction.

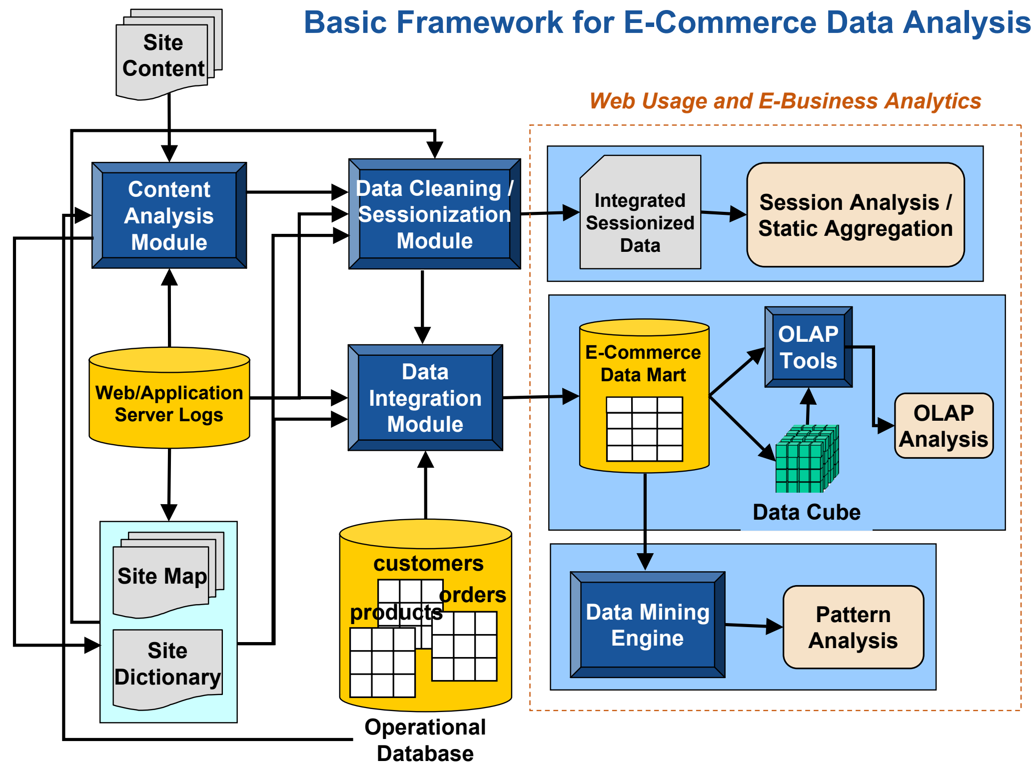


## Semantic Web Mining

[SHB01,BSH02,BSH02] have proposed *Semantic Web Mining* as a focal concept to describe all research that takes semantics into account.

Semantic Web Mining refers to this *mining of the Semantic Web* and to *mining for the Semantic Web*, in particular for building up structures that annotate Web pages semantically.

**Join us tomorrow!**




## Components of E-Commerce Data Analysis Framework

### Content Analysis Module

- extract linkage and semantic information from pages
- potentially used to construct the site map and site dictionary
- analysis of dynamic pages includes (partial) generation of pages based on templates, specified parameters, and/or databases (may be done in real time, if available as an extension of Web/Application servers)

## Components of E-Commerce Data Analysis Framework

**Site Map / Site Dictionary**

- site map is used primarily in data preparation (e.g., required for pageview identification and path completion); it may be constructed through content analysis and/or analysis of usage data (e.g., from referrer information)
- site dictionary provides a mapping between pageview identifiers / URLs and content/structural information on pages; it is used primarily for "content labeling" both in sessionized usage data as well as integrated e-commerce data



## Components of E-Commerce Data Analysis Framework

**Data Integration Module**

- used to integrate sessionized usage data, e-commerce data (from application servers), and product/user data from databases
- user data may include user profiles, demographic information, and individual purchase activity
- e-commerce data includes various product-oriented events, including shopping cart changes, purchase information, impressions, click-throughs, etc.
- primarily used for data transformation and loading mechanism for the Data Mart

## Components of E-Commerce Data Analysis Framework

**E-Commerce Data mart**

- this is a multi-dimensional database integrating data from a variety of sources, and at different levels of aggregation
- can provide pre-computed e-metrics along multiple dimensions
- is used as the primary data source in OLAP analysis, as well as in data selection for a variety of data mining tasks (performed by the data mining engine



## Web Usage and E-Business Analytics

### Different Levels of Analysis

- Session Analysis
- Static Aggregation and Statistics
- OLAP
- Data Mining

## Session Analysis

**Simplest form of analysis: examine individual or groups of server sessions and e-commerce data.**

**Advantages:**

- **Gain insight into typical customer behaviors.**
- **Trace specific problems with the site.**

**Drawbacks:**

- **LOTS of data.**
- **Difficult to generalize.**



## Static Aggregation (Reports)

**Most common form of analysis.**

**Data aggregated by predetermined units such as days or sessions.**

**Generally gives most "bang for the buck."**

**Advantages:**

- **Gives quick overview of how a site is being used.**
- **Minimal disk space or processing power required.**

**Drawbacks:**

- **No ability to "dig deeper" into the data.**

| Page View | Number of Sessions | Average View Count per Session |
|---|---|---|
| Home Page | 50,000 | 1.5 |
| Catalog Ordering | 500 | 1.1 |
| Shopping Cart | 9000 | 2.3 |

## Online Analytical Processing (OLAP)

Allows changes to aggregation level for multiple dimensions.

Generally associated with a Data Warehouse.

Advantages & Drawbacks

- Very flexible
- Requires significantly more resources than static reporting.

| Page View | Number of Sessions | Average View Count per Session |
|---|---|---|
| Kid's Stuff Products | 2,000 | 5.9 |

| Page View | Number of Sessions | Average View Count per Session |
|---|---|---|
| Kid's Stuff Products | | |
| Electronics | | |
| Educational | 63 | 2.3 |
| Radio-Controlled | 93 | 2.5 |



## Data Mining: Going Deeper (I)

### Frequent Itemsets

- The “Home Page” and “Shopping Cart Page” are accessed together in 20% of the sessions.
- The “Donkey Kong Video Game” and “Stainless Steel Flatware Set” product pages are accessed together in 1.2% of the sessions.

### Association Rules

- When the “Shopping Cart Page” is accessed in a session, “Home Page” is also accessed 90% of the time.
- When the “Stainless Steel Flatware Set” product page is accessed in a session, the “Donkey Kong Video” page is also accessed 5% of the time.

### Sequential Patterns

- add an extra dimension to frequent itemsets and association rules - time
- “x% of the time, when A appears in a transaction, B appears within z transactions.”
- Example:The “Video Game Caddy” page view is accessed after the “Donkey Kong Video Game” page view 50% of the time. This occurs in 1% of the sessions.

## Data Mining: Going Deeper (II)

**Clustering: Content-Based or Usage-Based**

- Customer/visitor segmentation
- Categorization of pages and products

**Classification**

- "Donkey Kong Video Game", "Pokemon Video Game", and "Video Game Caddy" product pages are all part of the Video Games product group.
- customers who access Video Game Product pages, have income of 50K+, and have 1 or more children, should be get a banner ad for Xbox in their next visit.



## Agenda

- Introduction
- Data Acquisition and Data Preparation
- Evaluation of Web Site Success
- Applications and KDD Techniques for them
- Privacy Concerns
- Research Issues and Future Directions

## What does Success mean?

Before talking of success:

- Why does the site exist? — Business goals
- Why should someone visit it? — Value creation
- Why should someone return to it? — Sustainable value

After answering these questions:

- Does the site satisfy its owner? — Business-centric measures
- Does the site satisfy its users? — User-centric measures
- ALL the users? — User types

Value creation: [Kuhl96]



## What does Success mean?

Before talking of success:

- Why does the site exist? — Business goals
- Why should someone visit it? — Value creation
- Why should someone return to it? — Sustainable value

After answering these questions:

- Does the site satisfy its owner? — Business-centric measures
- Does the site satisfy its users? — User-centric measures
- ALL the users? — User types

## Business Goals of a Site (I)

Personalization

1. Sale of products/services on-line

Amazon sells books (etc) online.
The site should help the users find the most suitable books for their needs, identify further related products of interest and, finally purchase them in a secure and intuitive way.

Site design

Cross/Up-Selling

2. Marketing for products/services to be acquired off-line

Insurances, banks, application service providers etc: providers of services based on a long-term relationship with the customer do not sell on-line to unknown users.
The site should persuade the users on the quality of the product/service and on the trustworthiness of its owner and initiate an off-line contact.



## Business Goals of a Site (II)

3. Reduction of internal costs

Some banks offer online banking. Some insurances support case registration online. This reduces the need for human-preprocessing and the likelihood of typing errors.
The site should help the users locate and fill the right forms and submit them in a secure and intuitive way.

4. Information dissemination

Google, AltaVista, IMDB offer information by means of a search engine over a voluminous archive of high quality data.
The site should help the users find what they search for, ensure them upon the quality (precision and completeness) of the information provided, and also motivate them to take advantage of the products/services of the sponsors.

4. Networking

5. Public relations

## What does Success mean?

Before talking of success:

- Why does the site exist? — Business goals
- Why should someone visit it? — Value creation
- Why should someone return to it? — Sustainable value

After answering these questions:

- Does the site satisfy its owner? — Business-centric measures
- Does the site satisfy its users? — User-centric measures
- ALL the users? — User types



## User-Centric Measures

- User-centric measures quantify usability.

  A product's usability is high if users

  - achieve their goals / perform their tasks in little time,
  - do so with a low error rate,
  - experience high subjective satisfaction.

  cf. ISO definition, given in [Usab99], [Niel01]

- The exact measure(s) chosen depend on the questions of the analysis, and also on the site's purpose, see also [Spen99].
- Consider information utility and entertainment value! [Eigh97].

## Usability on the Web

**Usability is a special concern on the Web because**

**"In product design and software design, customers pay first and experience usability later.**

**On the Web, users experience usability first and pay later."**

**[Niel00, pp. 10f.]**



## Design decisions that influence usability

**Design = page design + site design**

**Page design concerns issues like:**

- **Screen real estate, links, graphics+animation, cross-platform design; content design (writing for *hyper*media)**

**Site design / Information architecture concerns issues like:**

- **Hierarchical / network-like content organization, metaphors**
- **Navigation**
  - **Where am I? Where have I been? Where can I go?**
  - **Navigation is user-controlled!**

**Further issues: Users with disabilities, international audiences**

## General: Common usability mistakes

In 1996 and 1999, Jakob Nielsen investigated the "Top Ten Mistakes in Web Design." [Niel96,Niel99]

"All ten mistakes from 1996 are still mistakes in 1999."

7 out of 10 were still "severe" or "very severe" problems:

- Slow download times
- Bleeding-edge technology
- Scrolling text and looping animations
- Outdated information
- Lack of navigation support
- Non-standard link colors
- Complex URLs



## General: Principles of successful navigation

Navigation that works should [Flem98, pp. 13f.]

