Tutorial: Data Mining for the Web

Myra Spiliopoulou

http://www.wiwi.hu-berlin.de/~myra

Institute of Information Systems, Faculty of Economics

Humboldt University Berlin

PKDD'99

Prague, 15.9.99

Contents

1 Introduction

Mining any data: The standard procedure

1. Problem specification
2. Data gathering
3. Data preparation
4. Establishment of quality criteria
5. Data mining
6. Evaluation and interpretation of the results
7. Putting the results into praxis
8. `GO TO 1`

Which problem do we pursue when mining the web?

Abstract goal:

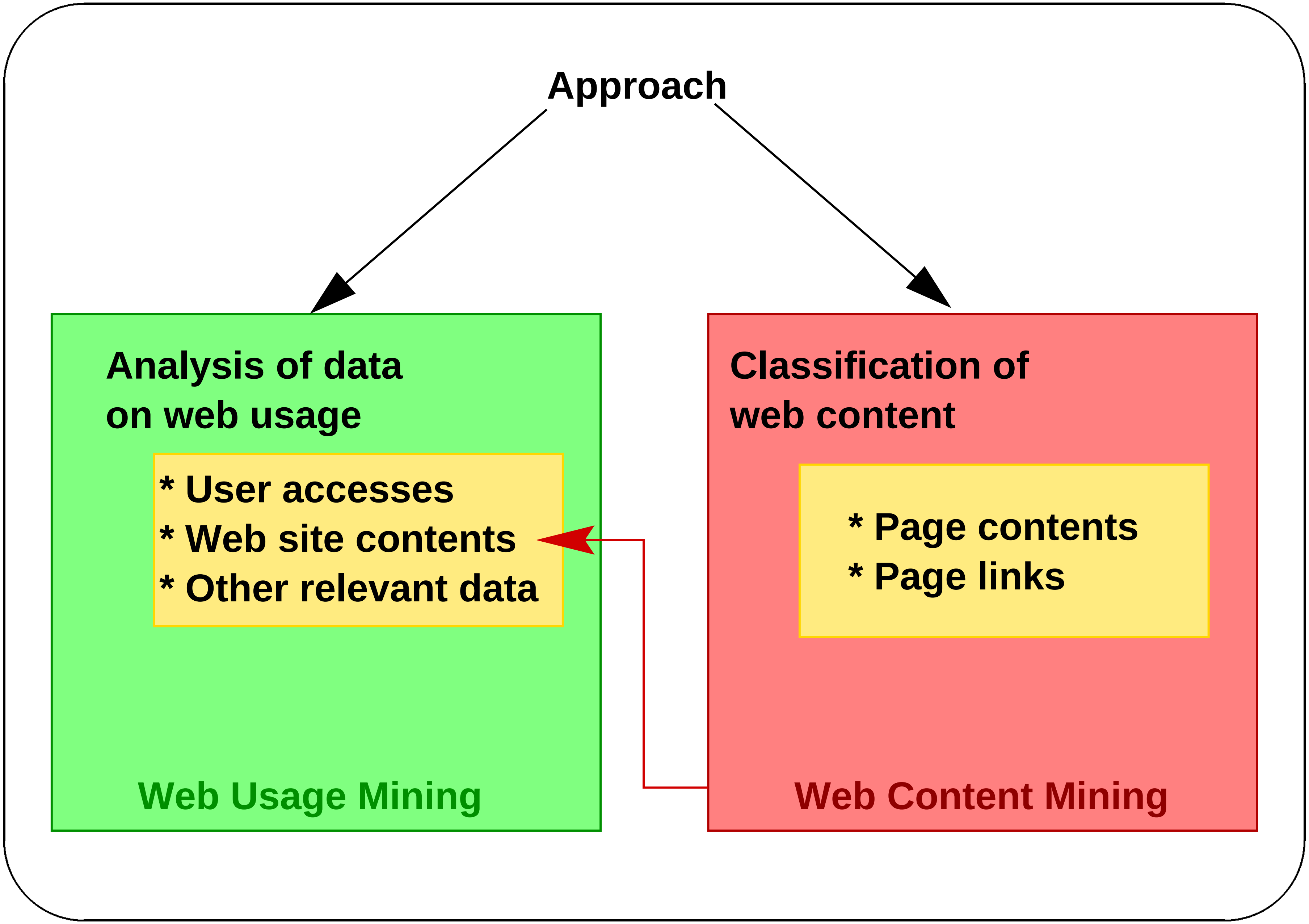
Assist the user in finding something interesting/important/useful for her and for the site owner.

User-centric problem specifications:

- ⋆ discovery of documents on a subject
- ⋆ discovery of semantically related documents
- ⋆ prevention of disorientation

Owner-centric problem specifications:

- ⋆ increase of contact and conversion efficiency (web marketing) [2]
- ⋆ targeted promotion of goods
- ⋆ development of personalized services


Approach
Analysis of data
on web usage
* User accesses
* Web site contents
* Other relevant data
Web Usage Mining
Classification of
web content
* Page contents
* Page links
Web Content Mining

2 Web Content Mining

IR - Viewpoint: The web site consists of hyper documents

DB - Viewpoint: The website consists of semi structured documents

2.1 Web content mining — IR Viewpoint

We try to discover the semantics of the documents by exploiting

- ⋆ their contents and
- ⋆ the links to and from them
- ⋆ the knowledge of experts
- ⋆ the selections of the users

Approach: Text Mining ($\surd$) extended to HyperText Mining

Exploiting Links to Classify Documents

Motivation:

The content of a document is reflected in:

1. the documents it cites
2. the documents that cite it

similarly to articles and books.

Methodology:

Incorporation of text or terms from the document's neighbours into the mining process

- ⋆ treating these new data similarly to the document's data
- ⋆ treating these new data separately

The approach of Chakrabarti et al in TAPER [5]

Goal:

Indexing of web documents
using text classification techniques
and exploiting information on the neighbours

Observation:

Simple exploitation of information from the neighbours
decreases the quality of the results

Methodology:

Classification based on

* the text of the document
* the classes of its pre-classified neighbours

classifying neighbours iteratively

The approach of Perkowitz and Etzioni in PageGather [14]

Goal:

Discovery of groups of pages that are

- semantically related according to the users
- not linked to each other

Visit-coherence assumption

Method:

Grouping of documents into overlapping clusters

Algorithm:

- Establishment of a similarity matrix on the co-occurence frequency of non-linked pages
- Construction of a graph over the matrix
- Discovery of the cliques in the graph, subject to size and pair similarity constraints
- Construction of an index page for the pages in each cluster

2.2 Web content mining — DB Viewpoint

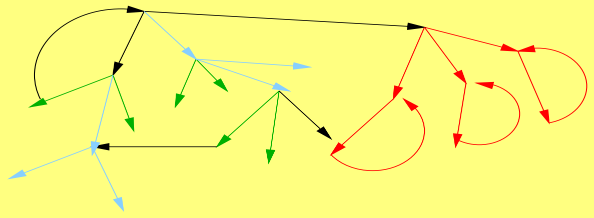
We try to extract the schema of the documents, in which at least the most important semantics are reflected.

