Discovering Interesting Usage Patterns in Web-based Learning Environments

Karin Becker, Mariângela Vanzin
[kbecker ,mvanzin]@inf.pucrs.br

Pontifícia Universidade Católica do Rio Grande do Sul – PUCRS
Porto Alegre - Brazil

PUCRS

Agenda

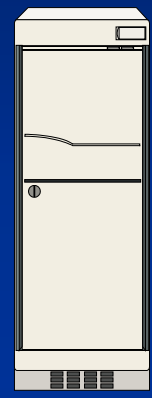
- Web Mining;
  - Web Mining Phases
- Case Study – PUCRS VIRTUAL;
- How to discovery interesting patterns;
- Perspectives

Web Applications

URI Requirements

HTTP

HTML page



Client
Browser
Script

Server
CGI, ASP, PHP API

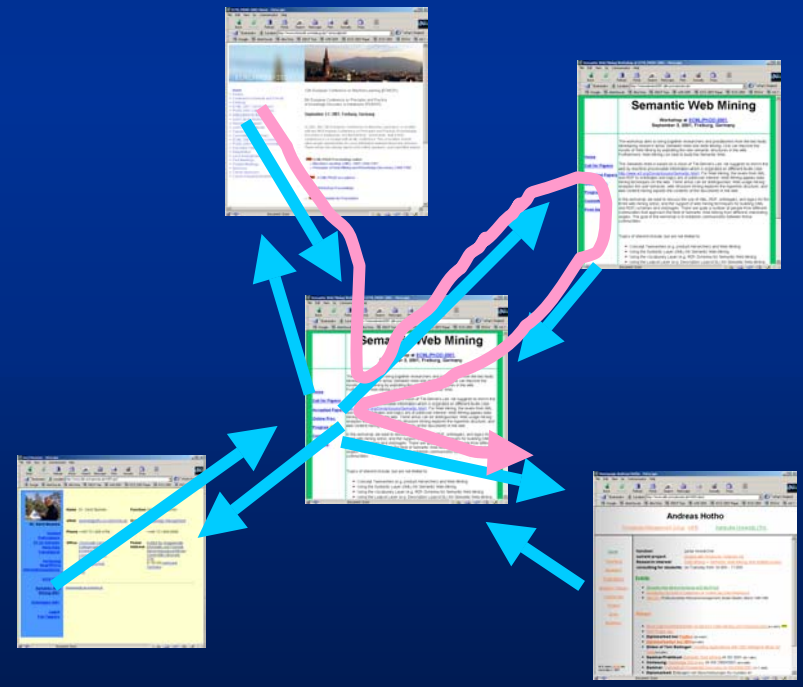
Application

Web Mining

- The use of data mining techniques to extract information from web documents and services
- Problems that can be addressed:
  - Finding relevant information
  - Creating new knowledge out of the information available on the web
  - Personalization of the information
  - Learning about consumers and individual users

Web Mining

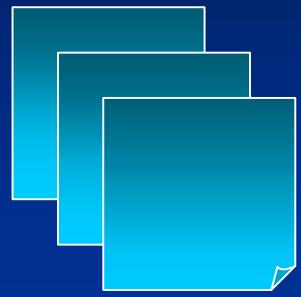
- **Web Mining Areas:**
  - **Content Mining;**
  - **Structure Mining;**
  - **Web Usage Mining**



Web Usage Mining (WUM)

- **Goal**
  - **To identify users' navigation patterns;**
- **Benefits**
  - **know costumers preferences;**
  - **determine cross marketing strategies;**
  - **restructure a Web site in order to better fit its users' needs;**
  - **site personalization;**
  - **performance improvement;**
  - **etc**