- Be easily learned
- Remain consistent
- Provide feedback
- Appear in context
- Offer alternatives
- Require an economy of action and time
- Provide clear visual messages
- Use clear and understandable labels
- Be appropriate to the site's purpose
- Support users' goals and behaviors

## Site-specific usability issues: Example I

**Navigation "requires an economy of action and time."**

**=> Pages that are frequently accessed together should be reachable with one or very few clicks.**

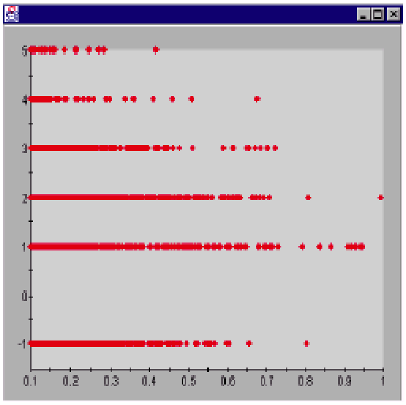
**[KNY00] compared :**

- **page co-occurrence in user paths (*x* axis), with**
- **hyperlink distance (*y* axis; -1 = distance > 5)**

**Results help to identify**

- **linkage candidates (top right)**
- **redundant links (bottom left)**

**=> Action: modify site design**






## Site-specific usability issues: Example II

**Navigation should "support users' goals and behaviors ."**

**=> Search criteria that are popular should be easy to find+use.**

**[BS00,Ber02] investigated search behavior in an online catalog:**

- **Search using selection interfaces (clickable map, drop-down menue) was most popular.**
- **Search by location was most popular.**
- **The most efficient search by location (type in city name) was not used much.**

**=> Action: modify page design.**

## General: Usability of personalized pages

In a personalized systems, page and site design may be different for each user. Studies on adaptive interfaces show pros and cons:

[Bel00]: Better performance and higher subjective satisfaction if users

- understand how the system works and generates its suggestions,
- have control over the use (or not) of suggestions,
- trust the system.

[Brus97,BE98] - Survey of research on adaptive educational software:

- Interfaces changing over time are difficult to learn.
- Adaptive information depth improves comprehension, reduces reading time
- Adaptive link annotation reduces no. of visited pages and learning time, encourages non-sequential navigation.
- Adaptive link ordering reduces search time;confusing for novices.
- The more users agree with the system's suggestions, the better their test results.



## How can usability be measured?

Usability is tested using different methods [Shne98, Jane99]:

- Reactive methods
  - Expert reviews and surveys ask for attitudes / assessments.
  - Usability testing employs experimental methods to investigate behavior and self-reports.
- Non-reactive methods
  - Based on data collection via Web log files
    - To assess user behavior
    - To simulate expected / measured user behavior [CPCP01]

Continuing assessments to parallel changes!

Issues: cost, practicality, expressiveness of results

## Mining for usability assessment: Caveats for interpretation

However, care should be taken when interpreting Web log data as indicative of users' experience with the site:

\+ Users act in a natural environment, and in a natural way.

– little or no control of variables that may influence behavior:

- User intentions and intervening factors (work environment, …)
- Context (e.g., online + offline competition, market developments)
- Often, several characteristics of the site are changed simultaneously, e.g., product offerings and page design.

=> Causality is hard to assess!

=> Use mining as an exploratory method, to be complemented by other methods that allow for more control.



## What does Success mean?

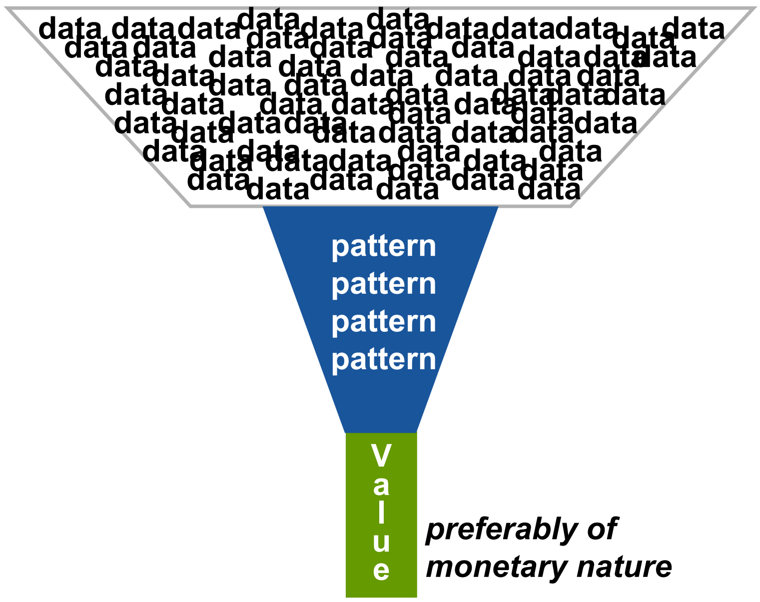
Before talking of success:

| | |
|---|---|
| ■ Why does the site exist? | Business goals |
| ■ Why should someone visit it? | Value creation |
| ■ Why should someone return to it? | Sustainable value |

After answering these questions:

| | |
|---|---|
| ■ Does the site satisfy its owner? | Business-centric measures |
| ■ Does the site satisfy its users? | User-centric measures |
| ■ ALL the users? | User types |

## The Purpose of Business-Centric Measures






## Business-Centric Measures

- User satisfaction is a pre-requisite for the success of a Web site.
- User satisfaction does not imply that the Web site is successful.

Business venues evaluate their achievements on the basis of industry- and application-specific measures.

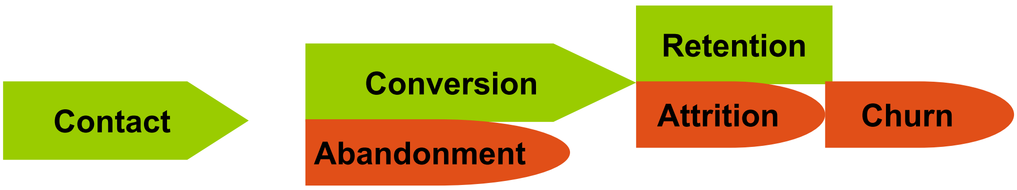
Some of these measures have been adapted for Web applications:

- e-Marketing measures for online sales of products/services
- e-Marketing measures for commodities that are sold offline
- e-measures adjusted to/alienated for other business goals

## Business-Centric Measures for Sales of Products/Services

The interaction of the potential customer with the company goes through three phases:


The ratio of persons going from one phase to the next is the basis for a set of positive and negative measures:






## Online Sales of Products/Services: Customer-Oriented Business-Centric Measures (I)

Early realization of the marketing measures for Web sites [BPW96]:


- Conversion efficiency := Customers / Active investigators
- Retention efficiency := Loyal Customers / Customers

whereby:

- Active investigators are visitors that stay long in the site.
- Customers are visitors that buy something.
- Loyal customers are customers that come to buy again.

## Online Sales of Products/Services: Customer-Oriented Business-Centric Measures (II)

**The model of [SP01] defines contact and conversion efficiency at the page(type) level:**

- **Active investigator is a user that invokes an action page**
- **Customer is an active investigator that invokes a target page**

**whereby:**

**Target page := any page corresponding to the fullfillment of the site's objectives**

- **purchase of a product**
- **registration to a service**

**Action page := any page that must be visited before invoking a target page**

- **product impression**
- **catalog search**



## Online Sales of Products/Services: Customer-Oriented Business-Centric Measures (II cntd.)

**Conversion efficiency can then be defined for**

- **any action page A : ratio of customers that invoked any target page via A to all visitors of A**
- **any pair of action and target pages: ratio of customers that invoked target page T via action page A to all visitors of A**
- **routes between A and T: ratio of customers that reach T from A via a specific route to all visitors of A**
  - **Route which is no longer than 3 pages**
  - **Route across offers-of-the-month only**

## Online Sales of Products/Services: Customer-Oriented Business-Centric Measures (II cntd.)

Conversion efficiency can then be defined for

- any action page A
- any pair of action and target pages
- routes between A and T

as the contribution of a page to the fulfilment of the site objectives.

Routes over sessions can be defined in the template-based mining language MINT of the web usage miner WUM [SF99,Spi99]



## Online Sales of Products/Services: Customer-Oriented Business-Centric Measures (III)

The model of [LPS+00] considers four steps until the purchase of a product:

- Product impression
- Click through
- Basket placement
- Product purchase

and introduces micro-conversion rates for them:

- look-to-click rate: click-throughs / product impressions
- click-to-basket rate: basket placements / click-throughs
- basket-to-buy rate: product purchases / basket placements
- look-to-buy rate: product purchases / product impressions

## Online Sales of Products/Services: Site-Oriented Business-Centric Measures (I)

In [DZ97], site efficiency is defined in terms of:

- Number of page requests
- Duration of site visits (sessions)

In [Sul97], site quality is defined in terms of:

- Response time
- Number of supported navigation modi
- Discoverability of a page
  finding out that a page on a given subject exists
- Accessibility of a page
  finding the page, after discovering that it exists
- Pages per visitor
- Visitors per page



## Online Sales of Products/Services: Site-Oriented Business-Centric Measures (II)

Site-oriented measures are

- statistics on the traffic of the Web site
- values based on the characteristics of the site from a designer's perspective

trying to capture the user perception of the site, without asking the user.

They do not consider the owner's intentions, i.e. the business goals of the site.

Combination of business-oriented and site-oriented measures

## Online Sales of Products/Services: Hybrid Business-Centric Measures

The e-metrics model of [CS00] is designed to

- compute values for customer-oriented measures

by

- allowing for an application-dependent definition of concepts
  - customer
  - conversion
  - loyalty
  - customer lifetime value

and by

- associating these concepts with site-oriented measures
- upon regions of the site

with some emphasis on online merchandizing.



## Online Sales of Products/Services: Hybrid Business-Centric Measures (cntd)

The e-metrics model of [CS00] encompasses:

- Site-centric measures for regions of a site, including:
  - $\text{Stickiness:=} \dfrac{\text{Total time spent in the region}}{\text{Number of visitors in the region}}$
  - Slipperiness := Stickiness
  - $\text{Focus:=} \dfrac{\text{Avg num of visited pages in the region}}{\text{Number of pages in the region}}$
- "Desirable value ranges" for each measure, depending on the purpose/objective of the region:
  - A region used during information acquisition should be sticky.
  - The pages accessed during the negotiation and transaction phase should be slippery.

## Business-Centric Measures Revisited

**Traditional business-centric measures:**

- are defined in terms of the application domain.
- are not directly reflected in the site usage.

**Site-oriented measures:**

- are based on site usage

**Hybrid business-oriented measures:**

- map site entities into business entities (customer, purchase)
- associate site usage with traditional measures

mostly in the domain of online sales of products/services.