Approach: Schema Discovery for semistructured documents [9, 18]

Discovering Structure and Semantics in a Web Site

TSIMMIS [9]

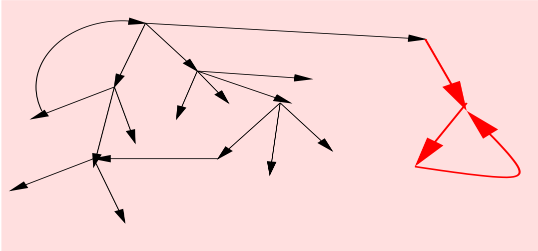
A web site is observed as a directed graph:
node = page
edge = link



Graph reduction by merging

Group of Ke Wang [18]

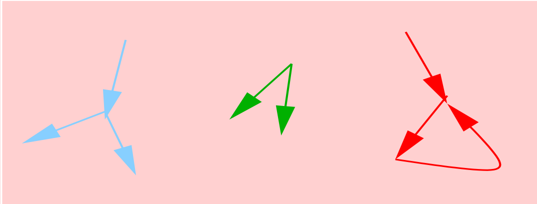
equivalent paths



The reduced graph is a quasi-schema.

frequent subgraphs

- Discovery of frequent nodes
- Expansion into frequent subgraphs



A frequent subgraph is a quasi-subschema.

Web Usage Mining

1. We record the behaviour of the user.

2. We mine the records, which reflect how the web site is (subjectively!) perceived

3. We try to
   - ⋆ adjust the site to the perception of the user.
   - ⋆ bias the user's behaviour towards satisfying the objective goals of the site.

Web usage mining can be:

Impersonalized

The user is observed as a unit of unknown identity, for at most some properties might be assessable from demographic data

cf. Market basket analysis

Personalized

The user is observed as a physical person, for which identifying information and personal data/properties are known

cf. Analysis of customer data

3 Web Usage Mining / Impersonalized

Recall: **The standard KDD procedure**

1. Problem specification $\surd$
2. Data gathering
3. Data preparation
4. Establishment of quality criteria $\rightarrow 6$
5. Data mining
6. Evaluation and interpretation of the results
7. Putting the results into praxis
8. `GO TO 1`

3.1 Data Gathering

i. Web Server Log

ii. Site description data

iii. Concept hierarchies

i. Web Server Log

According to the HTTP specifications of W3C

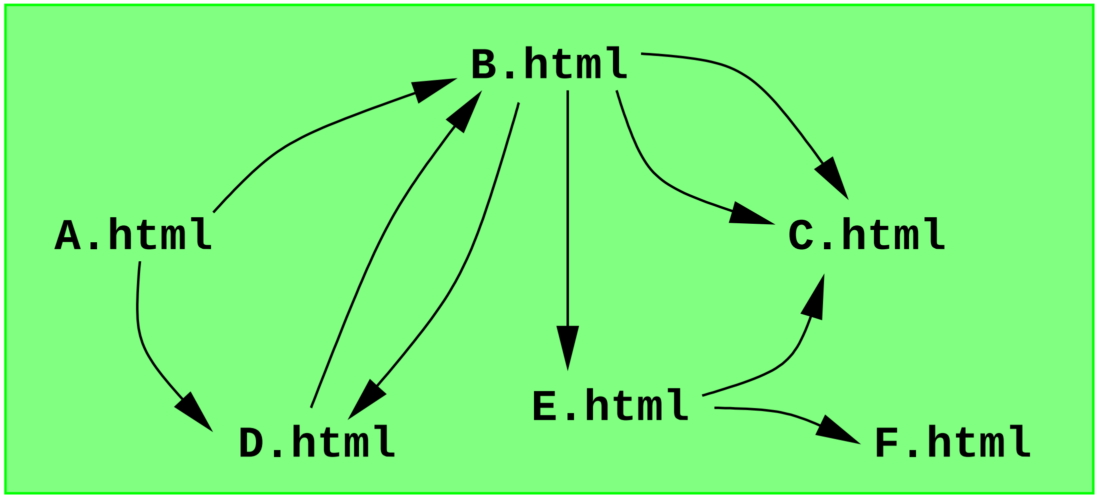
the Common Log contains

- IP address of the client's host
- Date and time of the access
- URL requested
- Modus of request (`GET`, `POST`, `HEAD`)
- Result code
- ...
- Username and password *if* the server requires user authentication

Optionally recorded: Referrer Log & Agent Log

ii. Site description data

⋆ Description of the site topology



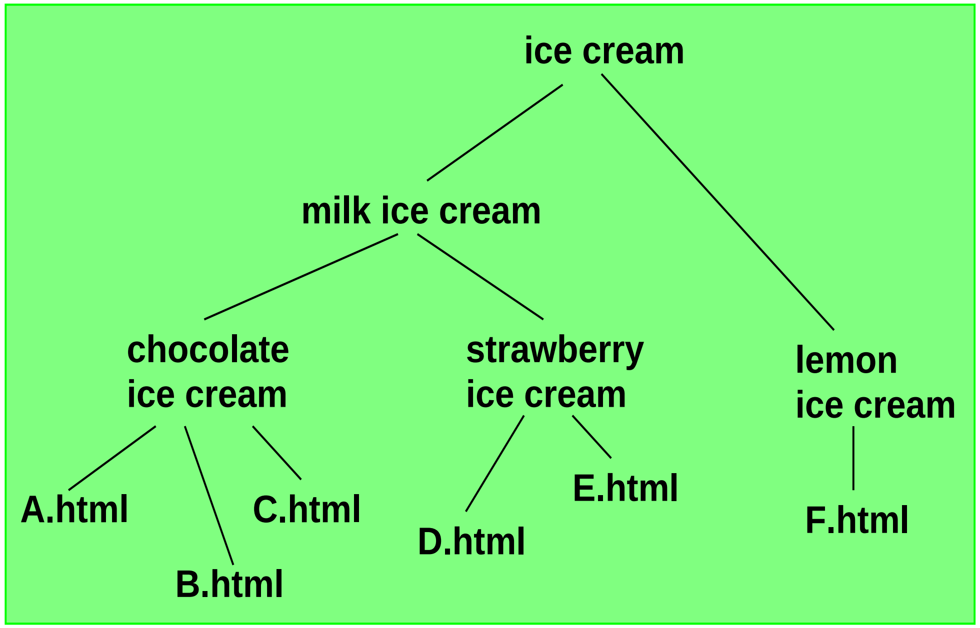
⋆ Meta-data on each web page, such as:

- page description records
- HTML tags on the page
- XML tags on the page, according to a known DTD

iii. Concept hierarchies

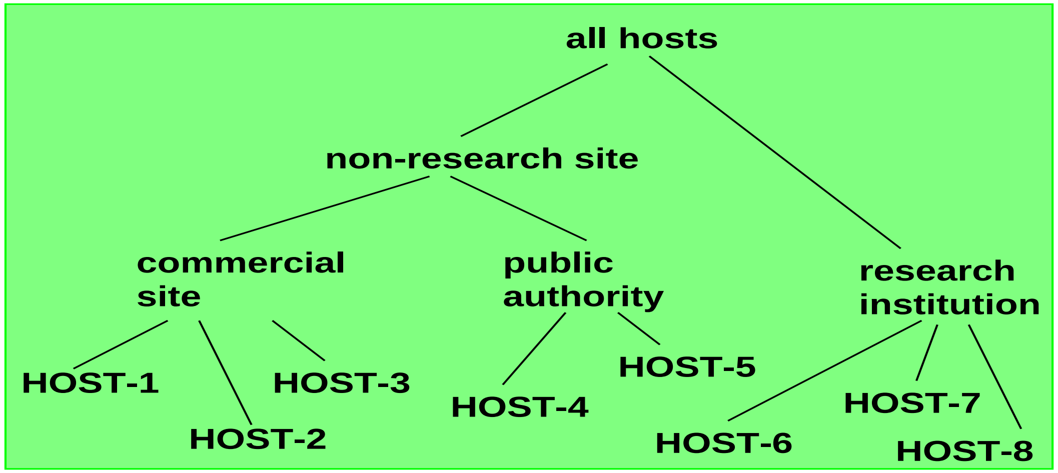
Concept hierarchies reflecting and generalizing the semantics of objects can be constructed upon:

