Ontologies in Web Personalization

Magdalini Eirinaki, Athens University of Economics and Business
Michalis Vazirgiannis, Athens University of Economics and Business
Ontologies in Web Personalization

Magdalini Eirinaki, Michalis Vazirgiannis

DB-NET research group, Dept of Informatics, Athens University of Economics and Business
WWW: http://www.db-net.aueb.gr
Talk outline

- The Web personalization process
- Semantic Web Personalization
- Mapping web content to ontologies
- Future Directions & Challenges
Talk outline

- The Web personalization process
- Semantic Web Personalization
- Mapping web content to ontologies
- Future Directions & Challenges
Introduction

- Continuous growth in the size and the use of WWW
- Large and complicated web structures
- Huge financial interests in attracting customers, leading them to relevant content

Objectives:
- Increase # of visits on a server to increase ads expenditure
- Convincing users to buy products and services

Need for predicting user needs in order to improve the usability and user retention of a web site

Web Mining for Personalization
Web Personalization - Definition

“any action that adapts information or services provided by a Web site to the needs of a user/set of users, taking advantage of the knowledge gained from the user’s navigational behavior”

Web Personalization – Impact (I)

Revenues

- global investment in personalization technologies will grow from $500m in 2001 to $2.1bn in 2006

- North America: 67% of personalization revenues (lead in eCommerce and CRM technologies)

- Europe: 25% ($131m) of global investment

[Datamonitor]
Web Personalization – Impact (II)

Research

- More than 10 workshops on Web personalization
  - AAAI SWP, IJCAI ITWP, etc.

- Many relevant workshops on web usage mining/web information management
  - ACM WEBKDD, PKDD EWMF, ACM WIDM, etc.

- Related conferences
  - KDD, PKDD, WWW, ICDM, etc.
Web Mining
The three cornerstones

WEB MINING

**USAGE**
- usage of web pages - stored in the web logs

**CONTENT**
- data presented to the end user - the web site's content

**STRUCTURE**
- the way content is organized - the web site's hyperlinks
Web Mining

- **Web usage mining (WUM):** The application of web log data analysis techniques to model users’ navigational behaviour.

  Use models for:
  - Decision making
  - Web personalization/recommendations

- **Integration of web mining techniques:** Combine web usage mining with content/structure mining techniques/data
Web Personalization

1. Data Collection
2. Data Preprocessing
3. Data Analysis - Pattern Discovery
4. Pattern Analysis
5. Personalization
Talk outline

- The Web personalization process
- Semantic Web Personalization
- Mapping web content to ontologies
- Future Directions & Challenges
Motivation (I)

- Apply web mining methods to Web log data (usage data)
- Use discovered (behavioral) patterns to personalize

www.sportal.com/events/ski.html,

www.sportal.com/travel/ski_resorts.html

www.sportal.com/equipment/ski_boots.html
Motivation (II)

PROBLEMS:

- Not enough hits in the usage log
- URI content updates

NEW URI:

- www.sportal.com/equipment/ski_boot_offers.html

- URIs with semantically relevant content ignored

www.sportal.com/weather/snowreport.html
Motivation (III)

What about visits to semantically similar pages?

www.sportal.com/sports/winter_sports/ski.html,
www.sportal.com/travel/winter/hotels.html

?
Semantic Web Mining (SWM)

- Fast-emerging research area
- How can Semantic Web and Web Mining benefit from each other? [BHS02]
  - Connective link: Ontologies
  - Apply web mining techniques to semi-automatically create ontologies for building Semantic Web sites
  - Exploit semantic structures such as ontologies to improve the results of Web Mining

Semantic Web Personalization (SWP)
SWM/SWP Approaches (I)

- “Service based” concept hierarchies: Analyze search behavior of users. Also use for grouping web pages [BS00, B02]

- Semi-Markov process over concept hierarchy of web site’s topics [AG03]
SWM/SWP Approaches (II)

- Web Personalizer [DM02]
- SEAL [OB+03]
- SEWeP [EVV03, EL+04]
Semantic Web Mining & Personalization Systems:

Web Personalizer

Web Personalizer

- Create usage/content profiles
  - Weighted item vectors
- Abstract items mapping them to ontology terms
- Focus on ontology creation
Current User Model at Domain-Level

Instantiation

Recommendations of Items

Approach 1: Personalization based on current user model at domain-level
Approach 2: Personalization based on domain-level usage profile
Semantic Web Mining & Personalization Systems:

SEAL

SEAL Framework

- SEmantic Web PortAL
- Use:
  - Ontology (portal’s backbone)
  - RDF/RDF Schema
  - Web pages & content metadata
  - Semantic Log File
Application: KA2 Portal
Concept extraction

- Ontology underlying semantic web portal
- Static pages: Inherent RDF annotations
  - Use to map URLs to ontology terms
- Dynamic pages: Analyze query strings
Semantic Log File

Log File & feature vector:

<table>
<thead>
<tr>
<th>UserID</th>
<th>Time Int.</th>
<th>Person</th>
<th>Publication</th>
<th>name</th>
<th>author</th>
</tr>
</thead>
<tbody>
<tr>
<td>4711</td>
<td>$t_{4711,1}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4711</td>
<td>$t_{4711,2}$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0815</td>
<td>$t_{0815,1}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$u_{anon}$</td>
<td>$t_{anon}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Ontology concepts:
Semantic Web Mining

- Exploit concept hierarchy (taxonomy part):
  - Use generalizations/specializations to expand feature vector
- Generalize web usage mining queries
- Find underlying associations on mining results
Semantic Web Mining & Personalization Systems:

SEWeP

[EVV03] M. Eirinaki, M. Vazirgiannis, SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process, SIGKDD’03

[EL+04] M. Eirinaki, C. Lampos, S. Paulakis, M. Vazirgiannis, Web Personalization Integrating Content Semantics and Navigational Patterns, ACM WIDM 04

M.Eirinaki, M. Vazirgiannis
Ontologies in Web Personalization
SEWeP System

Expand the recommendation set using Web site’s content semantics

www.sportal.com/accessories/ski_boots.html
www.sportal.com/weather/snowreport.html
www.sportal.com/accessories/ski_boot_offers.html

ski, snow, winter
SEWeP Personalization

- Utilize:
  - Web site’s usage
  - Web page semantics
  - Domain-specific Taxonomy

- Map extracted keywords to taxonomy categories (concepts)

- Create C-Logs (concept logs) & semantic document clusters
Similarity Mapping

Semantic similarity

Similarity measure &

WORDNET

keywords

concept hierarchy
C-Logs Snapshot

- #Fields: c-ip cs-uri-stem Title Categories Keywords

- 213.249.0.155 /projects.htm projects of db-net database data system project research medicine database data management systems novel projects research local teams position medical private

- 213.249.0.155 /people.htm people of db-net student research phd erasmus person professor student research phd associate erasmus people professor
Semantic Document Clustering

- Web pages characterized with taxonomy terms
  
  `/accessories/ski_boots.html`
  `/weather/snowreport.html`
  `/accessories/ski_boot_offers.html`

- Discover document clusters based on similarity between *taxonomy terms*
- Use these clusters to expand recommendation set (with documents from the same cluster)
Category-based Recommendations

- Express user behavior using concepts
- Use *semantic* relationships between concepts as expressed by their topology in taxonomy
- Pattern matching to user’s navigational behavior no longer *exact*

```
cinema, music  => soundtrack
movie, singer  => soundtrack
```
Recommendation Engine

User visit:
/events/ski.html, /travel/ski_resorts.html

Original Recommendations:
=> /accessories/ski_boots.html

Semantic Recommendations (KDD 2003):
=> /weather/snowreport.html
=> /accessories/ski_boot_offers.html

Category-based Recommendations (WIDM 2004):
sports, ski, winter, travel
=> snow, ski accessories
## Semantic Web Mining/Personalization Systems

<table>
<thead>
<tr>
<th></th>
<th>Ontology - Taxonomy</th>
<th>Concept logs</th>
<th>Content semantics</th>
<th>SWM</th>
<th>SWP</th>
<th><strong>x</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Web Personalizer</td>
<td><strong>x</strong></td>
<td></td>
<td></td>
<td>?</td>
<td></td>
<td><strong>x</strong></td>
</tr>
<tr>
<td>SEAL</td>
<td><strong>x</strong></td>
<td><strong>x</strong></td>
<td>Embedded</td>
<td><strong>x</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEWeP</td>
<td><strong>x</strong></td>
<td><strong>x</strong></td>
<td>Automatic extraction</td>
<td><strong>(x)</strong></td>
<td><strong>x</strong></td>
<td></td>
</tr>
</tbody>
</table>
Talk outline

- The Web personalization process
- Semantic Web Personalization
- Mapping web content to ontologies
- Future Directions & Challenges
Mapping the web content to ontologies

- Semantics for web personalization
  - Content semantics + knowledge from usage logs to enhance web personalization
  - Semantics’ extraction process:
    - Feature extraction (content and links)
    - Mapping to ontologies
Mapping to ontologies

- Objective: produce valid semantic recommendations
- Need to classify pages to ontology terms ~ map pages/links features to ontology terms
Feature extraction process

- Content & Links
  - Concise representation of large documents
  - Usage of information preserving methods
    - Tf/idf, MI, x^2
  - Each page/link represented by a small number of features
  - Objective: map them to an ontology
Natural Language Ambiguity

- A word can have various meanings in different contexts.
- i.e. the words bank:
  - “They pulled the canoe up on the bank”
  - “He cashed a check at the bank”
- A more expressive representation can be achieved when the correct senses are known.