## Business-Centric Measures and Business Goals of a Site

**Recall some potential business goals of a site:**

1. Sale of products/services on-line
2. Marketing for products/services to be acquired off-line
3. Reduction of internal costs
4. Information dissemination
5. Networking
6. Public relations

**Success evaluation of these goals demands**

- mapping of traditional business measures upon usage data
- integration of site usage data with data from other channels
- unambiguous association of site usage with success events

**Application Case:**
**Success evaluation for a site of type 2.**

## What does Success mean?

Before talking of success:

- Why does the site exist? **Business goals**
- Why should someone visit it? **Value creation**
- Why should someone return to it? **Sustainable value**

After answering these questions:

- Does the site satisfy its owner? **Business-centric measures**
- Does the site satisfy its users? **User-centric measures**
- ALL the users? **User types**



## User Segmentation

Truisms:

- A site owner does not welcome all users equally.
- A site cannot satisfy all users accessing it.

Hence, sites

- are designed for some types of users
- serve different user types to different degrees

User types are the result of:

- User segmentation according to criteria of the site owner
- User segmentation on the basis of personal characteristics
- User segmentation with respect to recorded behaviour

## User Segmentation
## In Predefined Segments (I)

A company may partition its customers on the basis of

- the revenue it obtains or expects from them
- the (cost of) services it must offer them to obtain the revenue

There are different segmentation schemes,
based on

- the characteristics of the customers
- the company portfolio

and producing a set of predefined classes.

For a Web site application this means:

**Classification**

For each visitor:
(1) assign her to the right segment ASAP
(2) motivate her to move to a segment of higher revenue

**Association rules for cross selling & up selling**

**Recommendations Personalization**



## User Segmentation
## In Predefined Segments (II)

Web site visitors exhibit different types of navigational behaviour.

There are different visitor segmentation schemes,
based on

- the navigation facilities preferred
- the contents being visited
- the anticipated purpose of the visit

and producing a set of predefined classes.

For a Web site application this means:

**Classification**

For each visitor:
(1) assign her to the right segment ASAP
(2) motivate her to generate more revenue

**Association rules for cross selling & up selling**

**Recommendations Personalization**

## User Segmentation
## In Predefined Segments (II cntd.)

**Web site visitors exhibit different types of navigational behaviour.**

- **Model I (simplistic):**
  Some users navigate across links. Others prefer a search engine.
- **Model II [FGL+00]:**

| Simplifiers | Surfers | Connectors | Bargainers | Routiners | Sportsters |
|---|---|---|---|---|---|

  based on criteria like active time spent on-line and per page, pages and domains accessed etc.

- **Model III [Moe] for merchandizing sites:**

| Direct buying | Hedonic browsing | Search/ Deliberation | Knowledge building |
|---|---|---|---|

  based on criteria like purchase intention, time spent on the site, number of searches initiated, types of pages visited etc.



## User Segmentation
## In Unknown Segments

**Web site visitors can be grouped on the basis of their interests, characteristics and navigational behaviour without assuming predefined groups.**

There is much research on user **clustering** based on
- the properties and contents of the objects being visited
- the declared or otherwise known characteristics of the visitor
- (the order of the requests)

For a Web site application this means:

For each visitor:
(1) assign her to the right segment ASAP
(2) make suggestions based on the contents of the segment

**Recommendations**
**Personalization**

## A Summary on Success Evaluation

- The success of a site is defined with respect to its objectives.
- To be successful, a site should be satisfactory to its users. This is a necessary but not adequate condition for success.
- Success evaluation is performed by the site owner according to business-oriented measures.
- Traditional business-oriented measures have not been designed for site visitors. Site-oriented measures do not reflect the business goals of the site. Their combination is difficult but promising.
- A site cannot and should not treat all users equally. Users can be segmented with respect to their value for the site owner, with respect to their properties or behaviour.

The role of data mining:



## Data Mining and Success Evaluation

Data Mining methods are used to:

- Identify the user segments, upon which one evaluates the success of the site.
- Extract the patterns that contribute to the success of the site
  - association rules
  - groups of similar interests
  - prediction and recommendation of the next object
- Extract the patterns that describe the contribution of each site component on the overall success

Success evaluation methods should exploit data mining to:

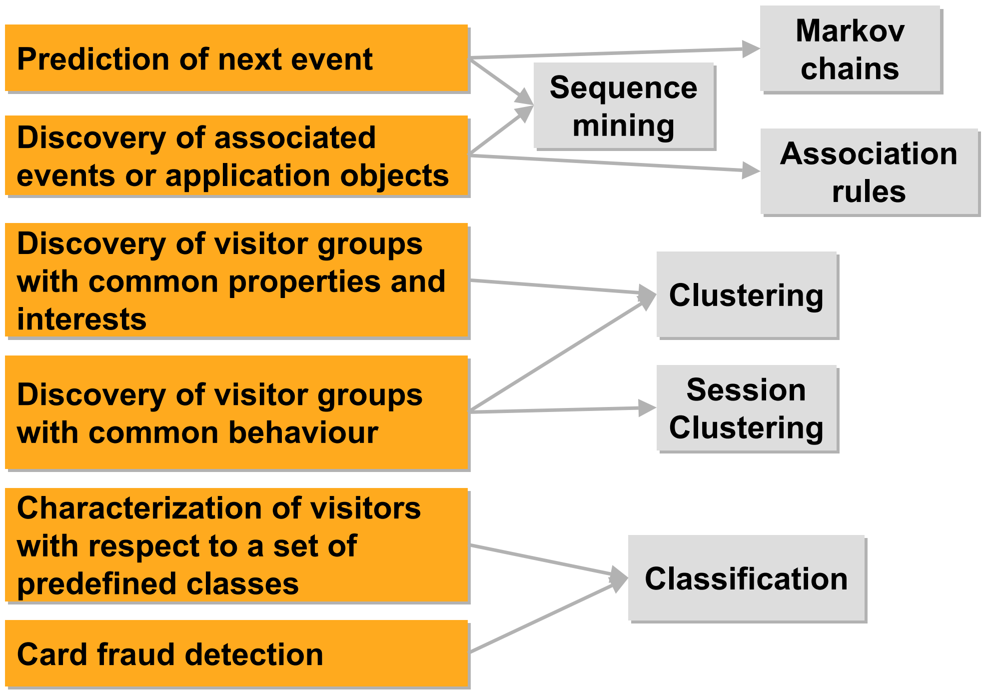
- Compute the success at a finer level than the whole population associated to a site

## Agenda

- Introduction
- Data Acquisition and Data Preparation
- Evaluation of Web Site Success
- Applications and KDD Techniques for them
- Privacy Concerns
- Research Issues and Future Directions



## Does the Analysis of Web Applications Require Special Mining Techniques?



## KDD Techniques for Web Applications

**Recall:**

- Problem specification
- Data collection
- Data preparation
- Data mining
- Presentation of the results
- Evaluation and Interpretation of the results
- Action upon the results

Web applications

**Whether a new technique is needed depends on the problem specification, which determines the kind of analysis to be done.**



## KDD Techniques for Web Applications: Examples (I)

**Calibration of a Web server:**

- **Prediction of the next page invocation over a group of concurrent Web users under certain constraints**
  - **Sequence mining, Markov chains**

**Cross-selling of products:**

- **Mapping of Web pages/objects to products**
- **Discovery of associated products**
  - **Association rules, Sequence Mining**
- **Placement of associated products on the same page**

## KDD Techniques for Web Applications: Examples (II)

Sophisticated cross-selling and up-selling of products:

- Mapping of pages/objects to products of different price groups
- Identification of Customer Groups
  - Clustering, Classification
- Discovery of associated products of the same/different price categories
  - Association rules, Sequence Mining
- Formulation of recommendations to the end-user
  - Suggestions on associated products
  - Suggestions based on the preferences of similar users



## Applications and KDD Techniques for them

Success Analysis for a Non-Merchandizing Site:
A dedicated technique for a sophisticated problem specification

Personalization:

Techniques for finding similar users and making suggestions to them

## Applications and KDD Techniques for them

Success Analysis for a Non-Merchandizing Site



## Success Analysis for a Non-Merchandizing Site [SPT02]

Web-server configuration:

- No cookies
- No proactive sessionization
- Agents are recorded

Sessionization with reactive heuristics:

- IP+agent for the assignment of sessions to individuals
- 30 minute session duration threshold

Resulting in 27,647 sessions, after the removal of

- robot entries (identified by IP+agent)
- sessions of personnel
- sessions of customers

## Success Analysis for a Non-Merchandizing Site: Description of the Site

The owner of the donor site D

- offers products/services in a long-term contractual basis
- for which personal contact and establishment of trust are prerequisites

The site serves as

- provider of information and services to customers
- marketing instrument for customer acquisition

The analysis should answer the following questions:

- What do the visitors ask for when accessing the site?
- What is the conversion rate for each type of visitors?

where conversion := establishment of contact



## Success Analysis for a Non-Merchandizing Site: Visitor types

In a merchandizing site:

- Some visitors enter with a list of products they want to buy, purchase them and leave. **Direct buying**
- Other visitors walk in to have a look; they make impulsive purchases. **Hedonic browsing**
- Some other visitors do not have specific products in mind; they want to be informed about what products satisfy a particular need. **Search/Deliberation**
- Still other visitors are interested in the shop and in the way it runs the business; their overall impression may motivate them to buy. **Knowledge-building**

Some of these visitor types are of interest for non-merchandizing sites, too.

Source: [Moe]

## Success Analysis for a Non-Merchandizing Site: Data preparation

**Observation:**

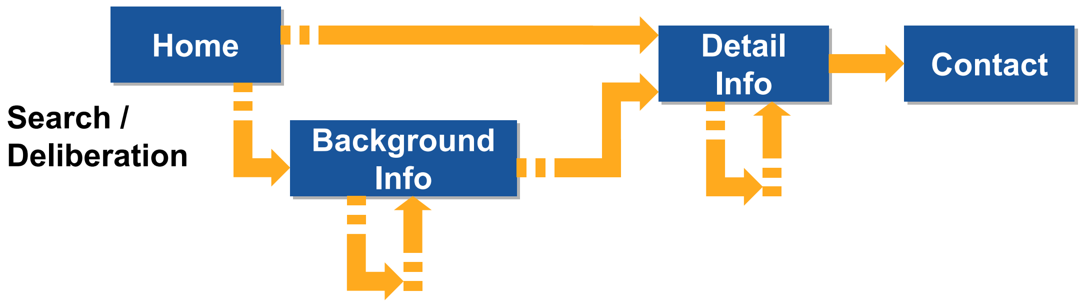
- Visitor types differ in the contents they acquire AND
- in the way they navigate.

**Actions:**

- The site's objects are mapped into concepts [PS02] associated with
  - the products/services **DetailInfo**
  - background info about the site owner **BackgroundInfo**
  - contact establishment **Contact**
- The expected navigational behaviour for each visitor type is mapped into a template that serves as input to the miner.