⋆ the contents of the web pages



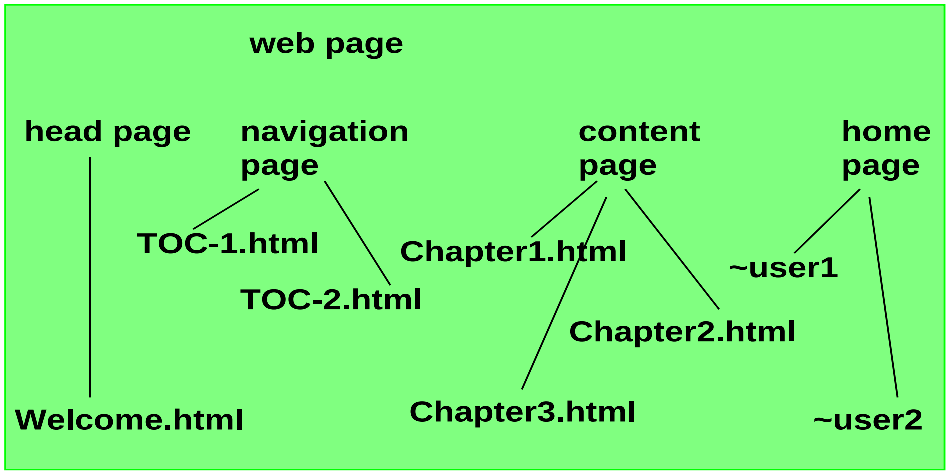
SurfAid discovers [10]
concept hierarchies of products from:
* page tags
* correlations among products
* analysis of page contents

⋆ the hosts of the users



In MIMIC, [4]
hosts can be organized into a concept hierarchy using OLAP technology.

⋆ the types of the web pages



In WebSift, [7]
pages are categorized according to their intended usage and links/content ratio.

3.2 Data Preparation

Goal:

Restore the users' activities in the web server log in a reliable way

i. Distinguishing among users

ii. Building sessions

iii. Restoring the contents of a session

iv. Removing undesirable entries

i. Distinguishing among users

Interleaved activities of multiple users from the same host (or proxy) are recorded as activities of a single user.

Avoiding the problem

- ⋆ User authentication
- ⋆ Cookies
- ⋆ Personalized pages

vs privacy and security concerns

Remedying the problem using:

- ⋆ the optional logs
- ⋆ the site topology
- ⋆ some rules of thumb

in heuristics

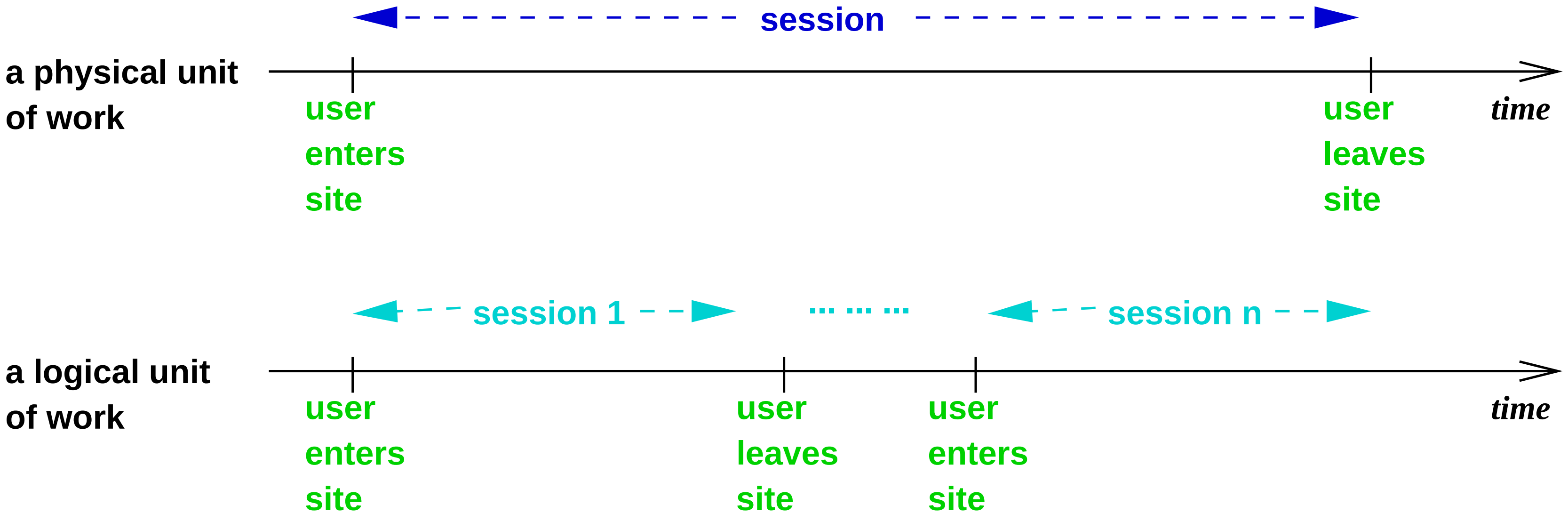
Some heuristics to distinguish among users ← **WebSift** [7]

Consecutive entries in the web server log originate from different users in the same host/proxy if:

1. Different agents are used — Agent log
2. URL requests are performed from different referrers — Referrer log
3. A URL is requested, which is not linked to any previously visited page — Web site topology

ii. Building sessions

A session as:



Heuristics identifying the end of a session

⋆ *Time-based criteria*

- upper limit on the duration of page visit
- upper limit on the total session duration

⋆ *Context-based criteria*

- a page is accessed that is not reachable via the previously visited page(s)
- a page of a specific type is reached
- a conceptual unit of work is completed

iii. Restoring the contents of a session

Some activities of the user are not recorded due to caching.

Ignoring the problem when:

⋆ studying correlated purchases within a session

⋆ focussing on the first access to each page only

Avoiding the problem

⋆ Non-cachable pages *vs* performance decay

Remedying the problem using:

⋆ the site topology

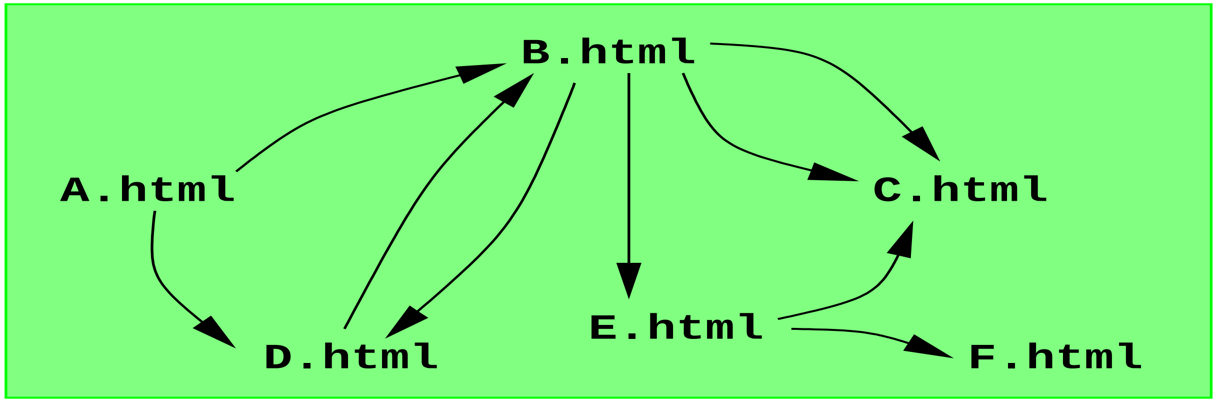
⋆ some rules of thumb in heuristics

A heuristic to restore session contents ← **WebSift** [7]