Word Sense Disambiguation
Ontology based WSD

- The goal of a WSD is the resolution of the ambiguity in natural language.

- A case of WSD utilizes the senses provided by Machine Readable Dictionaries - MRDs (i.e. Collins, WordNet).
  - Input: List of possible senses for a word provided by the MRD.
  - Output: The correct sense, based on contextual information.
  - Evaluation: Based on pre-annotated corpora (Senseval 2,3, SemCor 1.7.1).
WSD is computationally expensive

- The combinations of possible tuples of senses grows exponentially with the window size.
  - Techniques for reducing the search space should be employed (i.e. Simulated Annealing).

- WordNet particularities.
  - Consists of the 9 distinct hierarchies.
  - When senses are distributed over different hierarchies, semantic distances are not explicit.
Compactness based WSD

Main intuitions:

- Adjacent terms in a document are semantically close.
- Compact sub-hierarchies in the MRD reflect semantic similarity.
- For a group of adjacent terms the corresponding MRD senses should form a compact sub-hierarchy.

Preprocessing phase:
- Part of Speech Tagging (POS) (Stanford Tagger).
- Phrase detection based on WordNet.
- Retain only the Noun Phrases for disambiguation.

Disambiguation phase:
- Noun phrases are divided in non-overlapping windows of predefined size.
- For each window of terms, for all the possible tuples of senses the algorithm selects the tuple that exhibits the highest compactness in the hierarchical thesaurus.
Compactness Measure

- A popular distances between two senses in an Ontology: cost of the path connecting them through their least common ancestor (lca).

- Extension of this notion from a pair to a tuple of senses $S$, and is defined as the cost of the Steiner Tree of $S$ such that the final Tree contains all the pairwise distances of the senses.
  - Note that the restrictions impose the lca($S$) and a path from every sense to the lca($S$) to be included in the final Tree.

- The cost of the final Tree is a measure of semantic similarity for a tuple of senses.
Compactness - The role of the ontology

\{\text{wind, thunder}\} \text{ vs } \{\text{wind, quitar}\}
Compactness measure evaluation

- Conducted experiments with Brown and Senseval corpora
- The experimental results in Classification indicate that high precision (even at the cost of low recall) is a key element for applying WSD in classification
Some Classification Results

- The two largest categories of Reuters-21578 corpus (*acquisitions* and *earnings*).
  - 4,436 training documents
  - 1,779 test documents
- Preprocessing:
  - POS Tagging, Phrase detection based on WordNet.
- $W=3$, $L=0$ parameters used for WSD.
- We varied the training set size between 3 and 500 documents per topic.
- We varied the number of hypernyms used in the definition of the similarity measure.
- We report the differences of the macro-averaged F-Measure between the respective configuration and the baseline, using 10 iterations for Reuters
 Experimental Results

![Graphs showing experimental results](image-url)
Messages

- Semantics increase the value/success of recommendations in web personalization by expanding their semantic scope
- Mapping web content/links features to ontologies is a key feature to successful assignment of semantics to web pages/links
- WSD contributes to more accurate classification (i.e. mapping sets of features to ontologies) especially in the absence of sufficient training data – often the case in usage mining.
Talk outline

- The Web personalization process
- Semantic Web Personalization
- Mapping web content to ontologies
- Future Directions & Challenges
Future Directions & Challenges

- Web grows at a tremendous pace
  - 25% new links and 8% new pages per week [NCO04]
  - 15% of pages weekly updated [FMNW03]
- Web personalization: area with significant industrial and research impact
Future Directions & Challenges

- Integration of web and usage graphs
  - Higher quality recommendations
  - More realistic web page rankings!

- Integration of content semantics & structural knowledge in the personalization process

- Promote link semantics to 1st class citizens in the recommendation process …
Thank you!
References

References

- [EVV03] M. Eirinaki, M. Vazirgiannis, SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process, in Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD’03), Washington DC, August 2003