Web Usage Mining Phases

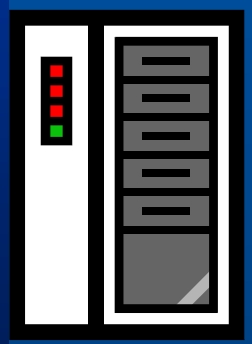


Content and Structure Data

Pre-processing

Pattern Discovery

Pattern Analysis



Raw Usage Data

Preprocessed Clickstream Data

Patterns

"Interesting" Patterns = Knowledge

Web Server Log

IP Address | User ID | Date/time | URL | Status | Bytes

Web Log

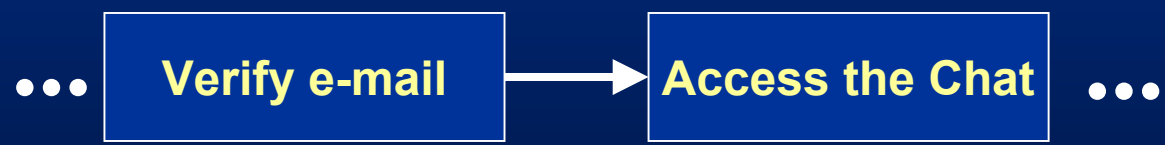
```
200.176.25.110 - aluno1 [10/Jan/2002:00:00:06 -0200] "GET /ESP_SE_01130/competencia/07_01/conselhos.doc HTTP/1.1" 200 31744
200.176.8.249  - -       [10/Jan/2002:00:10:17 -0200] "GET / HTTP/1.1" 200 189
200.176.8.249  - -       [10/Jan/2002:00:10:18 -0200] "GET /webct/public/home.pl HTTP/1.1" 200 1977
200.176.8.249  - aluno2  [10/Jan/2002:00:10:39 -0200] "GET /webct/homearea/homearea HTTP/1.1" 200 20032
200.176.8.249  - -       [10/Jan/2002:00:11:20 -0200] "GET /SCRIPT/Curso_DEF_07JAN/scripts/student/dropbox_stud_home pl HTTP/1.1" 401 899
200.248.5.164  - -       [10/Jan/2002:00:11:21 -0200] "GET /webct/homearea/homearea HTTP/1.1" 401 866
200.176.8.249  - aluno2  [10/Jan/2002:00:11:25 -0200] "GET /SCRIPT/Curso_DEF_07JAN/scripts/student/dropbox_stud_home.pl HTTP/1.1" 200 28552
200.176.8.249  - aluno2  [10/Jan/2002:00:11:26 -0200] "GET /SCRIPT/Curso_DEF_07JAN/scripts/student/serve_layout.pl?LOGO HTTP/1.1" 200 52
200.176.8.249  - aluno2  [10/Jan/2002:00:11:30 -0200] "GET /SCRIPT/Curso_DEF_07JAN/scripts/student/dropbox_stud_home.pl?START+++ HTTP/1.1" 200 8817
200.248.5.164  - aluno3  [10/Jan/2002:00:11:32 -0200] "GET /webct/homearea/homearea HTTP/1.1" 200 12498
200.248.5.164  - -       [10/Jan/2002:00:11:58 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/serve_home HTTP/1.1" 401 881
200.248.5.164  - aluno3  [10/Jan/2002:00:12:02 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/serve_home HTTP/1.1" 200 20172
200.248.5.164  - aluno3  [10/Jan/2002:00:12:05 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_home?_homepage+START HTTP/1.1" 200 4105
200.248.5.164  - aluno3  [10/Jan/2002:00:12:14 -0200] "GET /Curso_ABC_02JAN/AmbienteCurso.pdf HTTP/1.1" 200 16485
200.248.5.164  - aluno3  [10/Jan/2002:00:13:41 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_home?1010410417+view HTTP/1.1" 200 6340
200.176.25.110 - aluno1  [10/Jan/2002:00:13:42 -0200] "GET /ESP_SE_01130/competencia/07_01/paulo_freire_texto.pdf HTTP/1.1" 200 41563
200.248.5.164  - aluno3  [10/Jan/2002:00:13:49 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_home?1010412547+view HTTP/1.1" 200 7930
200.176.20.28  - aluno2  [10/Jan/2002:00:15:15 -0200] "GET /SCRIPT/Curso_DEF_07JAN/scripts/student/dropbox_view.pl?START+++1010521868 HTTP/1.1" 200 2148
200.176.25.110 - aluno1  [10/Jan/2002:00:16:38 -0200] "GET /ESP_SE_01130/competencia/07_01/paulo_freire_texto.pdf HTTP/1.1" 304 -
200.176.25.110 - aluno1  [10/Jan/2002:00:17:16 -0200] "GET /ESP_SE_01130/competencia/07_01/reuniões_prod.doc HTTP/1.1" 200 31744
200.248.5.164  - aluno3  [10/Jan/2002:00:17:39 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_home?1010431964+view HTTP/1.1" 200 8736
200.248.5.164  - aluno3  [10/Jan/2002:00:18:41 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_home?1010422540+view HTTP/1.1" 200 8498
200.248.5.164  - aluno3  [10/Jan/2002:00:18:51 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_mail?START HTTP/1.1" 200 5938
200.248.5.164  - -       [10/Jan/2002:00:18:52 -0200] "GET /web-ct/style/stylesul.txt HTTP/1.1" 200 2419
200.248.5.164  - aluno3  [10/Jan/2002:00:19:01 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_chat.pl?START+1010422540 HTTP/1.1" 200 570
200.248.5.164  - aluno3  [10/Jan/2002:00:19:03 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_chat.pl?CLIENT+1010422540 HTTP/1.1" 200 831
200.248.5.164  - aluno3  [10/Jan/2002:00:19:03 -0200] "GET /SCRIPT/Curso_ABC_02JAN/scripts/student/serve_chat.pl?SERVER+1010422540 HTTP/1.1" 200 714
200.248.5.164  - -       [10/Jan/2002:00:19:11 -0200] "GET /web-ct/code/Client.class HTTP/1.1" 200 6068
```

Web Server Log

- **From:**

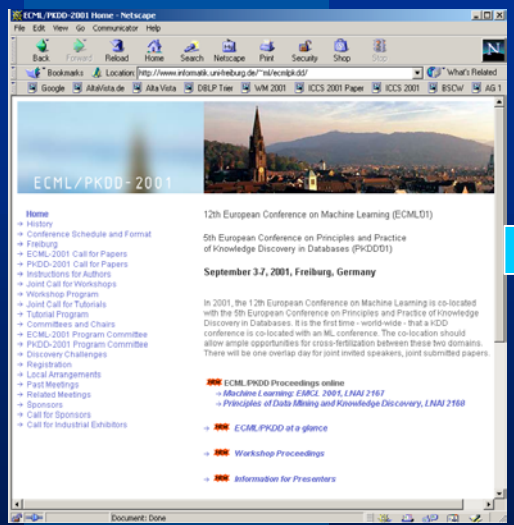
```
"GET/SCRIPT/Curso_ABC_02_Jan/scripts/student/serve_mail
?START

"GET/SCRIPT/Curso_ABC_02_Jan/scripts/student/serve_chat
.pl?START+1010422540
```

- **To:**



Pre-processing Phase

- laborious and difficult step;
- Influences the whole process
- Page view vs Log record;




```
200.248.5.164 - aluno [10/jan/20002:00:15:51 - 0200]
"GET/SCRIPT/Curso_ABC_02_Jan/scripts/student/Home_page" 200 500

200.248.5.164 - aluno [10/jan/20002:00:15:52 - 0200]
"GET/SCRIPT/Curso_ABC_02_Jan/scripts/student/city.jpg" 200 7938

200.248.5.164 - aluno [10/jan/20002:00:15:52 - 0200]
"GET/SCRIPT/Curso_ABC_02_Jan/scripts/student/sky.jpg" 200 3568

200.248.5.164 - aluno [10/jan/20002:00:15:52 - 0200]
"GET/SCRIPT/Curso_ABC_02_Jan/scripts/student/video.mpg"
```