A previously visited page / sequence of pages is accessed again if:

⋆ The requested URL is not linked to the page visited before it but can be reached through pages in the cache

Web site topology



Original session: A.html - B.html - C.html - D.html

Restored session: A.html - B.html - C.html - B.html - D.html

iv. Removing undesirable entries

The page accesses relevant for the analysis depend on the goal of the analysis

- ⋆ *Optimize the server performance*
  All entries are relevant
- ⋆ *Study the usage of the site*
  Accesses to images should be removed → double recording
- ⋆ *Study the behaviour of users*
  - Accesses to images should be removed → double recording
  - Robot accesses should be removed → robots are not humans

Heuristics to remove robot entries

Robot entries can be recognized and removed by:

- ⋆ Name
  http://info.webcrawler.com/mak/projects/robots/active.html
- ⋆ Any characteristic (known) string ⟵ WUM [16]
- ⋆ Odd behaviour:
  - repeated accesses on the same page ⟵ WUM [16]
  - consecutive accesses in very short time intervals ⟵ WUM [16]
  - deviation from behavioural patterns known to originate from humans ⟵ SurfAid [10]

3.3 Data Mining

Base Methodologies for the Analysis

- ⋆ OLAP
- ⋆ Analysis of usage trends
- ⋆ Association rules' discovery $\surd$
- ⋆ Navigation pattern discovery
- ⋆ Discovery of correlated pages $\surd$

Exploitation of OLAP services for:

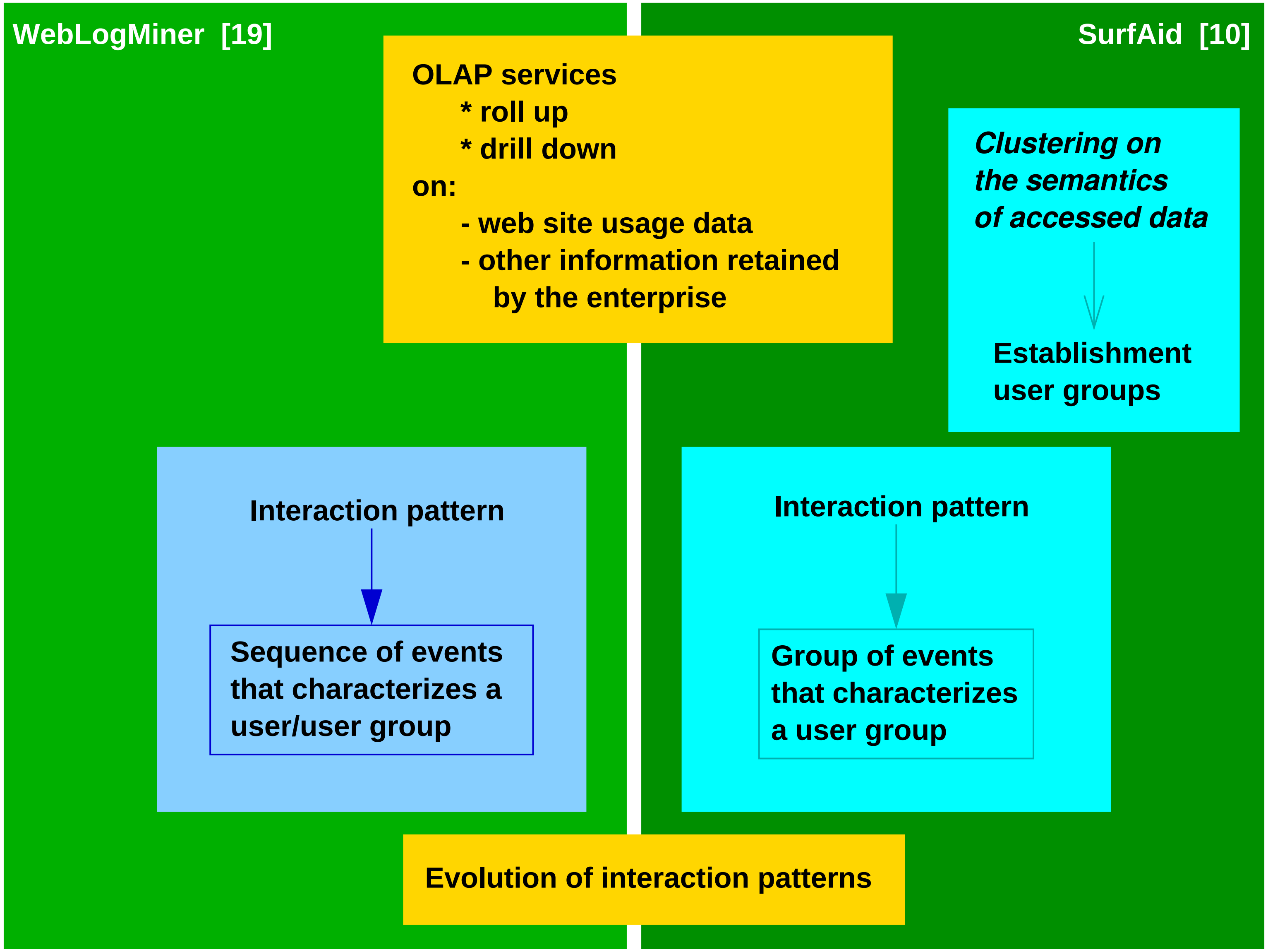
- ⋆ Computation of statistics across different dimensions
- ⋆ Establishment and observation of individual usage patterns
- ⋆ Computation and observation of trends for various objects
- ⋆ . . .

Establishment of a hypercube, whose dimensions describe:

- ⋆ Data on the site and its usage
  - URLs
  - Resources
  - Servers
  - Users
  - Access time
  - Access duration
  - Agents (of clients)
  - Client domains

  ← WebLogMiner [19]

- ⋆ domain knowledge
  - management strategies, policies and rules
  - customer data
  - product data
  - demographic data


WebLogMiner [19]
SurfAid [10]
OLAP services
* roll up
* drill down
on:
- web site usage data
- other information retained
by the enterprise
Clustering on
the semantics
of accessed data
Establishment
user groups
Interaction pattern
Sequence of events
that characterizes a
user/user group
Interaction pattern
Group of events
that characterizes
a user group
Evolution of interaction patterns

⋆ Observing pattern evolution in the **WebLogMiner** [19]

Users/Groups become more experienced in the usage of the site:

- Initially, lots of pages are investigated:
  The site is being explored.
- Later, some pages disappear from the interaction pattern:
  Only some of the pages are being revised.
- As time proceeds, the interaction pattern becomes compacter:
  The users concentrate on the accomplishment of specific tasks.

⋆ Evolution of rule patterns ← Chakrabarti et al [6]

Assumption: Not the correlations themselves are interesting, but the unexpected changes in their statistics over time.

Approach: Partitioning of the time axis into intervals, at which rule patterns show a remarkable change.

Analyzing patterns of **navigation** to:

⋆ study (and bias) the behaviour of the users

⋆ to improve the site's design

Navigation pattern discovery vs

· Association rules' discovery
· Sequence mining

From a dataset containing information about all activities of the users in the site

- ⋆ Association rules discover correlations between objects: pages, products, advertisements
- ⋆ Sequence miners return frequently occuring event sequences

Is this all we need to know?