## Success Analysis for a Non-Merchandizing Site: Templates and Mining Queries




- Mapping the interaction strategy into a MINT query [SF98]
- Navigation pattern discovery with the miner WUM [Spi99]

```
SELECT t
FROM NODES x, y, z
TEMPLATE #x * y * z AS t
WHERE x.url = 'Home'
AND y.url = 'DetailInfo'
and z.url = 'Contact'
and wildcard.y.url = 'BackgroundInfo'
and wildcard.z.url = 'DetailInfo'
```

## Success Analysis for a Non-Merchandizing Site: Statistics of the S/D strategy (I)

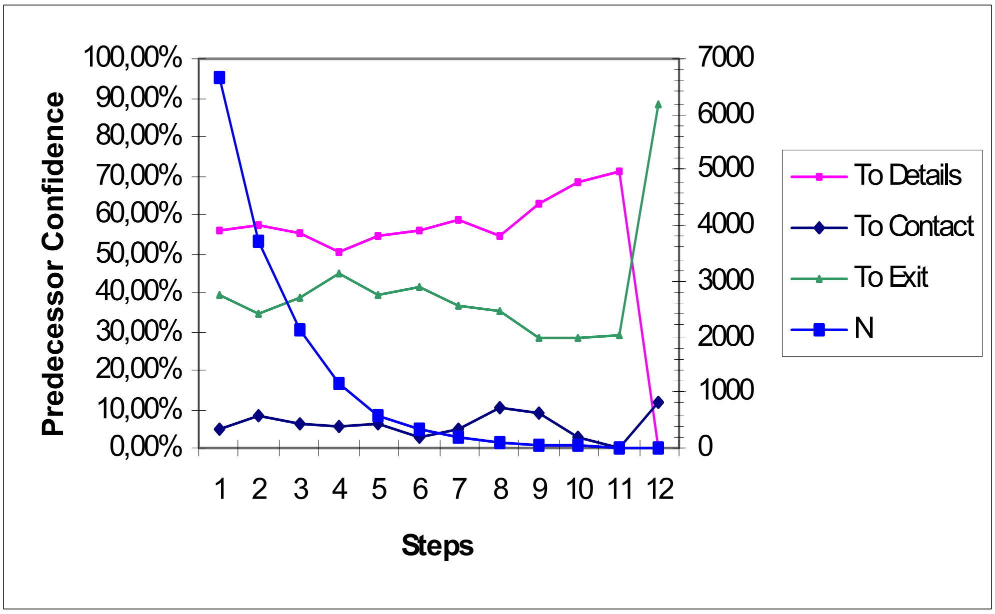
| Description | Num. of sessions | Confidence (%) | |
|---|---|---|---|
| Sessions starting at the Home page | 20815 | | 100.00 |
| Sessions invoking Detail Info at a later step | 6640 | 100.00 | 31.90 |
| Detail Info at Step 2 | 5839 | 87.93 | 87.93 |
| Sessions invoking Back-ground Info at a later step | 8929 | | 42.89 |
| Background Info at Step 2 | 3726 | | 41.72 |
| Sessions invoking Detail Info after Background Info | 801 | 12.06 | 8.97 |



## Success Analysis for a Non-Merchandizing Site: Statistics of the S/D strategy (II)

| Description | Num. of sessions | Confidence (%) | | |
|---|---|---|---|---|
| Sessions invoking Detail Info after the Home page | 6640 | 100.00 | | |
| Contact establishment after Detail Info | 896 | | 100.00 | |
| Contact at Step 1 after Detail Info | 324 | 4.88 | 36.16 | |
| Contact at Step 2 after Detail Info | 301 | | 33.59 | 8.11 |
| Detail Info at Step 1 after Detail Info | 3707 | 55.82 | 55.82 | 100.00 |

## Success Analysis for a Non-Merchandizing Site: Statistics of the S/D strategy (III)






## Success Analysis for a Non-Merchandizing Site: Lessons Learned

- The notion of success for a site
- The data to be analysed
- The metrics to be used
- The application-domain specific theories to be investigated

depend on the goals of the site and the associated objectives of the analysis.

For the specific site:

- The conversion rate is low.
- The retrieval of information assets leads to the acquisition of still further assets but its impact on conversion is limited.

## Applications and KDD Techniques for them

Personalization



## What is Web Personalization

**Web Personalization: "personalizing the browsing experience of a user by dynamically tailoring the look, feel, and content of a Web site to the user's needs and interests."**

**Why Personalize?**

- **broaden and deepen customer relationships**
- **provide continuous relationship marketing to build customer loyalty**
- **help automate the process of proactively market products to customers**
  - **lights-out marketing**
  - **cross-sell/up-sell products**
- **provide the ability to measure customer behavior and track how well customers are responding to marketing efforts**

## Standard Approaches

**Rule-based filtering**

- **provide content to users based on predefined rules (e.g., "if user has clicked on A and the user's zip code is 90210, then add a link to C")**

**Collaborative filtering**

- **give recommendations to a user based on responses/ratings of other "similar" users**

**Content-based filtering**

- **track which pages the user visits and recommend other pages with similar content**

**Hybrid Methods**

- **usually a combination of content-based and collaborative**



## Collaborative Filtering

**Example: users rate musical artists from like to dislike**

- **1 = detest; 7 = can't live without; 4 = ambivalent**

**Nearest Neighbors Strategy: Find similar users and predicted (weighted) average of user ratings**

- **Pearson *r* algorithm: weight by degree of correlation between user *U* and user *J***
- **1 means very similar, 0 means no correlation, -1 means dissimilar**

$$r_{UJ} = \frac{\sum (U-\bar{U})(J-\bar{J})}{\sqrt{\sum (U-\bar{U})^2 \cdot \sum (J-\bar{J})^2}}$$

**Average rating of user J on all items.**

- **Other similarity measures can be used (e.g., cosine similarity)**

## Collaborative Filtering
### (k Nearest Neighbor Example)

| | Star Wars | Jurassic Park | Terminator 2 |
|---|---|---|---|
| **Sally** | 7 | 6 | 3 |
| **Bob** | 7 | 4 | 4 |
| **Chris** | 3 | 7 | 7 |
| **Lynn** | 4 | 4 | 6 |
| **Karen** | 7 | 4 | 3 |

| K | Pearson |
|---|---|
| 1 | 6 |
| 2 | 6.5 |
| 3 | 5 |

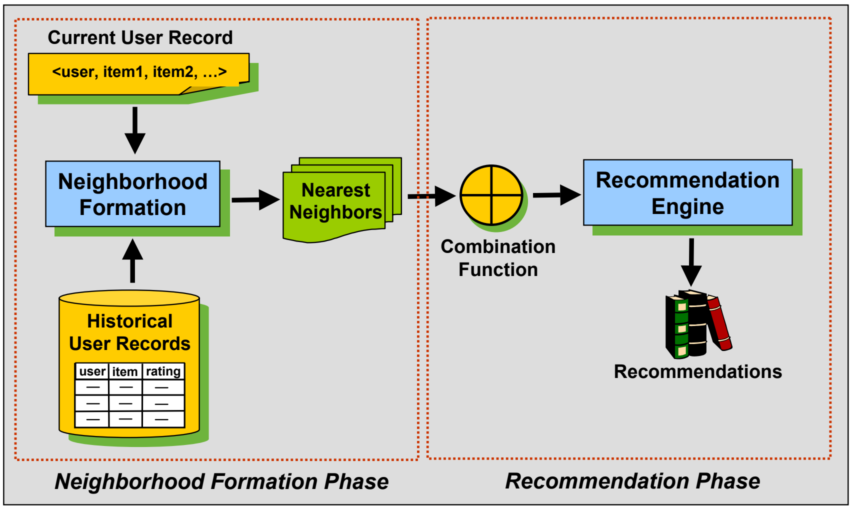
**K is the number of nearest neighbors used in to find the average predicted ratings of Karen on Indep. Day.**

Pearson(Sally, Karen) =

$$( (7-5.75)*(7-4.67) + (6-5.75)*(4-4.67) + (3-5.75)*(3-4.67) )$$
$$/ \ SQRT( ((7-5.75)^2 + (6-5.75)^2 + (3-5.75)^2) * ((7-4.67)^2 + (4-4.67)^2 + (3-4.67)^2))$$
$$= 0.82$$



## Basic Collaborative Filtering Process




**Both of the Neighborhood formation and the recommendation phases are real-time components**

## Collaborative Filtering: Pros & Cons

### Advantages

- Ignores the content, only looks at who judges things similarly
  - If Pam liked the paper, I'll like the paper
  - If you liked Star Wars, you'll like Independence Day
  - Rating based on ratings of similar people
- Works well on data relating to "taste"
  - Something that people are good at predicting about each other too
  - can be combined with meta information about objects to increase accuracy



## Collaborative Filtering: Pros & Cons

### Disadvantages

- major problem with CF is scalability: neighborhood formation is done in real-time; as number of users increase, nearest neighbor calculations become computationally intensive
- small number of users relative to number of items may result in poor performance
- because of the (dynamic) nature of the application, it is difficult to select only a portion instances as the training set
- In case of personalization based on clickstream data, explicit user ratings are not available
- early ratings by users can bias ratings of future users

## Content-Based Filtering Systems

**Track which pages the user visits and give as recommendations other pages with similar content**

- **Often involves the use of client-side learning interface agents**
  - **WebWatcher (Joachims, Freitag, Mitchell, 1997 - CMU) [JFM97]**
  - **Letizia (Lieberman, 1995 - MIT Media Labs) [Lieb95]**
- **May require the user to enter a profile or to rate pages/objects as "interesting" or "uninteresting"**
- **Profiles can be obtained implicitly by extracting content attributes from pages visited by the user**



## Content-Based Filtering Systems

Advantages:

- useful for large information-based sites (e.g., portals)
- can be easily integrated with "content servers"

Disadvantages

- may miss important semantic relationships among items (based on usage)
- not effective in small-specific sites or sites which are not content-oriented

## Personalization Based on Web Mining

### Basic Idea

- discover aggregate user profiles by automatically discovering user access patterns through Web usage mining (offline process)
  - aggregate profiles can be obtained via clustering of transactions, clustering of pageviews, association rule mining, or discovery of navigational or sequential patterns
- data sources for mining include server logs, other click-stream data (e.g., product-oriented user events), site content, and site structure
- match a user's active session against the discovered profiles to provide dynamic content (online process)



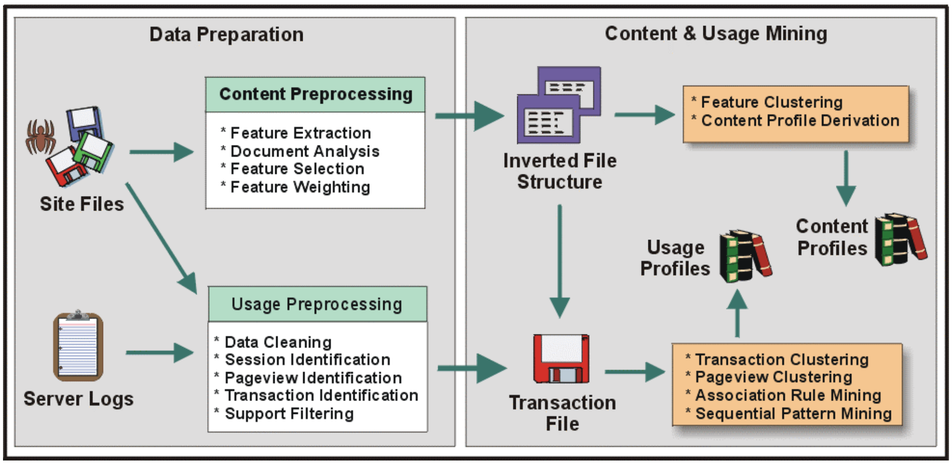
## Personalization Based on Web Mining

### Advantages / Goals

- profiles are based on objective information (how users actually use the site)
- no explicit user ratings or interaction with users is necessary
- helps preserve user privacy by making effective use of anonymous data
- captures relationships missed by content-based approaches
- can help enhance the effectiveness of collaborative or content-based filtering techniques (sometimes at the cost of reduced recommendation accuracy)