Pre-processing Phase

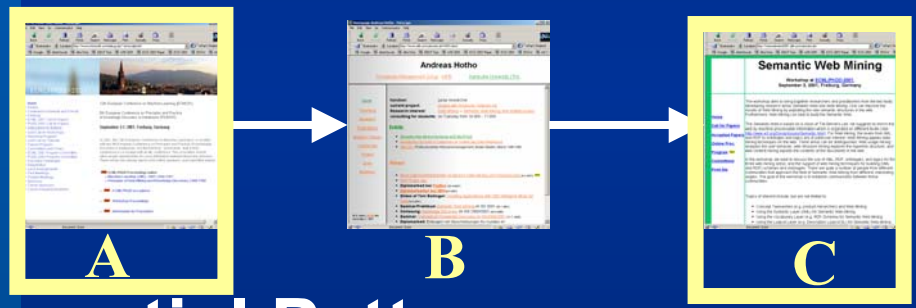
- **Includes:**
  - **Data Cleaning;**
  - **User Identification;**
  - **Session Identification;**
  - **Transaction Identification;**
  - **Path completion.**

Pattern Discovery Phase

- **Traditional mining techniques:**
  - **Association;**
  - **Sequence;**
  - **Classification;**
  - **Clustering;**

Pattern Discovery Phase

- **Association Patterns**
  - Relate pages that most often are referenced together in a user session.

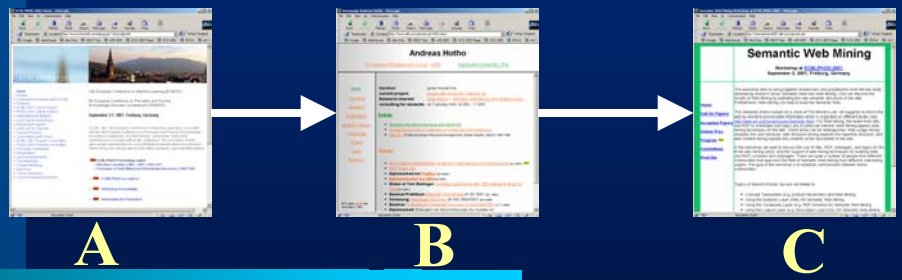


Patterns:

C → A

A → C

- **Sequential Patterns**
  - Describe related accesses in a specific order.



Patterns:

A → B → C

A → C

Pattern Analysis Phase

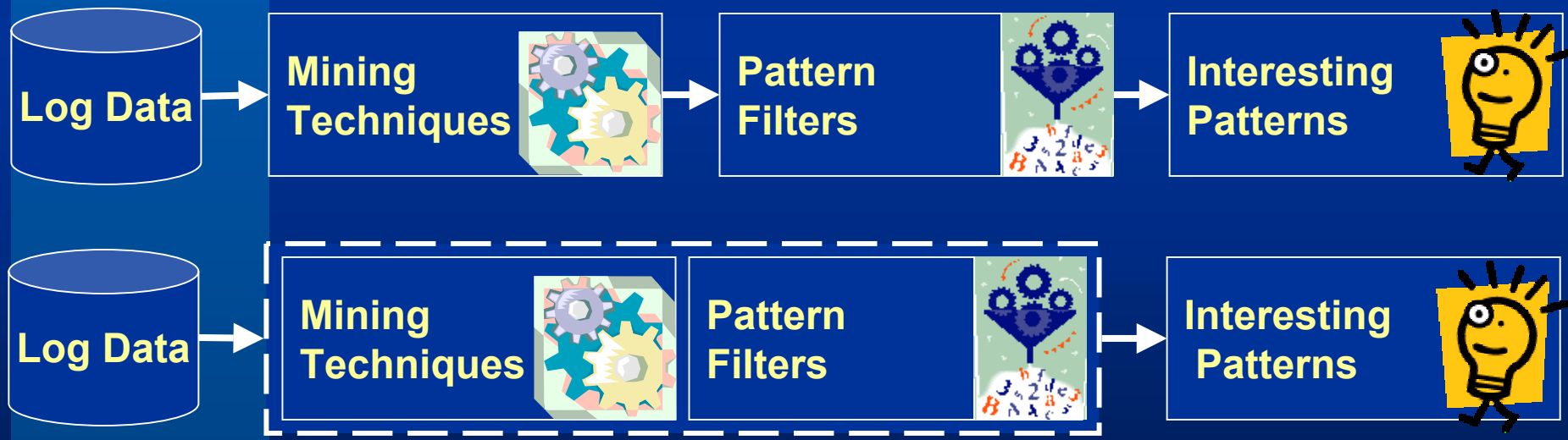
- Pattern + aggregated value = knowledge
- Aggregated value (”interesting”)
  - New
  - Understandable
  - Useful
  - Valid

Pattern Analysis Phase

- **It is difficult for the domain expert to identify interesting patterns:**
  - **The huge number of patterns discovered;**
  - **“interesting” is subjective;**
  - **patterns may not reveal the semantics from a user’s point of view (URL’s Pattern vs. Events patterns).**

Pattern Analysis Phase

- **Once interestingness is defined, it can be incorporated into WUM Process.**



How to Identify Interesting Patterns?

- Using objective or subjective measures;
- Defining patterns to filter or to generate pattern;
- Using domain beliefs;
- Requiring the user intervention to define if a rule is interesting or not.
- Using visualization functionality.