Fictious association rule (or sequence):

Rule	*support*	*confidence*	*improvement*	χ^2-*test*
Welcome.html orders.html	25%	50%	> 1	$+$

Interesting?

Actionable?

This rule says:

$\top$ At least 25% of the visitors purchase something.

$\surd$ Purchasing visitors often start at the Welcome page.

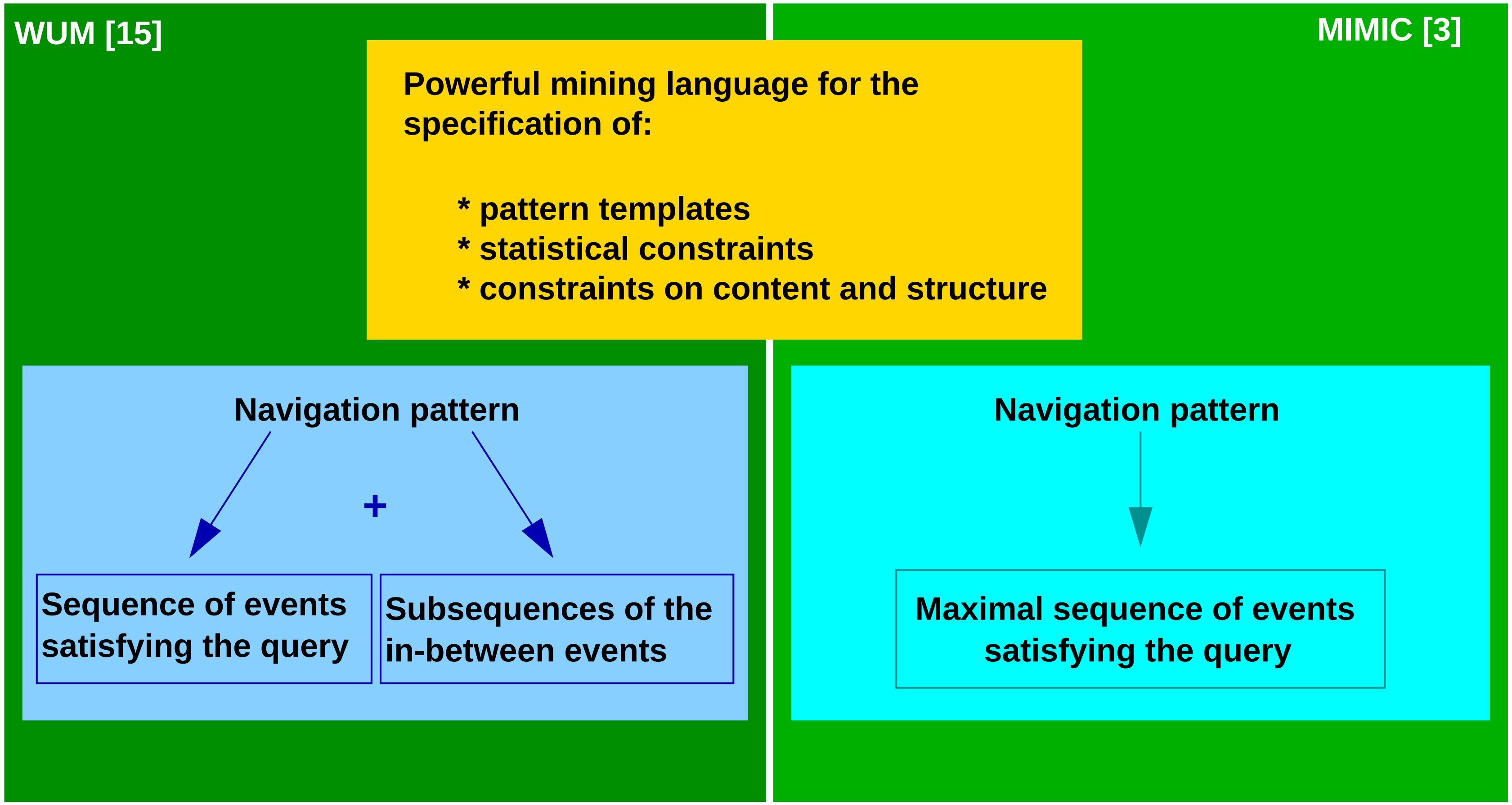
$\perp$ 50% of the visitors going through the Welcome page do not purchase anything.

Recall the problem specifications:

- ⋆ Increase the site's contact/conversion rate
- ⋆ Detect problems of disorientation
- ⋆ Detect discrepancies between expected and actual site usage
- ⋆ ...

What should the miner's results be to help the human solve such problems?

Web tailored miners for the discovery of navigation patterns



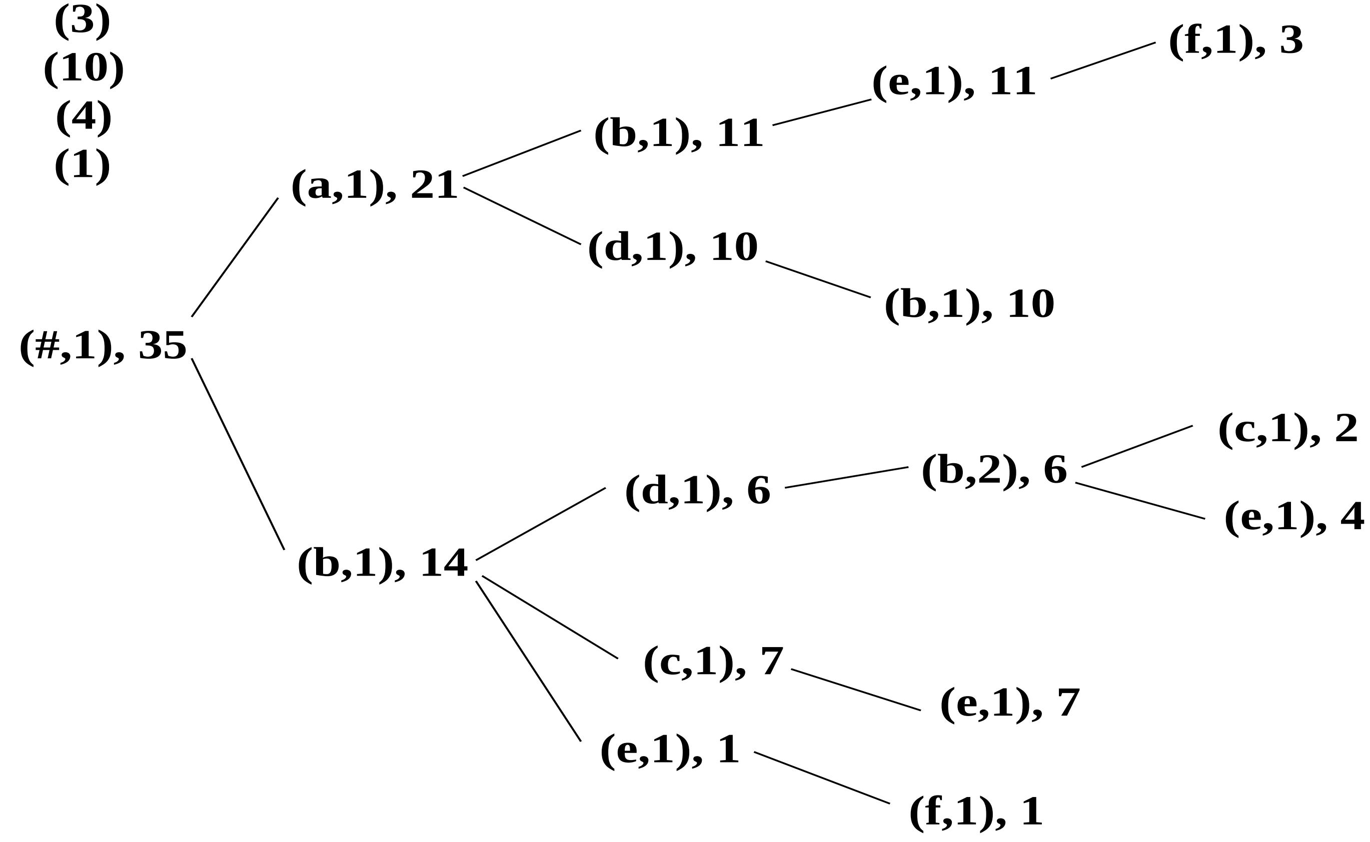
Example

Fictious web server log after preprocessing:

Session Num	*Sequence*	*Num of occurences*
s1	a b e	8
s2	b d b c	2
s3	b c e	7
s4	a b e f	3
s5	a d b	10
s6	b d b e	4
s7	b e f	1

The web server log in aggregated tree format:

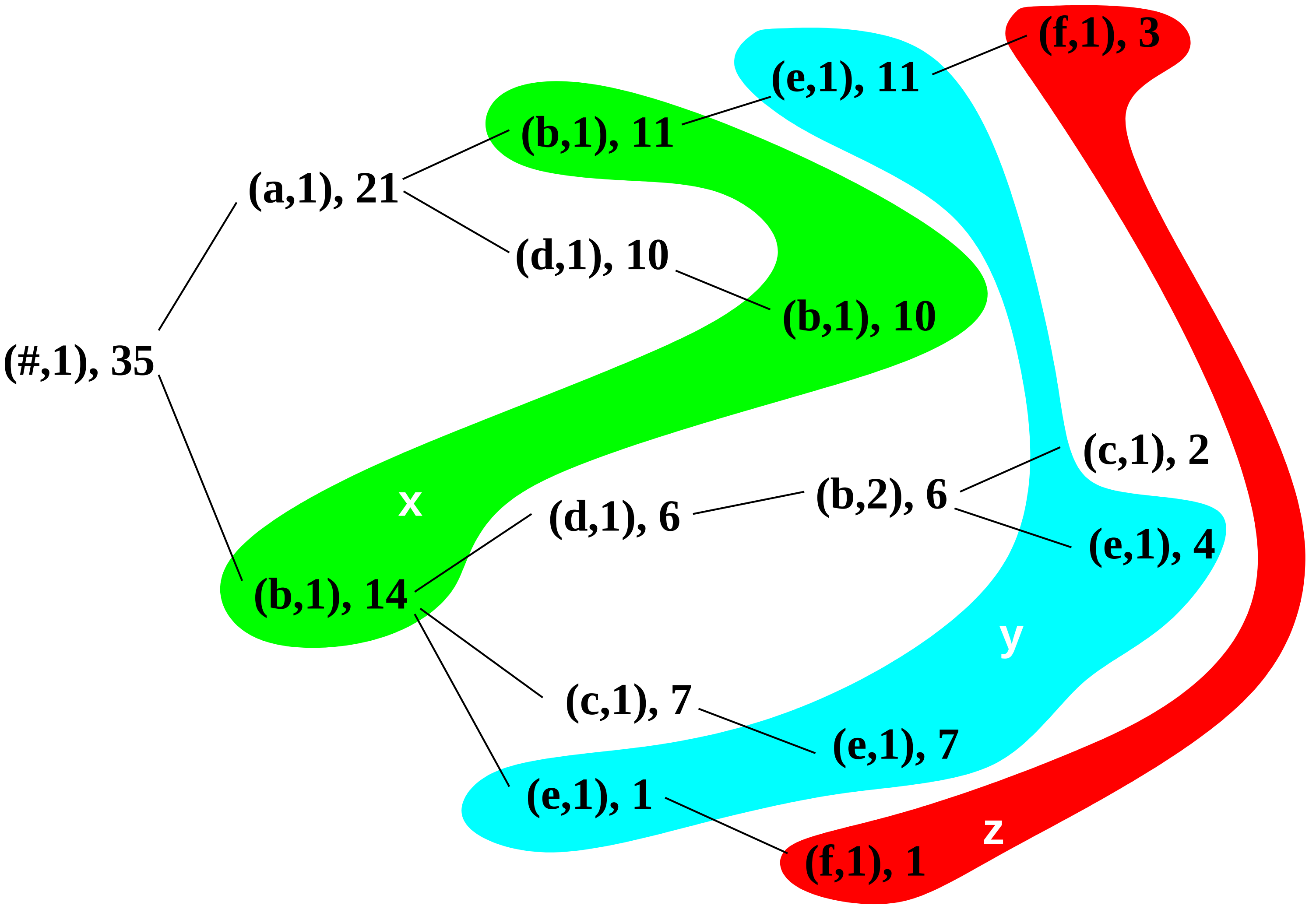
s1.	**a b e**	**(8)**
s2.	**b d b c**	**(2)**
s3.	**b c e**	**(7)**
s4.	**a b e f**	**(3)**
s5.	**a d b**	**(10)**
s6.	**b d b e**	**(4)**
s7.	**b e f**	**(1)**



A mining query in MINT (WUM's mining language) [16]:

SELECT t	template name
FROM NODE AS x y z,	3 named variables
TEMPLATE x * y * z AS t	structural
WHERE x.support >= 20	
AND (y.support / x.support) >= 0.5	statistical constraints
AND (z.support / y.support) >= 0.15	

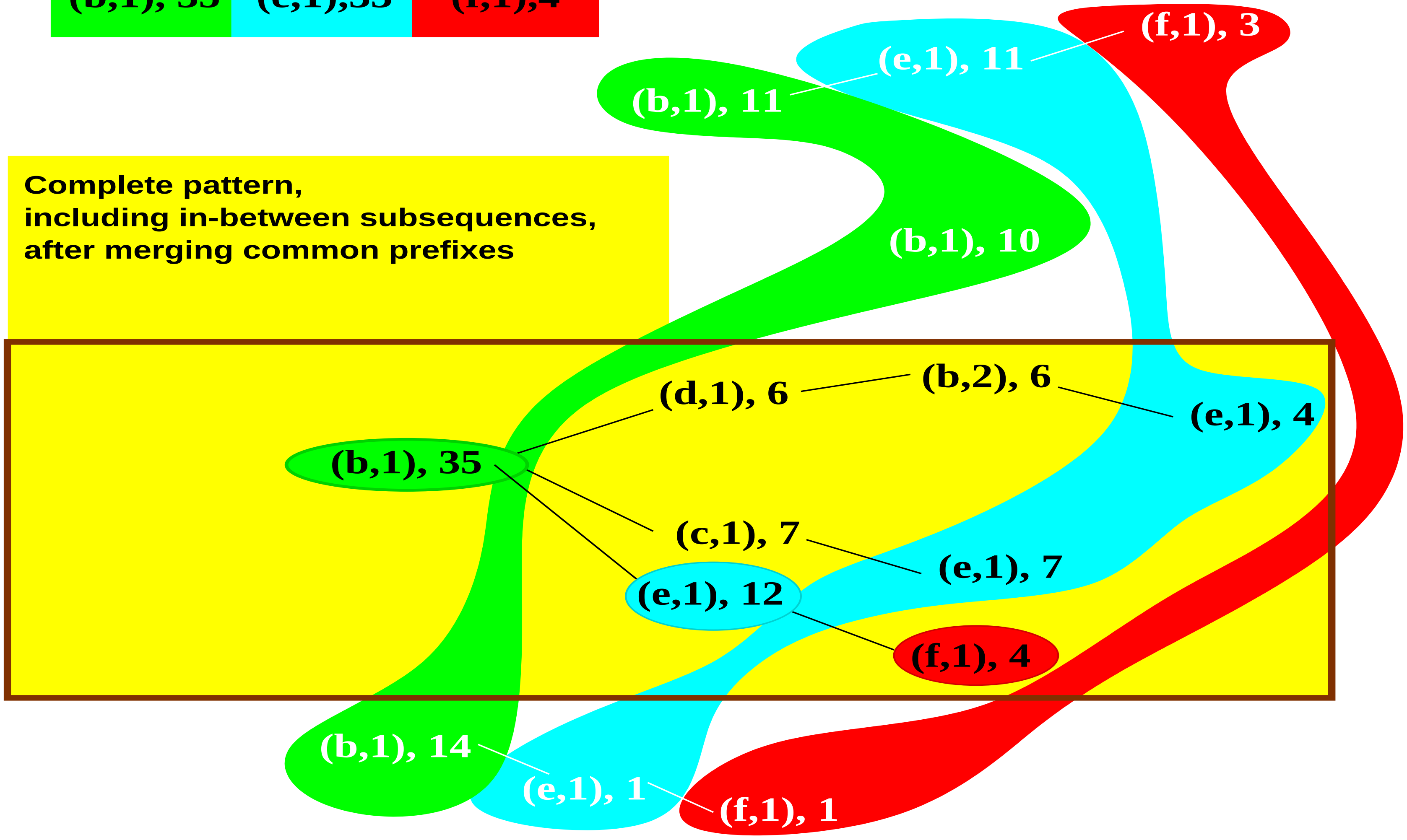
A triplet of variable bindings satisfying the query