**An important goal of usage-based recommender systems: improve the scalability (through offline pattern discovery) while maintaining recommendation effectiveness)**

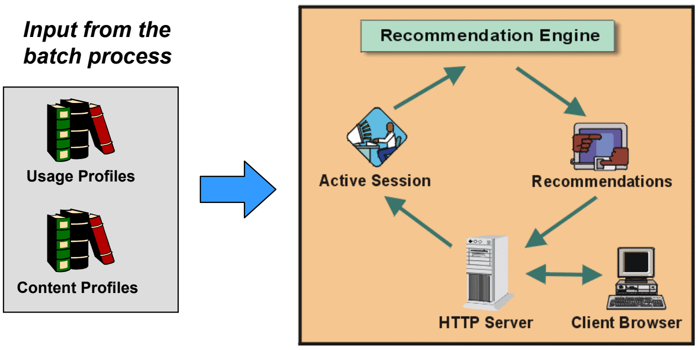
## Framework for Personalization Based on Web Mining




Offline Phase



## Framework for Personalization Based on Web Mining




Online Phase

## Discovery of Aggregate Profiles

Discovery of Profiles Based on Transaction Clusters

- cluster user transactions - features are significant items/pageviews identified in the preprocessing stage
- derive usage profiles (set of item-weight pairs) based on characteristics of each transaction cluster

Aggregate Profiles as Clusters of Items

- directly compute overlapping clusters of pageviews/items based on co-occurrence patterns across transactions
- features are user transactions, so dimensionality poses a problem for traditional clustering algorithms
- need techniques that can handle high-dimensional data, e.g., Association-Rule Hypergraph Partitioning



## Discovery of Aggregate Profiles

Association Rules as Aggregate Profiles

- match left-hand side of rules with the active user session and recommend items in the rule's consequent
- essential to store patterns in efficient data structures (the search of all rules in real-time is computationally ineffective)
- as in case of clustering, the ordering of accessed pages is not taken into account
- good recommendation accuracy, but the main problem is "coverage"
  - high support thresholds lead to low coverage and may eliminate important, but infrequent items from consideration
  - low support thresholds result in very large model sizes and computationally expensive pattern discovery phase

## Discovery of Aggregate Profiles

**Sequential / Navigational Patterns as Aggregate Profiles**

- similar to association rules, but the ordering of accessed items is taken into account
- Two basic approaches
  - use contiguous sequences (e.g., Web navigational patterns)
  - use general sequential patterns
- Contiguous sequential patterns are often modeled as Markov chains and used for prefetching (i.e., predicting the next user access based on previously accessed pages
- In context of recommendations, they can achieve higher accuracy than other methods, but may be difficult to obtain reasonable coverage



## Aggregate Profiles - The Clustering Approach

- the goal is to effectively capture common usage patterns from potentially anonymous click-stream data
- profiles are represented as weighted collections of pageviews
- weights represent the significance of pageviews within each profile
- profiles are overlapping in order to capture common interests among different groups/types of users

## Aggregate Profiles Based on Clustering Transactions (PACT) [MDL+00, MDLN02]

**Input**

- set of relevant pageviews in preprocessed log

$$P = \{p_1, p_2, \ldots, p_n\}$$

- set of user transactions

$$T = \{t_1, t_2, \ldots, t_m\}$$

- each transaction is a pageview vector

$$t = \langle w(p_1,t), w(p_2,t), \ldots, w(p_n,t) \rangle$$



## Aggregate Profiles Based on Clustering Transactions (PACT)

**Transaction Clusters**

- each cluster contains a set of transaction vectors
- for each cluster compute centroid as cluster representative

$$\vec{c} = \langle u_1^c, u_2^c, \ldots, u_n^c \rangle$$

**Aggregate Usage Profiles**

- a set of pageview-weight pairs: for transaction cluster C, select each pageview pi such that $u_i^c$ (in the cluster centroid) is greater than a pre-specified threshold

## Recommendation Engine for Clustering Approach

Match user’s activity against the discovered profiles

- a sliding window over the active session to capture the current user’s “short-term” history depth
- profiles and the active session are treated as vectors
- matching score is computed based on the similarity between vectors (e.g., normalized cosine similarity)

Recommendation scores are based on

- matching score to aggregate profiles
- “information value” of the recommended item (e.g., link distance of the recommendation to the active session)

- recommendations can be contributed by multiple matching aggregate profiles



## Association Rules & Personalization

An Approach Based on Association Rules [MDLN01]

- discovered frequent itemsets of are stored into an “itemset graph” (an extension of lexicographic tree structure of [AAP99])
  - each node at depth d in the graph corresponds to an itemset, I, of size d and is linked to itemsets of size d+1 that contain I at level d+1. The single root node at level 0 corresponds to the empty itemset.
- frequent itemsets are matched against a user's active session S by performing a search of the graph to depth |S|
  - recommendation generation can be done in constant time
  - does not require apriori generate association rules from frequent itemsets
- a recommendation r is an item at level |S+1| whose recommendation score is the confidence of rule S ==> r

## Example: Frequent Itemsets

**Sample Transactions**

T1: $\{ABDE\}$
T2: $\{ABECD\}$
T3: $\{ABEC\}$
T4: $\{BEBAC\}$
T5: $\{DABEC\}$

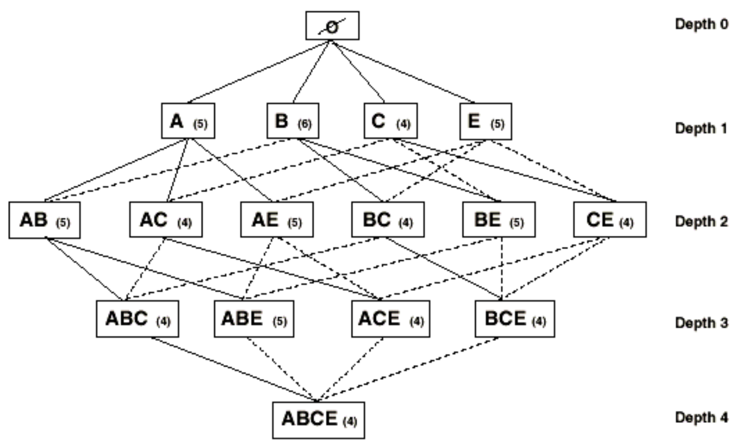
**Frequent itemsets (using min. support frequency = 4)**

| Size 1 | Size 2 | Size 3 | Size 4 |
|---|---|---|---|
| $\{A\}(5)$ | $\{A,B\}(5)$ | $\{A,B,C\}(4)$ | $\{A,B,C,E\}(4)$ |
| $\{B\}(6)$ | $\{A,C\}(4)$ | $\{A,B,E\}(5)$ | |
| $\{C\}(4)$ | $\{A,E\}(5)$ | $\{A,C,E\}(4)$ | |
| $\{E\}(5)$ | $\{B,C\}(4)$ | $\{B,C,E\}(4)$ | |
| | $\{B,E\}(5)$ | | |
| | $\{C,E\}(4)$ | | |



## Example: An Itemset Graph

**Frequent Itemset Graph for the Example**




**Given an active session window <B,E>, the algorithm finds items A and C with recommendation scores of 1 and 4/5 (corresponding to confidences of the rules {B,E }==>{A } and {B,E }==>{C} ).**

## Associations With Multiple Minimum Support

Multiple minimum supports can be used to capture associations involving “rare” but important items

Based on the work of Liu, Hsu, and Ma, 1999 [LHM99]

Particularly important in usage-based personalization:

- often references to deeper content or product-oriented pages occur far less frequently that those of top level navigation-oriented pages
- Yet, it is important to capture patterns and generate recommendations that contain these items
- Approach of Liu et al.:
  - user can specify different support values for each item
  - the support of an itemset is defined as the minimum support of all items contained in the itemset



## Quantitative Evaluation of Recommendation Effectiveness

Two important factors in evaluating recommendations

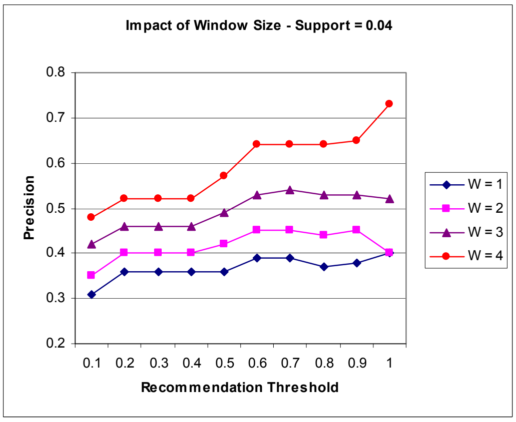
- Precision: measures the ratio of “correct” recommendations to all recommendations produced by the system
  - low precision would result in angry or frustrated users
- Coverage: measures the ratio of “correct” recommendations to all pages/items that will be accessed by user
  - low coverage would inhibit the ability of the system to give relevant recommendations at critical points in user navigation

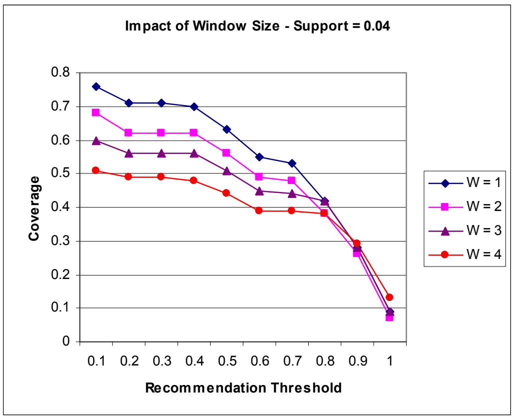
Transactions/sessions divided into Training & Evaluation Sets

- training set is used to build models (generation of aggregate profiles, neighborhood formation)
- evaluation set is used to measure precision & coverage

## Impact of Window Size

**Increasing window sizes (using larger portion of user's history) generally leads to improvement in precision**



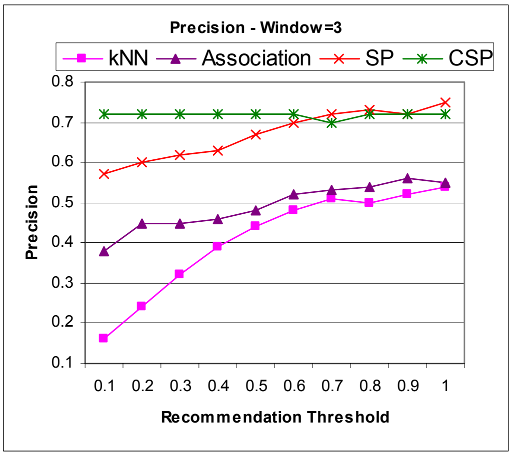



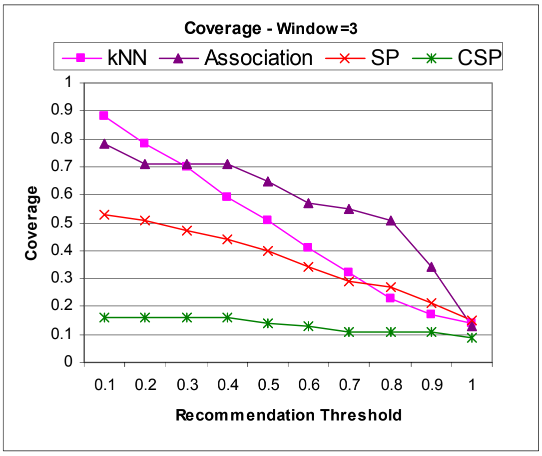

This example is based on the association rule approach



## Associations vs. Sequences

**Comparison of recommendations based on association rules, sequential patterns, contiguous sequential patterns, and standard *k*-nearest neighbor**



## Agenda

- Introduction
- Data Acquisition and Data Preparation
- Evaluation of Web Site Success
- Applications and KDD Techniques for them
- Privacy Concerns
- Research Issues and Future Directions



## Privacy issues

**Privacy is "the right to be let alone" [WB90].**

**This includes**

- **limits on the government's power to interfere with personal decisions**
- **physical privacy: limits on others' ability to learn things about a person by accessing their property**
- **information privacy: "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information aobut them is communicated to others" [West67]**

## Privacy and Web usage mining

Privacy is a special concern for Web usage mining for several reasons:

- "logical": Mining needs data
- legal: Not all data may be collected / used
- commercial:

  "The Internet industry is built on trust between businesses and their customers - and privacy is the number one ingredient in trust."