Web-based Learning Environments

- **Web-based learning environments are designed as a set of pages that constitute the educational site.**
- **Course management infrastructures (e.g. WebCT, Virtual-U, Ava, Ariane).**
- **Students with different learning styles access pages differently to reach the educational goals settled by the teacher.**

Evaluation of Web-based Learning Environments

- **Course management infrastructures present limited functionality for monitoring students behavior;**
- **Applying WUM to Web-based learning environments, it is possible:**
  - **To improve the evaluation of course content;**
  - **To identify how the students interact with the learning site.**
  - **To evaluate the effectiveness of site design for the learning process;**

E-learning vs. E-commerce

- **E-commerce:**
  - **E-commerce sites can be measured according to objective goals;**
  - **Well-structured processes;**
  - **Site usage and navigation can be oriented;**
- **E-learning:**
  - **Objectives and site effectiveness cannot be easily defined, nor measured;**
  - **Students interact with the site according to their own learning style;**
  - **Distinct students may reach a same learning goal distinctly;**

Case Study – PUCRS Virtual

- Distance education department of PUCRS
- Hybrid platform
  - Satellite
  - Internet
  - http://www.ead.pucrs.br/

Course Management Infrastructure

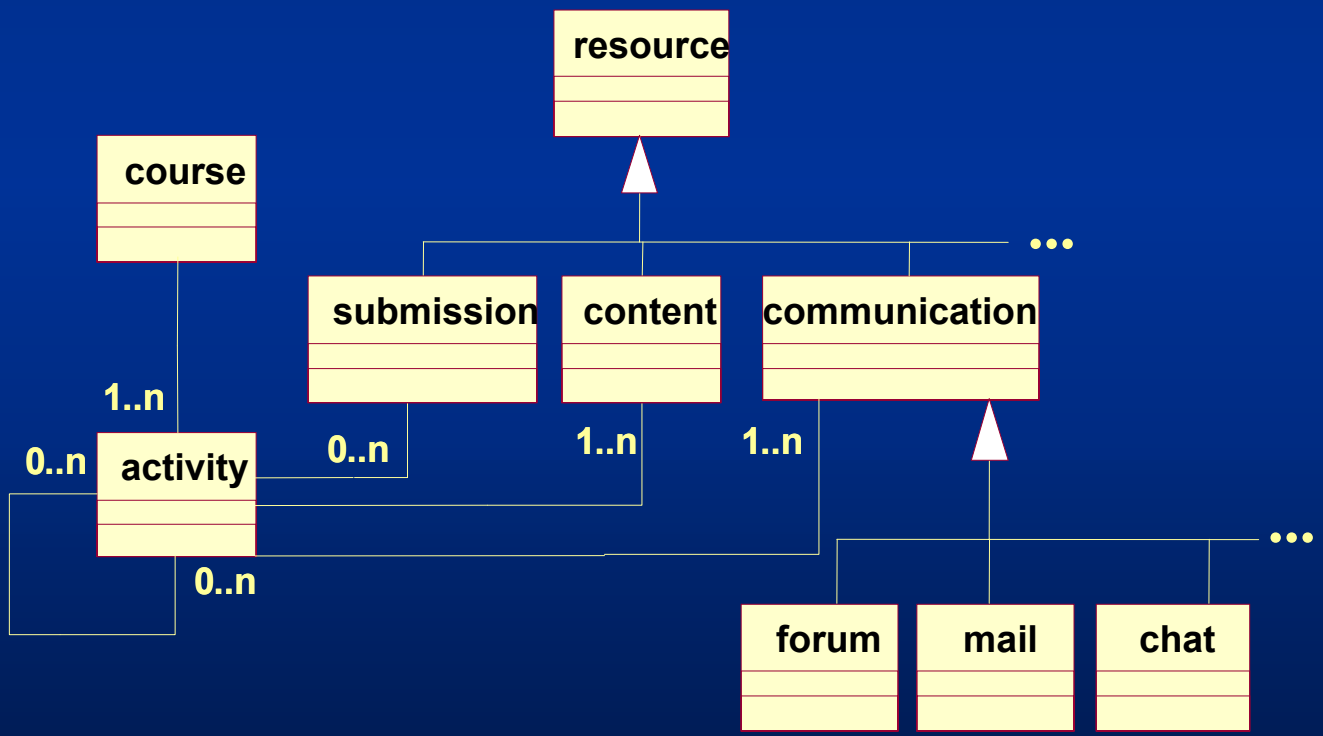
- **WebCT is an infrastructure that helps the development and management of web-based courses.**
- **Learning Resources for Students:**
  - **Content**
    - **Texts, images, sounds, etc.**
  - **Technological**
    - **Chat, e-mail, forum, assessment, activity submission, calendar, etc.**
- **Monitoring Resources for Instructors**
  - **Statistics**

Course Characteristics

- Intensive extracurricular course
- 15 students
- 11 days
- Web server access log: 15.953 access records.

Course characteristics

- View of the Web site design.