The resulting navigation pattern

Sequence of events satisfying the query:

(b,1), 35	(e,1),33	(f,1),4



3.4 Being confronted with the mining results

Web usage objects
are mapped into
application objects

web page

product

Exploitation of
conventional,
application-specific
measures

support,
confidence, ...

product
correlations

Interpretation of
the results
according to the practices
in the domain

Web usage objects
are analyzed in the framework of
web usage applications

+

Exploitation of
conventional generic
measures

Results that are
difficult to interpret
and to act upon

Web usage objects
are analyzed in the
framework of
web usage applications

Mapping of the measures
for conventional applications
into measures for
web applications.

* semantic coherency of web pages [14]
* conversion efficiency of a site [17]
* interestingness of navigation patterns [1,8,12]

... ...

Actionable results

4 Web Usage Mining / Personalized

Combining:

+ Web Usage Mining / Impersonalized

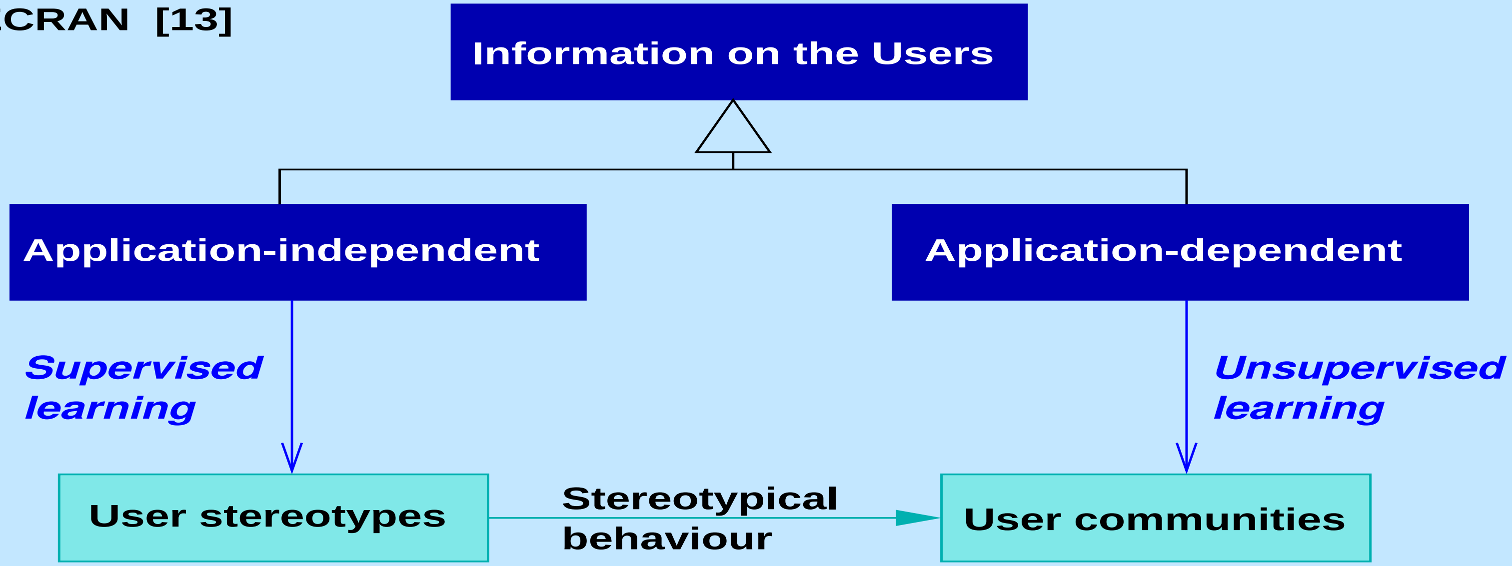
  with simplified data preparation: only caching problems must be dealt with

+ Recordings of user interests

  building user groups/signatures/profiles User Modelling

+ information on the site ← Web Content Mining

ECRAN [13]