  [Trus00]



## Effects of privacy violations on Web interaction

Careless dealing with privacy issues may inflict harm on a site in various ways:

- People report that their willingness to disclose information depends on how a site deals with privacy issues [ACR99].
- Contrary to their self-reports, even privacy-conscious users disclose highly personal information during interaction [SGB01]; discovery may lead to resentment [cf. Adam01].
- Abuse of user trust may lead to
  - Abandonment of the individual site
  - Loss of faith in the industry as a whole, lying that creates worthless data

## What data are privacy-sensitive?

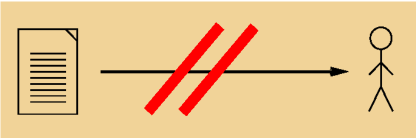
- Personal information
  - Information about a person: name, birth date, school, …
- Private information
  - Personal information that is not generally known, only sometimes protected by law (e.g., bank records),
  - Whether or not a particular piece of information is private frequently depends on the context.
- Personally identifiable information
  - Information that can be linked to a person's name or identity.



## Privacy-protecting data transformations

Basic idea: sever the link

data – physical person

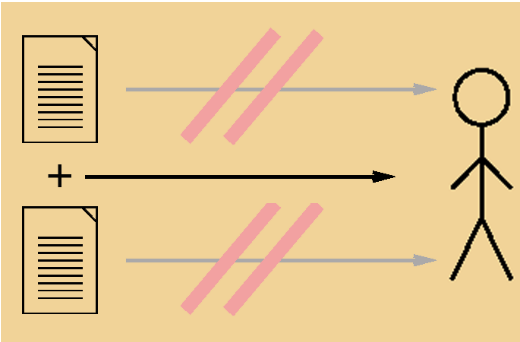


- Anonymized information
  - Information that is *not* personally identifiable
- Aggregate information
  - Statistical information combined from many individuals to form a single record.
  - Analysis / research in compliance with EU law: aggregate s.t. personal identification is impossible!

## Privacy-harming data re-transformations

**Problem: Often, triangulation is possible.**

- **= Combination of aggregate information and anonymized information to identify and reveal particular characteristics of an individual.**
- **Example: specification of one's US ZIP code + birthday – one combination applies to, on average, 8 people [GS02]**






## What data are transmitted during Web usage?

**The following data regularly are, or can be, transmitted by the browser and other technologies:**

- **IP address, domain name (may include the organization)**
- **referrer address**
- **platform: browser type and version**
- **Cookies, clear GIFs ("web bugs")**
- **query strings, form fill-ins**
  - **= any user-supplied data**
  - **"By far, the greatest kind of personal information on the Web today is the information provided by customers when they register at web sites." [GS02, p. 208]**

**Many of these data are, or can be, personally identifiable!**

## Basic approaches to privacy protection

**Who can protect privacy, and how?**

- **The state, using laws**
  - **Main European approach**
  - ***Data parsimony* is a basic EU principle: collect only what is needed**
- **The transaction parties, using market mechanisms**
  - **Main US approach**
  - ***Law of contract* is a basic US principle: transaction parties decide**
  - **Self-governance as a voluntary form of market self-regulation**
- **The users themselves, using technology**
  - **Increasing importance, development of a new market**



## Roots: The Code of Fair Information Practices

**Five principles, first formulated in a 1973 report from the Department of Health, Education, and Welfare 1973 [from GS02]**

- **There must be no personal data record-keeping systems whose very existence is secret.**
- **There must be a way for a person to find out what infor-mation about the person is in a record and how it is used.**
- **There must be a way for a person to prevent information about the person that was obtained for one purpose from being used or made available for other purposes without the person's consent.**
- **There must be a way for a person to correct or amend a record of identifiable information about the person.**
- **Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for its intended use and must take precautions to prevent misuses of the data.**

## Basic principles of European legislation

According to the EU directive 95/46/EC …

- Personal and personally identifiable information may only be collected with the informed consent (opt-in!) about
  - who : who collects the data
  - what for : for what purpose
  - how much : quality and amount necessary for purpose
- Data may then only be used as specified along these dimensions.
- Individuals can inspect and correct their data, and disallow usage.
- No data transfer to countries with inadequate data protection.
- Independent institutions overlook data protection in member countries.



## Basic principles of US legislation

Principles of the 1999 FTC discussion document "Elements of effective Self Regulation for the Protection of Privacy and Questions Related to Online Privacy" [cited from GS02]

- *Notice*: Consumers should have a right to know how an organization treats and collects personal information.
- *Choice*: A consumer should have an option to withhold personal information.
- *Access*: A consumer should have a right to view personal information that has been collected.
- *Security*: Online services should employ security measures to prevent the unauthorized release of or access to personal information.

What is missing (relative to the Fair Information Practices): the principle that people be allowed to challenge incorrect data about themselves.

General observations on current US privacy legislation

- "a piecemeal approach" [GS02]: separate legislation for financial, medical, educational, … data
- Information privacy gets protection from *law of contract* (which applies only to the parties to a contract [Volo00]).

## Legislation: Implications for Web usage mining

**General: Limitation of the data and combinations for analysis**

**Implications of EU law:**

- Opt-in is the basic principle!
- Legitimate to analyze non-personally-identifiable usage data
- Cookies are legally controversial [Maye97], but are legitimate as long as users are made aware of their presence [EU02]
- Safe Harbor principles bind non-European companies
  - US enterprises that collect + process data from EU voluntarily subject themselves to principles that correspond to EU standard; FTC-control [EU00]

**Implications of US law:**

- Opt-out is the basic principle!
- Generally, fewer restrictions



## Privacy policies

A privacy policy is a text (usu. publicized on a Web page) that

> "explains the responsibilities of the organization that is collecting personal information and the rights of the individual who provided the personal information." [Epic97]

However, …

- Of the 100 most popular shopping Web sites in 1999,
  - 18 did not display a privacy policy,
  - 35 of the sites have profile-based advertisers, and 86 use cookies,
  - not one of the companies adequately addressed all the elements of Fair Information Practices,
  - privacy policies are often confusing, incomplete, and inconsistent.

  "We concluded that the current practices of the online industry provide little meaningful privacy protection for consumers." [Epic99]
- Some sites clearly violate their stated policies [AE01].

**Implications for Web usage mining:** more planning *before* data collection?!

## Self-governance: privacy seals

A privacy seal is a measure to enforce voluntary privacy policies:

- A voluntary membership organization that polices its member companies
- Members can display a small logo, or *seal*, on compliant Web sites
- Popular examples: http://www.truste.org, http://www.bbbonline.org
- users can bring their complaints to the seal program
- The Web site must respond; the seal program aims at mediating a resolution:
  - change in company practice, or in posted policy; third party audit; refer case to government authorities, usually FTC
- Main problems:
  - The existence of a statement about privacy practices does not imply that these actually protect privacy [GS02]
  - Voluntary nature; main pressure arises from the adverse public relations consequences of privacy violations

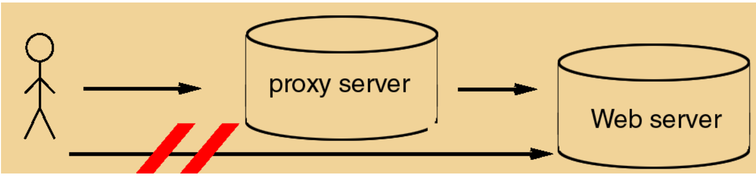
Implications for Web usage mining: same as privacy policies



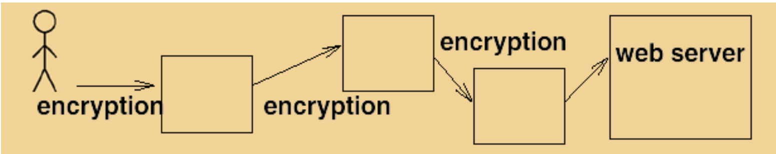
## Privacy-protecting technology (I): Anonymization

Basic principle: Obscure or delete your traces

Proxies or anonymizers *(example: www.anonymizer.com)*




Mix networks, crowds *(ex.s: www.freedom.net, anon.inf.tu-dresden.de)*




- Decentralization of knowledge about browsing histories, goal: complete non-reconstructability (problem: proxies know user ID!)

Implications for Web usage mining: lower-quality data ?!
(cf. “sessionization” problem above)

## Privacy-protecting technology (II): P3P

Basic principle: Negotiate your traces

P3P [W3C00]

- enables Web sites to express their privacy practices in a standard format that can be retrieved and interpreted by user agents
- is an initiative of W3C and industry partners, including Microsoft
- allows the user agent to warn the user, or block communication altogether, if a selected Web site's privacy policy does not comply with user preferences
- rests on XML elements including: who *<RECIPIENT>*, what for *<PURPOSE>*, how much (categories)

Main problems:

- Currently, insufficient browser support and adoption by sites
- Is the legal framework sufficient? What constitutes a violation?
- Is codification possible? (cf. "soft interaction", [SGB01])

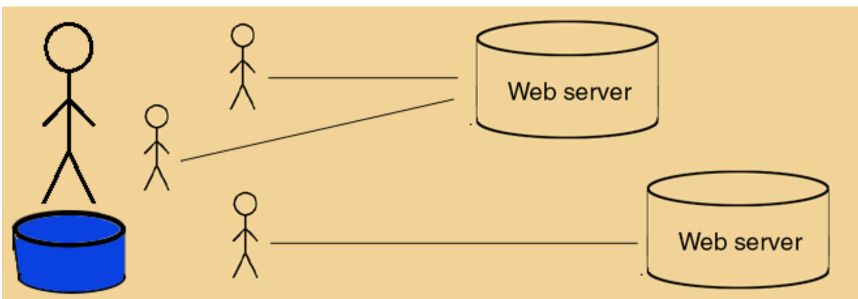
Implications for Web usage mining: same as privacy policies



## Privacy-protecting technology (III): Client-side profiling, pseudonymity, identity management

Basic principle: Control your traces, negotiate your disclosure behavior

- Client-side profiles [SH01]:
  - Users let privacy agents record *all* interactions with *all* Web sites.
  - At the user's discretion, parts of that profile can be made available to marketers or peer networks -> managed via privacy metadata.
- privacy agent should also provide identity management [JM00]:
  - Use new pseudonyms when entering sites, and/or re-use old ones



## Privacy-protecting technology (IV): Client-side profiling, pseudonymity, identity management

- The user privacy agent should also
  - manage default settings,
  - monitor third-party services to bring problems to user's attention,
- Issues to be resolved:
  - Need advanced interfaces to help users adopt a complex technology
  - Requires a well-functioning system of market surveillance, which is fed back to the user agents => a large enough user + contributor base

Implications for Web usage mining: higher-quality data

[cf. PZK01] ?!