WUM Process

Course

Resource

Learning Activities

Web Server

Taxonomy

Activity

Activity Session

Web Log

Click Stream

Pre-Processing

Templates

Mining

Patterns

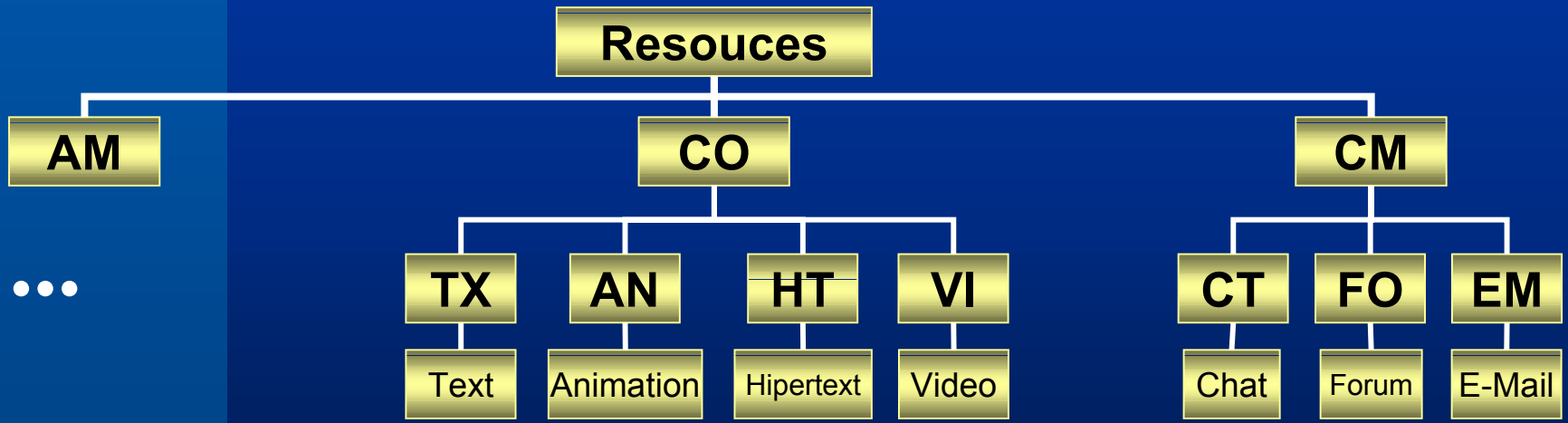
Pattern Discovery

Pattern Analisys

Interesting Patterns

Taxonomy

- Hierarchy of concepts.
- The taxonomy represents the mapping of the URL's into the resources available for execution of learning activities.

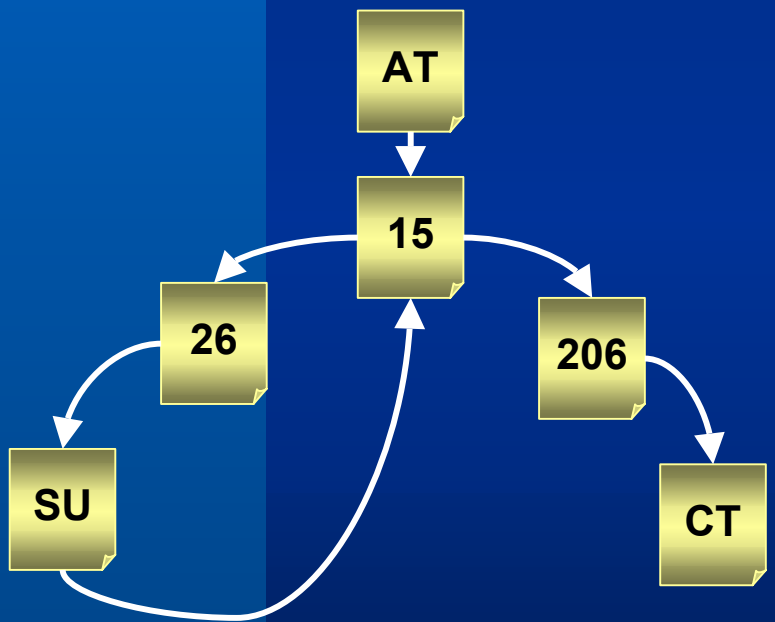


Mining techniques

- **Association**
  - **Combined use of resources**
- **Sequence**
  - **Navigation on the site for accessing resources**

Results (Example)

- Interesting Patterns



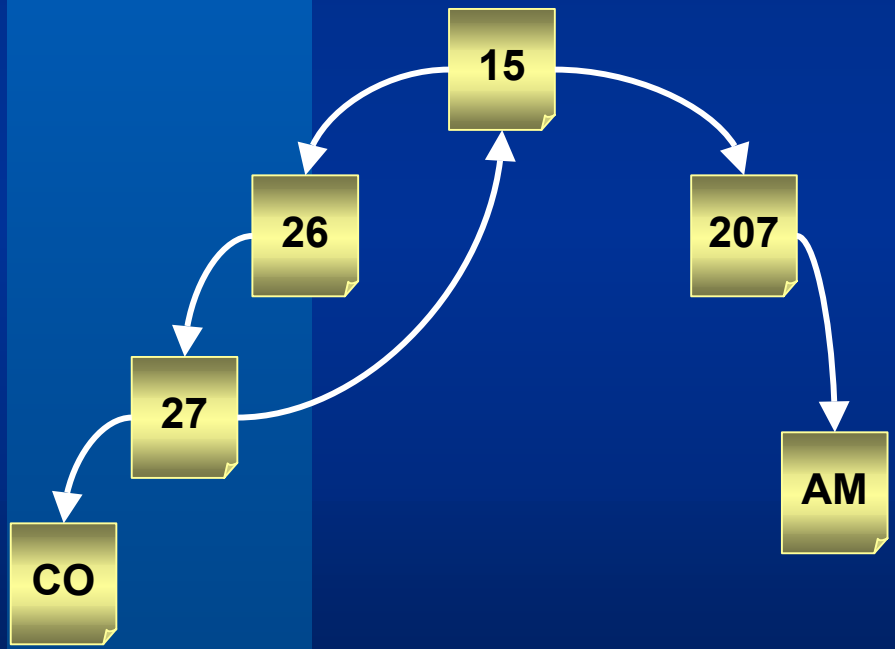
It represents a sequence of accesses that includes the submission of an assignment (class SU) followed by the use of the chat (class CT)

Submission functionality is too complicated for most students (they use the chat to seek help)

Rule Support = 87,8%

Results (Example)

- Interesting Patterns



It shows a sequence of accesses involving Content Material (CO) and Additional Material pages (AM). Additional learning material involves glossary, tutorials, virtual library, etc