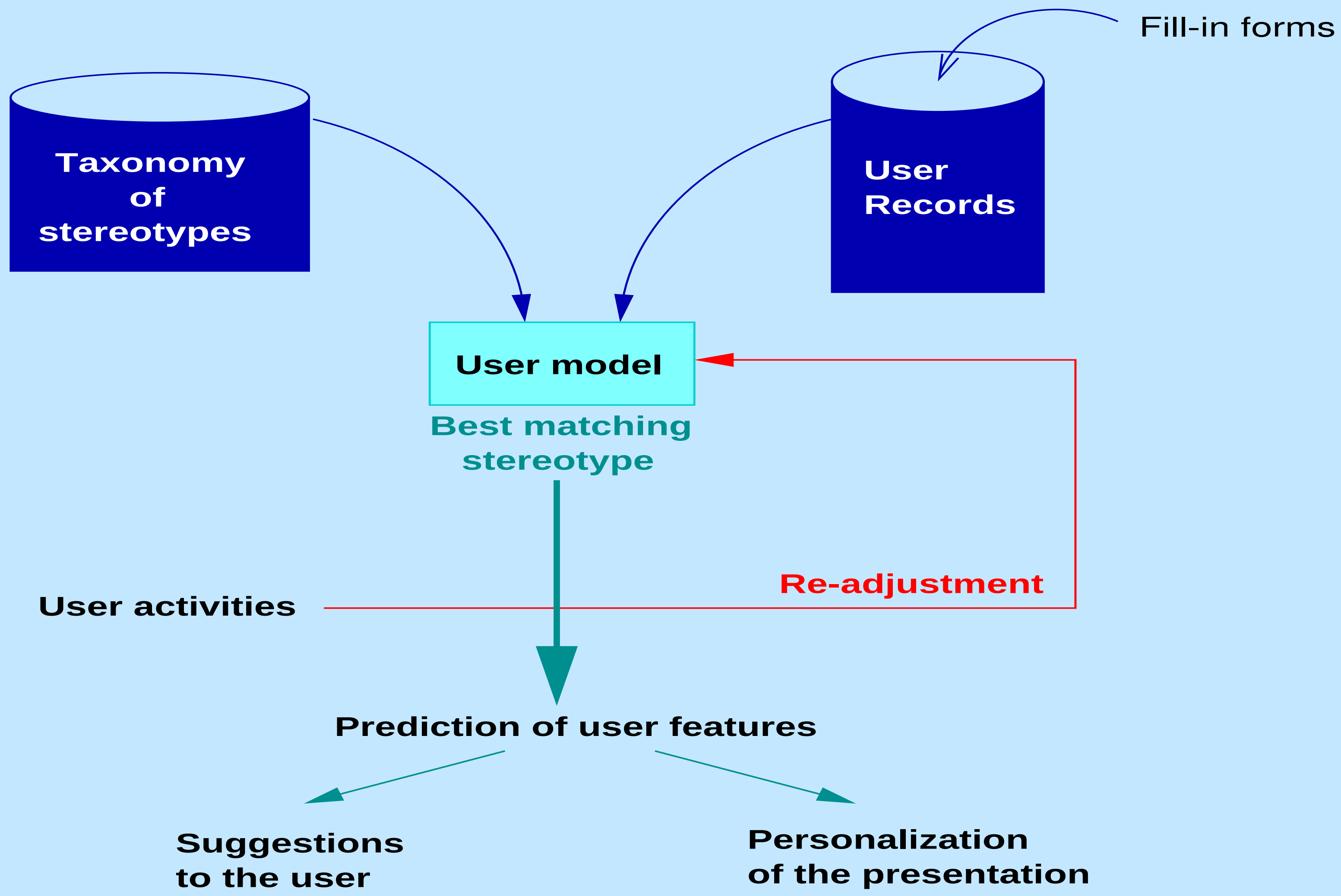
Algorithm:
Decision trees

Goal: Discovery of representative communities

Algorithm:
hierarchical clustering

Metric:
Frequence of appearance of data properties within each community
vs
within the whole dataset

SETA [20]


Fill-in forms
Taxonomy of stereotypes
User Records
User model
Best matching stereotype
Re-adjustment
User activities
Prediction of user features
Suggestions to the user
Personalization of the presentation

5 Conclusions

Concepts

⋆ Based on the data being analyzed:

  – Web Content Mining

  – Web Usage Mining / Impersonalized

  – Web Usage Mining / Personalized

⋆ Based on the goals of the analysis:

  Web Content Mining

  ⟶ Web Usage Mining / Impersonalized ⟶

  ⟶ Web Usage Mining / Personalized

We have

Data

Quantity: very high

Quality: low to very low

Data preparation mechanisms that produce acceptable results if

the site topology is known!

some domain expertise is available

Data mining environments combining:

data preparation mechanisms

data warehousing tools

general-purpose and dedicated miners

mechanisms for exploiting human expertise

We still need

* Precise problem specifications

  Mapping the conventional mining problems into the web context?

  Adjusting quality measures from marketing and other domains of intensive mining to mine the usage of a web site?

* Standardization of input data — cf. Efforts by W3C

* Models alleviating the biases in the data

* Treatment of evolution of
  · the web site
  · the users' behaviour

* Cooperation with the HCI and User Modelling communities

and probably web-oriented miners (?) .

References

[1] Gediminas Adomavicius and Alexander Tuzhilin. Discovery of actionable patterns in databases: The action hierarchy approach. In *KDD*, pages 111–114, Newport Beach, CA, Aug. 1997.

[2] Pierre Berthon, Leyland F. Pitt, and Richard T. Watson. The world wide web as an advertising medium. *Journal of Advertising Research*, 36(1):43–54, 1996.

[3] A. G. Büchner, M. Baumgarten, S. S. Anand, M. D. Mulvenna, and J. G. Hughes. Navigation pattern discovery from internet data. In *WEBKDD'99*, San Diego, CA, Aug. 1999.

[4] Alex G. Büchner and Maurice D. Mulvenna. Discovering internet marketing intelligence through online analytical web usage mining. *ACM SIGMOD RECORD*, pages 54–61, Dec. 1998.

[5] Soumen Chakrabarti, Byron Dom, and Piotr Indyk. Enhanced hypertext categorization using hyperlinks. In *SIGMOD'98*, pages 307–318, Seattle, WA, Jun. 1998.

[6] Soumen Chakrabarti, Sunita Sarawagi, and Byron Dom. Mining surprising patterns using temporal description length. In Ashish Gupta, Oded Shmueli, and Jennifer Widom, editors, *VLDB'98*, pages 606–617, New York City, NY, Aug. 1998. Morgan Kaufmann.

[7] Robert Cooley, Bamshad Mobasher, and Jaidep Srivastava. Data preparation for mining world wide web browsing patterns. *Journal of Knowledge and Information Systems*, 1(1), 1999.

[8] Alex Freitas. On objective measures of rule surprisingness. In *PKDD'98*, number 1510 in LNAI, pages 1–9, Nantes, France, Sep. 1998. Springer-Verlag.

[9] Roy Goldman and Jennifer Widom. DataGuides: Enabling query formulation and optimization in semistructured databases. In *VLDB'97*, pages 436–445, Athens, Greece, Aug. 1997.

[10] David Martin. IBM SurfAid project: Transactive analysis and prediction. Invited Talk in WEBKDD'99, Aug. 1999. see also http://surfaid.dfw.ibm.com/.

[11] Brij Masand and Myra Spiliopoulou, editors. *KDD'99 Workshop on Web Usage Analysis and User Profiling WEBKDD'99*, 1999. online archive at http://www.acm.org/sigkdd/proceedings/webkdd99/.

[12] Balaji Padmanabhan and Alexander Tuzhilin. A belief-driven method for discovering unexpected patterns. In *KDD'98*, pages 94–100, New York City, NY, Aug. 1998.

[13] Georgios Paliouras, Vangelis Karkaletsis, Christos Papatheodorou, and Constantine Spyropoulos. Exploiting learning techniques for the acquisition of user stereotypes and communities. In *7th Intl. Conf. on User Modelling, UM'99*, pages 169–178, Banff, Canada, June 1999.

[14] Mike Perkowitz and Oren Etzioni. Adaptive web pages: Automatically synthesizing web pages. In *Proc. of AAAI/IAAI'98*, pages 727–732, 1998.

[15] Myra Spiliopoulou. The laborious way from data mining to web mining. *Int. Journal of Comp. Sys., Sci. & Eng., Special Issue on "Semantics of the Web"*, Mar. 1999.

[16] Myra Spiliopoulou and Lukas C. Faulstich. WUM: A Tool for Web Utilization Analysis. In *extended version of Proc. EDBT Workshop WebDB'98*, LNCS 1590. Springer Verlag, 1999.

[17] Myra Spiliopoulou, Carsten Pohle, and Lukas C. Faulstich. Improving the effectiveness of a web site with web usage mining. In *Proc. of KDD Workshop WEBKDD'99*, San Diego, CA, Aug. 1999.

[18] Ke Wang and Huiqing Liu. Discovering typical structures of documents: A road map approach. In *SIGIR'98*. ACM Press, Aug. 1998.

[19] Osmar Zaïane, Man Xin, and Jiawei Han. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In *Advances in Digital Libraries*, pages 19–29, Santa Barbara, CA, Apr. 1998.

[20] Liliana Ardissono and Anna Goy. Tailoring the interaction with users in electronic shops. In *7th Intl. Conf. on User Modelling, UM'99*, pages 35–44, Banff, Canada, June 1999.