## After 9-11

Developments during the last year have led to

- More data being collected by government / official agencies
  - Example: Carnivore
  - Stronger requirements for public and private organisations to archive data, and to hand it over to official agencies under circumstances of suspicion
  - Example: EU Online Privacy Directive
- Technological advancements, e.g. in biometrics

What does this mean

- For the individual user?
- For Web sites and Web usage mining analysts?

## Agenda

- Introduction
- Data Acquisition and Data Preparation
- Evaluation of Web Site Success
- Applications and KDD Techniques for them
- Privacy Concerns
- Research Issues and Future Directions



## Research issues / future directions

Research must (continue to) address the whole cycle of mining:

- The specification of analysis goals
- Data acquisition and preparation
- The mining techniques themselves
- How to put the results into practice
- Societal issues

## Challenges for the specification of goals (I)

The basic goal of Web usage mining is to map the goal of a Web site to questions that can be answered by statistical patterns.

This requires operational definitions of goals

In marketing, examples include:

- Customer segmentation
- Maximizing conversion rates
- Maximizing customer loyalty

even though these do not always provide well-defined indices

Finding operational definitions is more difficult in other areas, such as imaging, brand awareness, …



## Challenges for the specification of goals (II)

Procedures that help to operationalize goals require an interdisciplinary approach, including

- Researchers and practitioners from the application domain to specify application concepts and metrics
- Experts on Web design, customer psychology, etc. to specify knowledge and assumptions concerning user behavior

## Challenges for data acquisition and preparation

- **Data selection / acquisition:**

  The growing importance of multi-channel business models and the observed changes in user expectations and behavior imply

  - challenges for data preparation, modelling, and analysis
  - privacy issues
  - *and also ...* challenges for the overall conception of a site

- **Data cleaning:**
  - Improve sessionization, data reconstruction, robot detection

- **Integration of context and structure:**
  - Mapping requests to the organization's conceptual system
  - Support ad hoc, problem-dependent concept systems, beyond the one implemented in the corporate data warehouse

- **Interactive graphical methods to support these tasks**



## Challenges for mining techniques (1)

**Two complementary dimensions:**

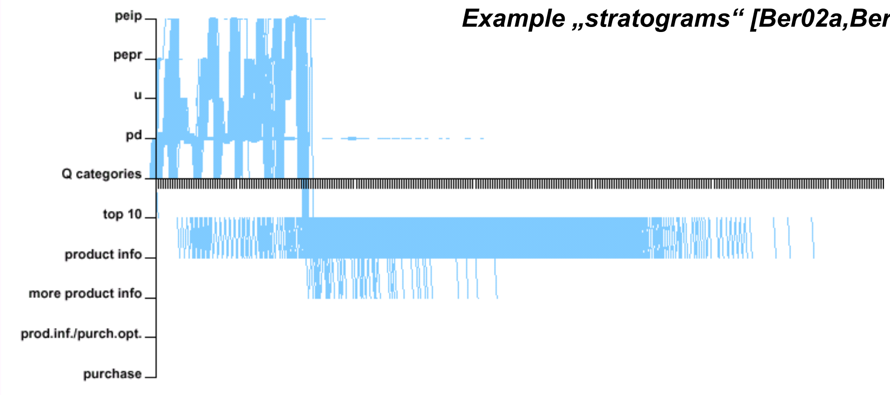
- **Adapt existing mining algorithms to the domain "Web usage"**
- **Develop new algorithms for this domain, to deal with particularities in**
  - problem specification
  - data structures
  - data preprocessing requirements
  - evaluation functions
  - runtime extraction and exploitation of patterns

## Challenges for mining techniques (2): Domain-specific challenges

- **Compare *intended* usage to *actual* usage, given the site structure**

  **Visualization tools that "overlay" intended usage with actual usage, present a good solution to this problem.**

***Example „stratograms" [Ber02a,Ber02b]***






## Challenges for mining techniques (3): More domain-specific challenges

- **Combine and exploit complex data and information structures**
  - **Sets, sequences, parallel user activities -> well-known**
  - ***But …* also need inf. on temporal structure, context,user characteristics**

    **Example: Need to understand user navigation preferences (search vs. browsing, using index structures, etc.) for positioning banner ads, for displaying product catalogs**

## Challenges for mining techniques (4): Integration of background knowledge

Background knowledge can refer to products, site properties, and users:

- It can be explicit (cf. Semantic Web Mining) or implicit expert knowledge
- It can be expressed as interestingness measures, beliefs on usage compared to actual usage [e.g., AT01,Cool02]

posing a strong need for

- statistical and grammatical templates
- interactive tools

to record and exploit it.



## Challenges for Exploitation: Putting the results into practice

Pattern identification: Operationalisation of business terminology for mining

Pattern maintenance: How to describe and store patterns, data warehousing for patterns

Pattern updating: incremental mining [cf. BS02]

## Societal issues

- Do we need more or less E-privacy?
- How can we develop Web usage mining methods to contribute to more (or less) E-privacy?
- Is “opt-in with incentives” (permission marketing) a good idea?
- Do we want to create “digital haves and have-nots”?



## Further issues

Research also needs to address new questions, such as:

- The growing role of the Web in the support of different activities (example E-learning)
- The shift from a focus on Web-related activities in isolation to Web-related activities embedded into a wider context of life
  - Example: mobile computing
- Perceptions change: privacy *vs.* security

## References


[AAP99] R. Agarwal, C. Aggarwal, and V. Prasad. A tree projection algorithm for generation of frequent itemsets. In Proceedings of the High Performance Data Mining Workshop, Puerto Rico, 1999.

[ACR99] Ackerman, M.S., Cranor, L.F., and Reagle, J. Privacy in E-commerce: Examining user scenarios and privacy preferences. In Proceedings of the ACM Conference on Electronic Commerce EC'9 (Denver, CL, Nov). 1999, 1-8.

[Adam01] Adams, Anne. Users' Perceptions of Privacy in Multimedia Communications. PhD Thesis, University College London. 2001. http://www.cs.mdx.ac.uk/RIDL/aadams/thesis.PDF. Access date: 20 June 2002.

[AE01] Antón, A.E. and Earp, J.B. (2001). A Taxonomy for Web Site Privacy Requirements. NCSU Technical Report TR-2001-14, 18December 2001. http://www.csc.ncsu.edu/faculty/anton/pubs/antonTSE.pdf. Access Date: 10 July 2002.

[AT01] Adomavicius, G. and Tuzhilin, A., Expert-driven validation of rule-based user models in personalization applications. Data Mining and Knowledge Discovery, 5 ( 1 / 2), 33-58, 2000.

[BE98] Brusilovsky, P., and Eklund, J. (1998). A study of user model based link annotation in educational hypermedia. Journal of Universal Computer Science, 4 , 429-448.

[Bel00] Belkin, N.J. (2000). Helping people find what they don't know. Communications of the ACM, 43 (8), 58-61.

[Ber02a] Berendt, B. (2002). Using site semantics to analyze, visualize, and support navigation. Data Mining and Knowledge Discovery, 6, 37-59.

[Ber02b] Berendt, B. (2002b). Detail and context in Web usage mining: coarsening and visualizing sequences. In R. Kohavi, B. Masand, M. Spiliopoulou, & J. Srivastava (Eds.), Extended Proceedings of WEBKDD 2001 - Mining Log Data Across All Customer TouchPoints. Berlin etc.: Springer, LNAI 2356.

[BHS02] Berendt, B., Hotho, A., & Stumme, G. (2002). Towards Semantic Web Mining. In I. Horrocks & J. Hendler (Eds.), The Semantic Web - ISWC 2002 (Proceedings of the 1st International Semantic Web Conference, June 9-12th, 2002, Sardinia, Italy) (pp. 264-278). LNCS, Heidelberg, Germany: Springer.




## References


[BMNS02] Berendt, B., Mobasher, B., Nakagawa, M., & Spiliopoulou, M. (2002). The impact of site structure and user environment on session reconstruction in Web usage analysis. In Proceedings of the WebKDD 2002 Workshop at KDD 2002. July 23rd, 2002, Edmonton, Alberta, CA.

[BMSW01] Berendt, B., Mobasher, B.,Spiliopoulou, M. & Wiltshire, J. (2001). Measuring the accuracy of sessionizers for web usage analysis. In Proceedings of the Workshop on Web Mining at SIAM Data Mining Conference 2001 (pp. 7-14). Chicago, IL, April 2001.

[BPW96] P. Berthon, L.F. Pitt and R.T. Watson. The World Wide Web as an advertising medium. Journal of Advertising Research, 36(1), pp. 43-54, 1996.

[Brus97] Brusilovsky, P. (1997). Efficient techniques for adaptive hypermedia. In C. Nicholas and J. Mayfield (Eds.), Intelligent hypertext: Advanced techniques for the World Wide Web, Berlin: Springer. 12-30.

[BS00] Berendt, B. & Spiliopoulou, M. (2000). Analysing navigation behaviour in web sites integrating multiple information systems. The VLDB Journal, 9, 56-75.

[BSH02] Berendt, B., Stumme, G., & Hotho, A. (Eds.) (2001). Proceedings of the Workshop "Semantic Web Mining" at the 13th European Conference on Machine Learning (ECML'02) / 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02), Helsinki, Finland, 20 August 2002. http://ecmlpkdd.cs.helsinki.fi/semwebmine-2002.html

[BSM02] Baron, S. and Spiliopoulou, M., Monitoring the results of the KDD process: An overview of pattern evolution. In J.M. Meij (Ed.) Dealing with the Data Flood: Mining data, text and multimedia. Den Haag, Chapter 5, 2002.

[CMS99] Cooley, R., B. Mobasher, J. Srivastava. 1999. Data preparation for mining world wide web browsing patterns. Journal of Knowledge and Information Systems 1, 5-32.