Direct access to auxiliary material avoids disorientation

Rule Support = 64,7%

Discussion

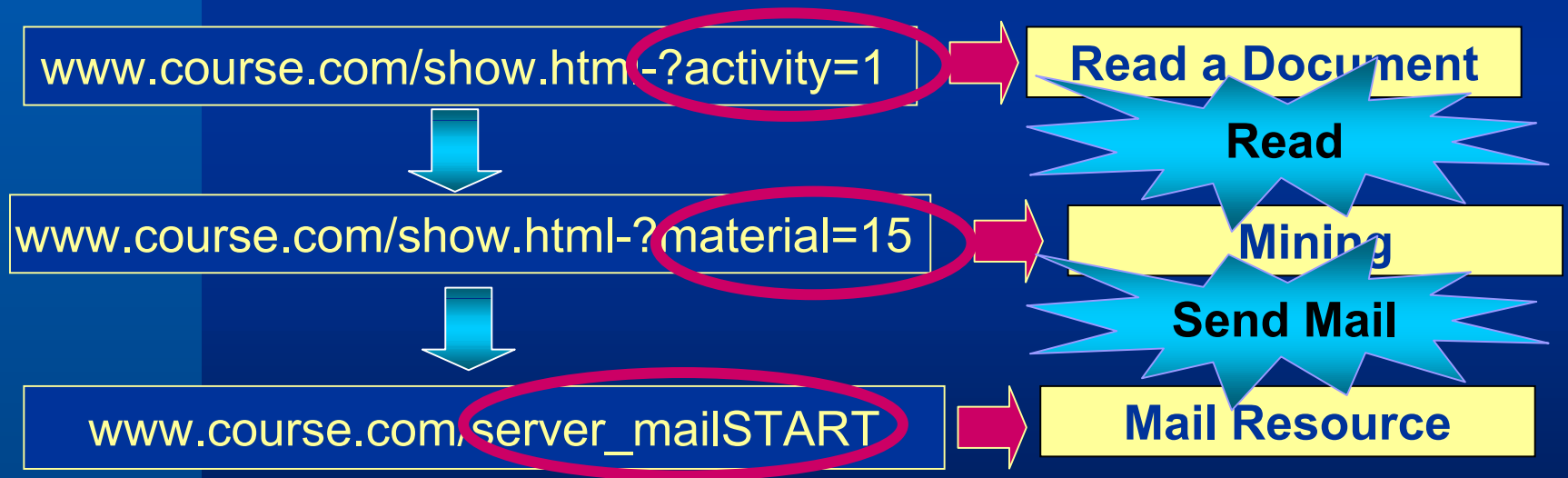
- WUM allows analyzing student navigation behavior in the context of a learning process;
- WUM allows understanding resource and site usage patterns;
- Finding among the patterns the ones that are interesting patterns is hard.

Discussion

- There is a semantic gap between the events that users perform and the events recorded in the web log;
- The patterns are formed by a set of URL's;
- There are URL's syntactically different, but similar in essence (e.g. sending an mail);
- The structure and content of a site is important :
  - input to pre-processing algorithms;
  - Filtering patterns in the Analysis Phase.

Discussion

- **Application experts are interested in application domain *events*.**



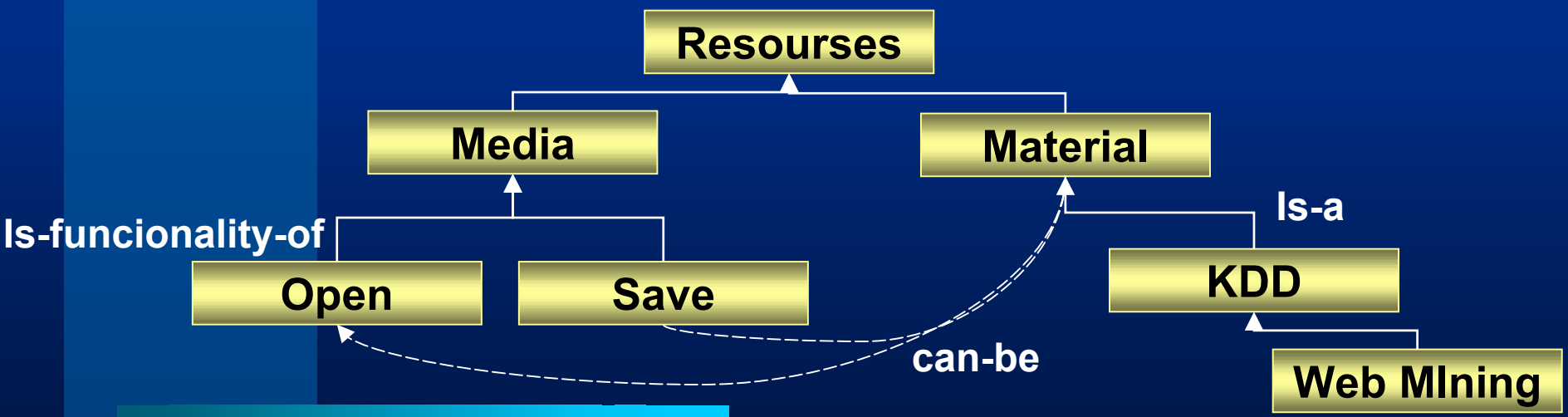
Knowledge Representation

- **Taxonomy**
  - **Primitive form of knowledge representation;**
  - **Hierarchy of concepts**
  - **Srikant et al., Klemetinen et al.**
- **Ontology**
  - **Ontology is defined as a formal, explicit specification of a shared conceptualization [Gruber 93].**

