Semantic Web Mining: Concepts & Applications

Sebastián A. Ríos
srios@dii.uchile.cl
Industrial Engineering Department - Univ. of Chile
Outline

- Motivation
- The Evolution of Web Mining
- Semantic Web Usage Mining
- Real Application
  - Concept-based Web Usage Mining
- Conclusion
1. Motivation
E-Commerce Revolution

Organizations realized of the huge WWW power

- Many organizations create their web sites and begin to use the technology to provide on-line services

- e-commerce begins
  - Web Banking, e-learning, e-retail, etc.
Problems Arose

- MANY WEB SITES PROVIDE SAME SERVICES
- HOW TO IMPROVE SALES
- VISITORS CHANGE SITE WITH MINIMUM EFFORT
- HOW TO REACH MORE PEOPLE

1st step
Keep customers (visitors) interested in the web site

THIS IS EASY TO SAY BUT HARD TO DO
Keep Visitors’ Interest

Many ways exist

- Create new and innovative services
  - You never know what visitors really want

- Provide the visitor the information that needs
  - Information in the sites must be **simple/clear** and **complete**
  - Visitor must find the information **fast**! (if not he/she will change site)

Enhance the site **Contents and Organization**
How to enhance a Site

We can use only the Analyst/Expert Intuition or Criteria

Use Web Mining Techniques to help the Analyst

This way we can help the analyst to discover information that his intuition/criteria alone could not discover
Type of Enhancements

Web Site

Documents

On-line Enhancements

- USUALLY THEY ARE PERFORMED AUTOMATICALLY
- THEY INCREASE THE WEB SITE COMPLEXITY
- RETRIEVAL TIME IS INCREASED

We need to understand the data in order to improve the web site!!

Off-line enhancements

- USUALLY THEY ARE NOT PERFORMED AUTOMATICALLY
- THEY DON’T INCREASE THE WEB SITE COMPLEXITY
- RETRIEVAL TIME IS NOT INCREASED
General Process of WM

Web Site Off-line Enhancements

Web Site Documents

Data Selection

Data Preprocessing

Data Generalization

Analysis/Evaluation

Analyst
The Evolution of Web Mining
What is Web Mining?

Web mining focuses in discover **patterns** in a web site

WE FIND SIMILAR CHARACTERISTICS AMONG OBJECTS (DOCUMENTS, SESSIONS, ETC.)
Main WM Types

We have three main kinds of Web Mining Techniques

- Web Text Mining
- Web Structure Mining
- Web Usage Mining
Web Text Mining (WTM)

We want to find text patterns

- This help to **discover keywords** for web documents labeling

- Keywords are a few words that describe the web page content

- Since 90’s until today search engines use key words to index web pages (yahoo, google, pubgene.org, etc)
Web Structure Mining (WSM)

We find patterns of links on web pages
- This help to improve the organization of the web site by discovering:
  - Neighbor pages
  - Directly linked pages
There is a second type of WSM

This analyzes the inner structure of a web page to find similarities among other documents

THIS IS ALSO CALLED WEB CONTENT MINING (WCM)
Web Usage Mining (WUM)

We analyze visitor’s sessions when browsing a site

- A session is a sequence of web pages

- This way we can find what sequence of pages are most common
  - We can use the information for marketing campaigns, enhance the contents and the structure of the site
In order to enhance the results there were combined Sessions with Contents on Web Pages.
TODAY, A WEB USAGE MINING (WUM) APPROACH IS THE MOST USED TO PERFORM OFF-LINE ENHANCEMENTS...
Semantic Usage Web Mining
The current Web...

- Today, the web is syntactic or structured using HTML.
- Links created a web of documents rich of multimedia contents.
- However, there is no much information for machines to work.
There are some problems...

The main problems of traditional WTM & WUM are:

- Poor and Hard Interpretation of Results
- Time consuming analysis process
- Poor quality of results (patterns found)
- Difficult Evaluation (non standard)

First three are produced by missing the meaning of web documents.
What does this mean?