[Cool00] Cooley, R. (2000). Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data.University of Minnesota, Faculty of the Graduate School: Ph.D. dissertation. http://www.cs.umn.edu/research/websift/papers/rwc_thesis.ps

[CPP01] Chi, E.H., Pirolli, P., Pitkow, J.E. (2000). The scent of a site: a system for analyzing and predicting information scent, usage, and usability of a Web site. In Proceedings CHI 2000 (pp. 161-168).

## References


[CPCP01] Chi, E.-H., Pirolli, P., Chen, K., & Pitkow, J.E. (2001). Using information scent to model user information needs and actions and the Web. In Proceedings CHI 2001 (pp. 490-497).

[CS00] M. Cutler and J. Sterne. E-metrics — Business metrics for the new economy. Technical report, NetGenesis Corp., http://www.netgen.com/emetrics (access date: July 22, 2001)

[DK00] M. Deshpande and G. Karypis. Selective Markov models for predicting Web-page accesses. Technical Report #00-056, University of Minessota, 2000.

[DM02] Dai, H., & Mobasher, B. (2002). Using ontologies to discover domain-level Web usage profiles. In [BSH02].

[DZ97] X. Dreze and F. Zufryden. Testing web site design and promotional content. Journal of Advertising Research,37(2), pp. 77-91, 1997.

[Eigh97] Eighmey, J. (1997). Profiling user responses to commercial web sites. Journal of Advertising Research , 37(2), 59-66.

[Epic97] Electronic Privacy Information Center (1997). Surfer Beware: Personal Privacy and the Internet. http://www.epic.org/reports/surfer-beware.html. Access Date: 10 July 2002.

[Epic99] Electronic Privacy Information Center (1999). Surfer Beware III: Privacy Policies without Privacy Protection. http://www.epic.org/reports/surfer-beware3.html. Access Date: 10 July 2002.

[EU95] Directive 95/46/EC of the European Parliament and the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. http://europa.eu.int/comm/internal_market/en/dataprot/law/. Access date: 10 July 2002.

[EU00] Safe Harbor Privacy Principles. http://europa.eu.int/eurlex/en/consleg/pdf/2000/en_2000D0520_do_001.pdf, http://www.ita.doc.gov/td/ecom/menu.html, and http://www.export.gov/safeharbor/. Access Date: 10 July 2002.




## References


[FBH00] X. Fu, J. Budzik, and K. J. Hammond. Mining navigation history for recommendation. In Proc. 2000 International Conference on Intelligent User Interfaces, New Orleans, January 2000. ACM.

[FGL+00] J. Forsyth and T. McGuire and J. Lavoie. All visitors are not created equal. McKinsey marketing practice. McKinsey & Company. Whitepaper. 2000.

[Flem98] Fleming, J. (1998). Web Navigation. Designing the User Experience. Sebastopol, CA: O'Reilly.

[GS02] Garfinkel, S., with Spafford, G. (2002). Web Security, Privacy & Commerce. 2nd Ed. Sebastopol, CA: O'Reilly.

[Jane99] Janetzko, D. (1999). Statistische Anwendungen im Internet. Daten in Netzumgebungen erheben, auswerten und präsentieren. München, Germany: Addison-Wesley.

[JFM97] T. Joachims, D. Freitag, and T. Mitchell. Webwatcher: A tour guide for the world wide web. In the 15th International Conference on Artificial Intelligence, Nagoya, Japan, 1997.

[JM00] Jendricke, U. and Gerd tom Markotten, D. Usability meets security - The Identity Manager as your personal security assistant for the Internet. In Proceedings of the 16th Annual Computer Security Applications Conference (New Orleans, LA, Dec.). 2000.

[KNY00] Kato, H., Nakayama, T., & Yamane, Y. (2000). Navigation analysis tool based on the correlation between contents distribution and access patterns. In Working Notes of the Workshop "Web Mining for E-Commerce - Challenges and Opportunities." 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. August 20-23, 2000. Boston, MA. pp. 95-104. Available at http://robotics.stanford.edu/~ronnyk/WEBKDD2000/papers/kato.pdf. Access Date: 10 July 2002.

[Kuhl96] R. Kuhlen. Informationsmarkt: Chancen und Risiken der Kommerzialisierung von Wissen. 2nd edition, 1996 (on German)

[LAR00] W. Lin, S.A. Alvarez, C. Ruiz. Collaborative recommendation via adaptive association rule mining. In Proceedings of the Web Mining for E-Commerce Workshop (WebKDD'2000), August 2000, Boston.

## References


[LHM99] B. Liu, W. Hsu, and Y. Ma. Association rules with multiple minimum supports. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-99, poster), San Diego, CA, August 1999.

[Lieb95] H. Lieberman. Letizia: An agent that assists web browsing. In Proc. of the 1995 International Joint Conference on Artificial Intelligence, Montreal, Canada, 1995.

[LPS+00] Junghoung Lee, M. Podlaseck, E. Schonberg, R. Hoch and S. Gomory. Analysis and visualization of metrics for online merchandizing. In "Advances in Web Usage Mining and User Profiling: Proc. of the WEBKDD'99 Workshop", LNAI 1836, Springer Verlag, pp. 123-138, 2000.

[Maye97] Mayer-Schönberger,V.1997.The Internet and privacy legislation: Cookies for a treat? West Virginia Journal of Law & Technology 1. http://www.wvu.edu/~wvjolt/Arch/Mayer/Mayer.htm. Access Date: 10 July 2002.

[MDL+00] B. Mobasher, H. Dai, T. Luo, Y. Su, and J. Zhu. Integrating web usage and content mining for more effective personalization. In E-Commerce and Web Technologies , volume 1875 of LNCS . Springer Verlag, Sept. 2000.

[MDLN01] B. Mobasher, H. Dai, T. Luo, M. Nakagawa. Effective personalization based on association rule discovery from Web usage data. In Proceedings of the 3rd ACM Workshop on Web Information and Data Management (WIDM01), held in conjunction with the International Conference on Information and Knowledge Management (CIKM 2001), ACM Press, Atlanta, November 2001.

[MDLN02] Mobasher, B., H. Dai, T. Luo, and M. Nakagawa 2002. Discovery and evaluation of aggregate usage profiles for Web personalization. Data Mining and Knowledge Discovery 6, 61-82.

[Moe] W. Moe. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. In Journal of Consumer Psychology.

[Niel96] Nielsen, J. (1996). Top Ten Mistakes in Web Design. Alertbox for May 1996. http://www.useit.com/alertbox/9605.html. Access Date: 10 July 2002.

[Niel99] Nielsen, J. (1999). "Top Ten Mistakes" Revisited Three Years Later. Alertbox, May 2, 1999. http://www.useit.com/alertbox/990502.html. Access Date: 10 July 2002.




## References


[Niel00] Nielsen, J. (2000). Designing Web Usability: The Practice of Simplicity. New Riders Publishing.

[Niel01] Nielsen, J. (2001). Usability Metrics. Alertbox, January 21, 2001. http://www.useit.com/alertbox/20010121.html. Access Date: 10 July 2002.

[Obe00] Oberle, D. Semantic Community Web Portals - Personalization. Studienarbeit. Universität Karlsruhe, 2000.

[PP99] J. Pitkow and P. Pirolli. Mining longest repeating subsequences to Predict WWW Surfing. In Proceedings of the 1999 USENIX Annual Technical Conference, 1999.

[PS02] C. Pohle, M. Spiliopoulou. Building and exploiting ad hoc concept hierarchies for Web log analysis. In Proc. of DaWaK 2002, Aix en Provence, France, Springer Verlag, Sept. 2002.

[PZK01] Padmanabhan,B.,Z.Zheng,S.O.Kimbrough.2001.Personalization from incomplete data: What you don't know can hurt. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,San Francisco,CA.154-163.

[SA95] Srikant, R., & Agrawal, R. (1995). Mining Generalized Association Rules. In Proceedings of the 21st International Conference on Very Large Databases (pp. 407-419). Zurich, Switzerland, September 1995.

[SF99] Spiliopoulou, M., L.C. Faulstich. 1999. WUM: a tool for Web utilization analysis. In Proceedings EDBT (Workshop WebDB'98), LNCS 1590, Berlin, Germany: Springer. 184-203.

[SGB01] Spiekermann, S., Grossklags, J., and Berendt, B. E-privacy in 2nd generation E-Commerce: privacy preferences versus actual behavior. In Proceedings of the ACM Conference on Electronic Commerce (EC'01). (Tampa, FL, Oct.). 2001, 38-47.

[SH01] Shearin, S. and Liebermann, H. Intelligent profiling by example. In Proceedings of the ACM Conference on Intelligent User Interfaces (Santa Fe, NM, January). 2001.

## References


[SHB01] Stumme, G., Hotho, A., & Berendt, B. (Eds.) (2001). Freiburg, Germany, 3 Proceedings of the Workshop "Semantic Web Mining" at the 12th European Conference on Machine Learning (ECML'01) / 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01), September 2001. http://semwebmine2001.aifb.uni-karlsruhe.de.

[Shne98] Shneiderman, B. (1998). Designing User Interface. Strategies for Effective Human-Computer Interaction. 3rd edition. Reading, MA: Addison-Wesley.

[SMBN03] Spiliopoulou, M., Mobasher, B., Berendt, B., & Nakagawa, M. (2003). A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analyis. To appear in INFORMS Journal on Computing, 15.

[SP01] M. Spiliopoulou,C.Pohle. Data mining for measuring and improving the success of Web sites. In Journal of Data Mining and Knowledge Discovery, Special Issue on E-commerce, 5, pp. 85–114. Kluwer Academic Publishers. 2001

[Spen99] Spendolini, M. (1999). Customer Measurement Systems - Opportunities for Improvement. White paper, MJS Associates, accenture CRM Portal. http://www.crmproject.com/documents.asp?d_ID=753. Access Date: 10 July 2002.

[Spi99] M. Spiliopoulou. The laborious way from data mining to Web mining. Int. Journal of Comp. Sys., Sci. & Eng., Special Issue on "Semantics of the Web", 14, pp. 113:126, 1999.

[SPT02] Spiliopoulou, M., Pohle, C., and Teltzrow, M. (2002). Modelling and Mining Web Site Usage Strategies.To appear in Proceedings of the Multi-Konferenz Wirtschaftsinformatik, Nürnberg, Germany, 9-11 September.

[Sul97] T. Sullivan. Reading reader reaction: A proposal for inferential analysis of web server log files. Proc. of the Web Conference'97, 1997.

[Trus00] TrustE. (2000). TrustE Online Privacy Resource Book. http://www.truste.org/about/oprah.doc. Access Date: 10 July 2002.

[Usab99] The Usability Group. (1999). What is Strategic Usability? http://usability.com/umi_what.htm. Access Date: 10 July 2002.




## References


[Volo00] Volokh, E. (2000). Personalization and privacy. Communications of the ACM, 43(8), 84-88.

[WB90] Warren, S. and Brandeis, L. The right of privacy. Harvard Law Review, 4, 193.

[West67] Westin, A. (1967). Privacy and Freedom. Boston: Atheneum Press.

[W3C00] W3C. The Platform for Privacy Preferences 1.0 (P3P1.0) Specification. http://www.w3.org/TR/2000/CR-P3P-20001215 and http://www.w3.org/TR/P3P. Access Date: 10 July 2002.