Mapping URL´s to Ontologies

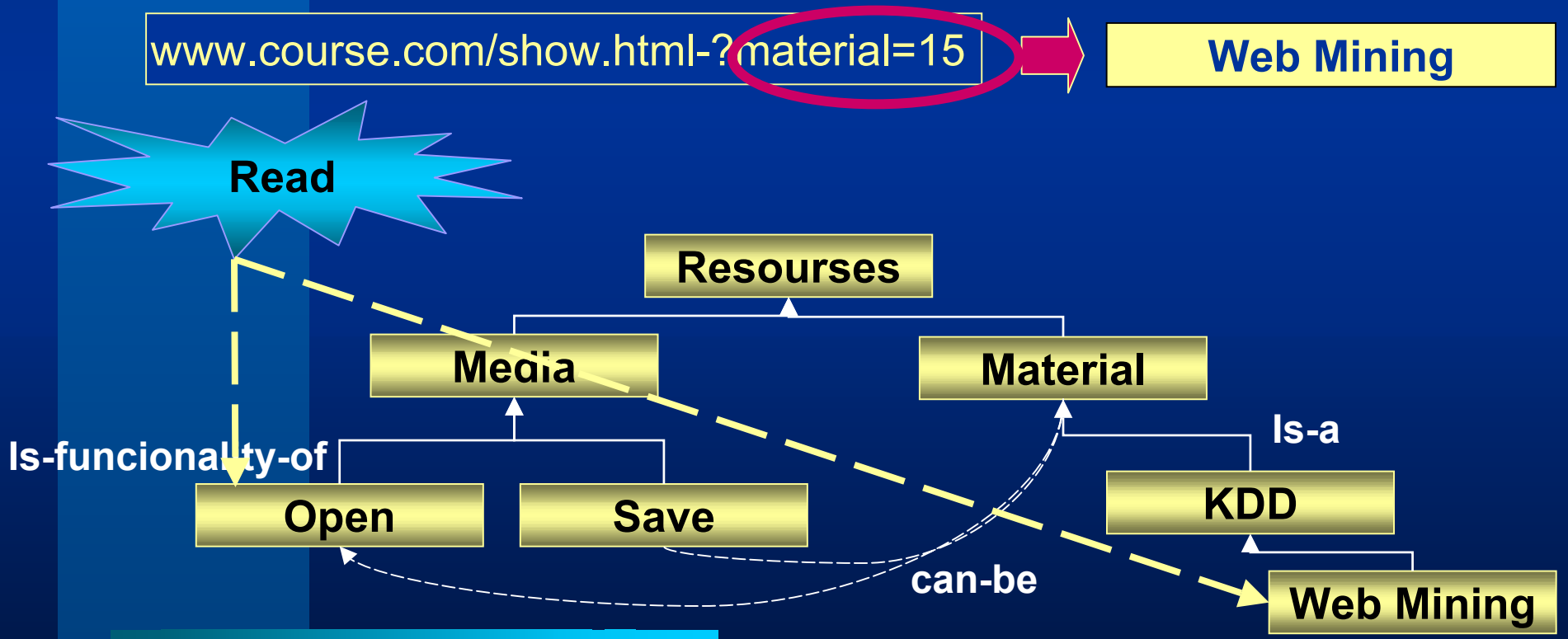
- **Mapping URL's to ontology concepts;**
- **Atomic Events:**
  - **Content**

www.course.com/show.html-?material=15 → **Web Mining**



Mapping URL´s to Ontologies

- **Atomic Events:**
  - **Services**



Mapping URL´s to Ontologies

- **Complex Events:**
  - **Sequences of events;**
  - **Problem-solving strategy, a canonical activity sequence pertaining to the domain, or a description of a behavior patterns observed by exploratory analysis.**

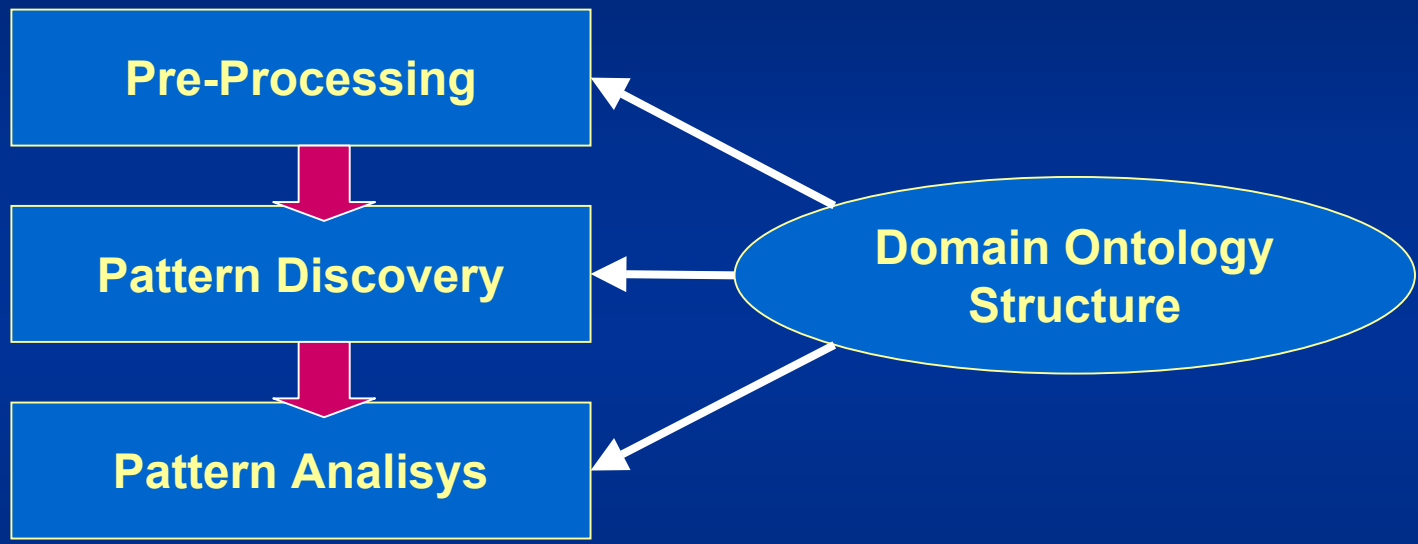
Applying Ontologies to WUM

- **Daí et al.**
  - Pre-processing Phase;
  - Page information are extracted using information retrieval techniques, a technique useful for adding semantics about page content or structure;
- **Berendt et al.**
  - Pattern Analysis Phase;
  - Concepts are extracted from a combination of scripts URI and database contents, based on site functionality interpretation.
- **Oberle et al.**
  - Pre-processing and Pattern Analysis Phase;
  - The log are mapped to ontological concepts, called as Semantic log.
  - The multitude of user interests can be captured - conceptual user tracking.