The pages of a web site are represented as a vector of words with some frequency. Later, this vectors are used by a clustering or classification algorithm in order to find some interesting patterns.

### A Web Page

<table>
<thead>
<tr>
<th>Component</th>
<th>TFIDF Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>School</td>
<td>0.0002604</td>
</tr>
<tr>
<td>Engineering</td>
<td>0.0002234</td>
</tr>
<tr>
<td>University</td>
<td>0.0001225</td>
</tr>
<tr>
<td>Student</td>
<td>0.0000023</td>
</tr>
<tr>
<td>Course</td>
<td>0.0234545</td>
</tr>
<tr>
<td>Lecture</td>
<td>0.0002667</td>
</tr>
<tr>
<td>Department</td>
<td>0.1093356</td>
</tr>
<tr>
<td>Alcohol</td>
<td>0.0000580</td>
</tr>
</tbody>
</table>

**TFIDF**

\[
w_{ij} = tf_{ij} \times \log_2 \left( \frac{N}{n} \right)
\]

Thousands of components (experimentally 5,000 components per web page)
...and the results?

POOR & HARD RESULTS’S INTERPRETATION

TIME CONSUMING RESULTS’S INTERPRETATION

RESULTS ARE MIXED UP (WRONG CLASSIFICATION)

SOM, K-MEANS, ETC.

A Neuron

<table>
<thead>
<tr>
<th>Label</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>School</td>
<td>0.0002604</td>
</tr>
<tr>
<td>Engineering</td>
<td>0.0002234</td>
</tr>
<tr>
<td>University</td>
<td>0.0001225</td>
</tr>
<tr>
<td>Student</td>
<td>0.0000023</td>
</tr>
<tr>
<td>Course</td>
<td>0.0234545</td>
</tr>
<tr>
<td>Lecture</td>
<td>0.0002667</td>
</tr>
<tr>
<td>Department</td>
<td>0.1093356</td>
</tr>
<tr>
<td>Alcohol</td>
<td>0.0000580</td>
</tr>
</tbody>
</table>

Thousands of components (experimentaly 5,000 components in WTM & 15,000 in WUM)
...and WUM Interpretation?

It's also hard to understand visitors' browsing behavior from such information.

Clustered Sessions' Documents

Session of three documents

People with different interests
...Why does this happen?

We obtained:  Similar  Similar  Similar
Should be:    Similar  Dissimilar  Dissimilar

MISSING THE MEANING

WE NEED TO CONSIDER THE MEANING
Semantic WUM

The word “semantic” stands for “the meaning of”

Therefore, the semantic of something is the meaning of something

The idea is to enhance the results by adding more semantics into the web mining process
Semantic Web Usage Mining

Clustered Sessions' Documents

With Semantic WUM it is possible to place together similar sessions with more related contents. Therefore is easier for analyst to understand the results and create off-line web site enhancements.
Different Approaches...

- Ontologies
  - General Purpose Ontologies (Wordnet, EuroWordNet)
  - Ontologies Developed for a particular Site

- Metadata & Concepts
  - Add metadata to web pages
  - Folksonomies - Cooperative Classification and Communication Through Shared Metadata

- Mathemtica Methods
  - Latent Semantic Analysis (used in LSI)

- Sintactic Analysis
A Real Application
Concept-based WUM
WUM Problems

Results were far from what visitors really wanted to find on the web site (survey to 100 students)

It was difficult to improve the sites’ organization or contents using this information

WUM take huge processing times

- Since we have millions of sessions and hundreds of pages
  - Chilean newspaper (LUN.COM) took about 28 days
  - Faculty of Science Physics and Mathematics took 24h (the faster experiment) (near 1 million sessions and 200 pages)
Concept-Based Approach

Using Experts Knowledge to Define Main Concepts or Topics of the Web Site

0. Organization’s Web Site Goals & Contents

1. Design Web Site Contents & Structure

2. Web Site Development

Web Site

- Section 1
  - Sub section 1.1
  - Sub section 1.2
  - Sub section 1.3
    - Sub Sub section 1.3.1
  - Sub section 1.4
- Section 2
- Section 3
- Section 4

Concept 1
Concept 2
Concept 3
...Concept N

Browsing Behavior Vector A

- t1
  - Page 1
- t12
  - Page 12
- t103
  - Page 103

Membership value: Degree that the page expresses a specific concept
Concept-based Framework

Combining a concept-based classification with the traditional WUM approach to perform better off-line enhancements on the web site structure and content.
We want to avoid...

Same color documents are conceptually related. However, with syntactic methods a cluster have sessions with mixed documents. Therefore, it is difficult for the analyst to understand the results to create off-line web site enhancements.

Clustered Sessions' Documents
Results of CWUM

With Concept-based WUM it is possible to place together similar sessions with more related contents. Therefore is easier for analyst to understand the results and create off-line web site enhancements.

Clustered Sessions' Documents
Does it really work??

- **NO ENHANCEMENTS**
- **- MIXED TOPICS DOCUMENTS**

Diagram:

- **Human Intuition**
  - Analyst without help decide Enhancements

- **Traditional Approach**
  - VSM + Sessions \(\rightarrow\) K-Means
  - Analyst

- **Traditional Approach**
  - VSM + Sessions \(\rightarrow\) SOFM
  - Analyst

- **LSISOM Approach**
  - VSM \(\rightarrow\) LSA \(\rightarrow\) SOFM \(\rightarrow\) SOFM + Sessions \(\rightarrow\) SOFM
  - Analyst

- **Concept-based Approach**
  - VSM + Concepts \(\rightarrow\) Fuzzy Reasoning + Sessions \(\rightarrow\) SOFM
  - Analyst

- **Web Site Document**
  - Off-line Enhancements
  - Interest
  - Utility

- Does it really work??
Structural Enhancement

Figure 4.7: New Link in Top Page
A JavaScript menu that contains links to all bylaws web pages. The menu appears when the mouse is over 'bylaws' button.

**Figure 4.6:** Bylaws Structure Modification
Content Enhancements

- New Section with contact information from all departments at the FCFM and organization from students

- We also keep a top navigation bar to access the same data to other organizations different to departments like students organizations

- This area is to display the information. Telephone, e-mail, name of responsible and charge, etc.

- We added a bar with shortcuts to find easily the information of the department or organization wanted

CONCEPT-BASED ENHANCEMENT
Enhancements obtained

Table 4.27: Enhancements Performed and its Code

<table>
<thead>
<tr>
<th>Code</th>
<th>Enhance Description</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Add Social Security in “About School” Menu</td>
<td>Concept-Based</td>
</tr>
<tr>
<td>M2</td>
<td>New Bylaws Menu</td>
<td>Concept-Based</td>
</tr>
<tr>
<td>M3</td>
<td>Creation of a “contact us” section - Organization</td>
<td>Concept-Based</td>
</tr>
<tr>
<td></td>
<td>contact information</td>
<td></td>
</tr>
<tr>
<td>M4</td>
<td>New menu to select between organizations information</td>
<td>Concept-Based</td>
</tr>
<tr>
<td>M5</td>
<td>New menu to select “other organizations” contact information</td>
<td>Concept-Based</td>
</tr>
<tr>
<td>M6</td>
<td>New Link to News Section</td>
<td>Concept-Based</td>
</tr>
<tr>
<td>M7</td>
<td>New Link to E-Movil service</td>
<td>Concept-Based, LSISOM</td>
</tr>
<tr>
<td>M8</td>
<td>New Link to Site Map</td>
<td>Concept-Based</td>
</tr>
<tr>
<td>M9</td>
<td>New Link to Social Security Area</td>
<td>Concept-Based, LSISOM</td>
</tr>
<tr>
<td>M10</td>
<td>New Link to Engineering Forum</td>
<td>Concept-Based</td>
</tr>
<tr>
<td>M11</td>
<td>New Link to BIA (grades report)</td>
<td>Concept-Based</td>
</tr>
<tr>
<td>M12</td>
<td>New Link to U-Cursos</td>
<td>Concept-Based, Human Intuition</td>
</tr>
<tr>
<td>M13</td>
<td>New Link to Personalized Schedule</td>
<td>Concept-Based</td>
</tr>
<tr>
<td>M14</td>
<td>New Link to ”Moises Mellado” Foundation</td>
<td>Human Intuition</td>
</tr>
<tr>
<td>M15</td>
<td>New Link to Open Science Magazine</td>
<td>Human Intuition</td>
</tr>
<tr>
<td>M16</td>
<td>New Link to WAP Site</td>
<td>Human Intuition</td>
</tr>
<tr>
<td>M17</td>
<td>New Link to PDA Site</td>
<td>Human Intuition</td>
</tr>
</tbody>
</table>
Better Performance