Different Usage Patterns

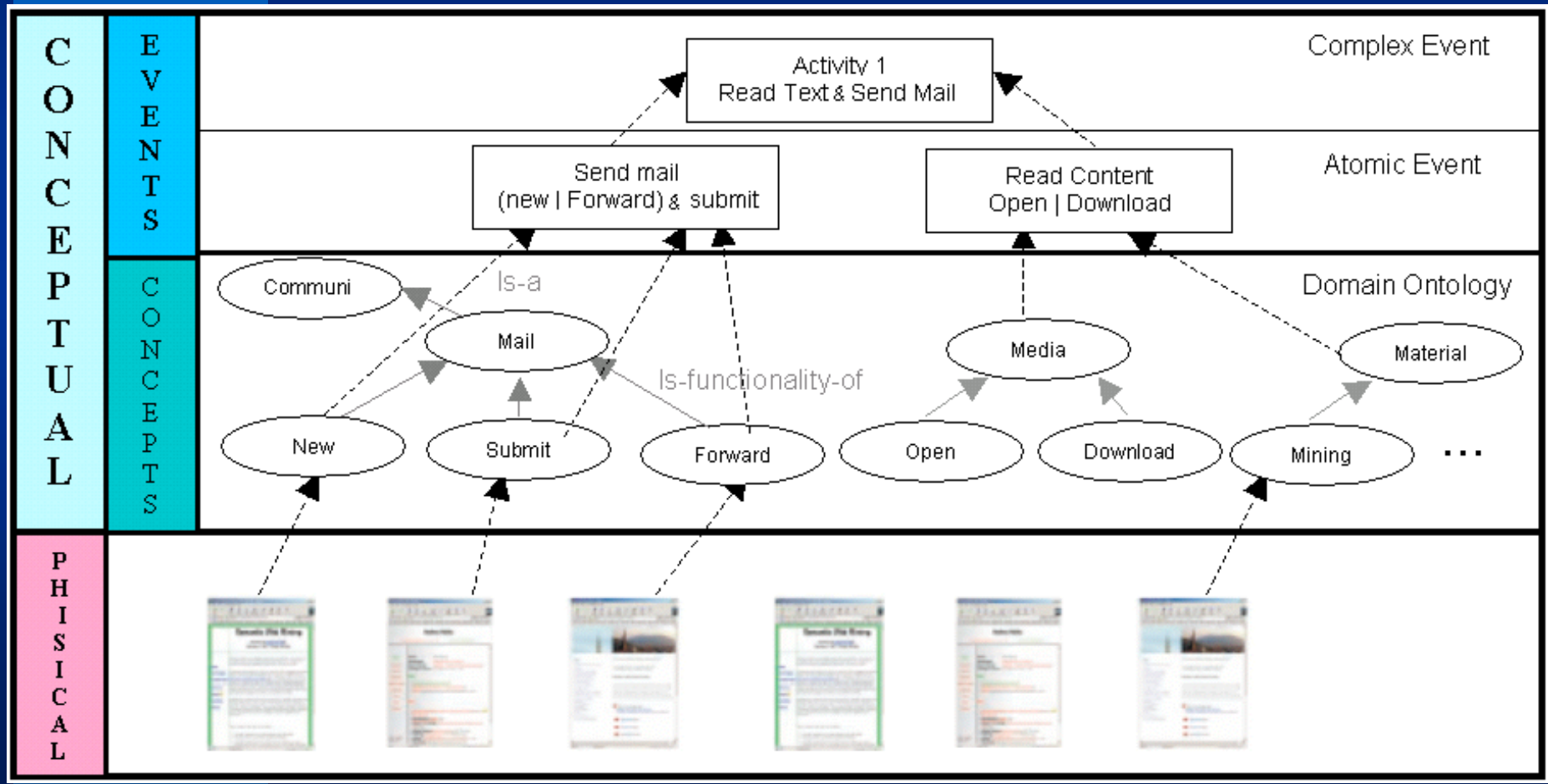
- **Structure: patterns that reveal how the structure of the site is navigated through;**
- **Content: patterns that reveal the subjects that are addressed during the navigation;**
- **Resource usage: patterns that reveal which resources are employed, and how these are employed.**

Perspectives on the Use of Ontologies



- **Benefits for Web Based Learning:**
  - **Pattern Interpretation;**
  - **Pattern Filtering;**
  - **Pattern Navigation – level of representation**

Perspectives - Domain Ontology Structure



Current Work

- **Preprocessing tool**
  - **Configuration of operators**
  - **Extensible**
- **Navigation mining algorithm and visualization**
  - **Extension of Spiloupoulou et al.**
- **Exploring Semantics**
  - **Pattern analysis**
  - **preprocessing**

Conclusions

- WUM helps understanding site usage and learning process
- Limited claims of WUM success
- Semantic gap between URL and site events
- Explicit domain knowledge
  - Multilevel
  - ontology
  - Preprocessing, interpreting, mining
- Semantic web
  - Semantic log?

Acknowledgments

PUCRS

This work is partially supported by DELL/PUCRS agreement and by Fundação de Amparo à Pesquisa do Rio Grande do Sul (FAPERGS- Brazil)