**Table 4.26:** Enhancements Obtained & Processing Times for Different Methods

<table>
<thead>
<tr>
<th>Algorithm</th>
<th># Modifications</th>
<th>Algorithm Processing Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Intuition</td>
<td>5</td>
<td>N/A</td>
</tr>
<tr>
<td>Traditional (SOFM)</td>
<td>0</td>
<td>~ 26h</td>
</tr>
<tr>
<td>Traditional (K-Means)</td>
<td>0</td>
<td>~ 5min</td>
</tr>
<tr>
<td>LSISOM</td>
<td>3</td>
<td>~ 11h</td>
</tr>
<tr>
<td>Concept-based</td>
<td>13</td>
<td>~ 15min</td>
</tr>
</tbody>
</table>
## Interest Correlation

### 100 STUDENTS SURVEY

<table>
<thead>
<tr>
<th>WUM Method</th>
<th>Interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Intuition</td>
<td>N/A</td>
</tr>
<tr>
<td>Traditional (SOFM)</td>
<td>44%</td>
</tr>
<tr>
<td>Traditional (K-means)</td>
<td>34.5%</td>
</tr>
<tr>
<td>LSISOM</td>
<td>50%</td>
</tr>
<tr>
<td>Concept-based</td>
<td>74%</td>
</tr>
</tbody>
</table>
Utility for Visitors

**Table 4.29: Global Usefulness Comparison for Different Methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>Subject Groups</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>L</td>
<td>P</td>
<td>L2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human Intuition</td>
<td>16.4%</td>
<td>37.3%</td>
<td>27.2%</td>
<td>15.5%</td>
<td></td>
<td>24.1%</td>
</tr>
<tr>
<td>Human Intuition*</td>
<td>0.2%</td>
<td>23.6%</td>
<td>13.5%</td>
<td>3.5%</td>
<td></td>
<td>10.2%</td>
</tr>
<tr>
<td>Traditional (SOFM)</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td></td>
<td>0.0%</td>
</tr>
<tr>
<td>Traditional (K-Means)</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td></td>
<td>0.0%</td>
</tr>
<tr>
<td>LSISOM</td>
<td>30.3%</td>
<td>54.6%</td>
<td>31.0%</td>
<td>62.3%</td>
<td></td>
<td>44.6%</td>
</tr>
<tr>
<td>Concept Based</td>
<td>56.6%</td>
<td>68.6%</td>
<td>50.1%</td>
<td>65.9%</td>
<td></td>
<td>60.3%</td>
</tr>
<tr>
<td>Concept Based*</td>
<td>54.6%</td>
<td>66.6%</td>
<td>47.4%</td>
<td>66.1%</td>
<td></td>
<td>58.7%</td>
</tr>
</tbody>
</table>

* In those methods was not considered enhance M12 in usefulness calculus.
Utility for Visitors

**Figure 4.16:** Utility of Enhancements by Percentage of People

- **Concept Based Method:**
  - Very Useful: 41.1%
  - Useful: 31.7%
  - Not very useful: 19.4%
  - Useless: 8.2%

- **LSISOM Based:**
  - Very Useful: 23.8%
  - Useful: 35.5%
  - Not very useful: 27.7%
  - Useless: 13.0%

- **Human Intuition:**
  - Very Useful: 37.1%
  - Useful: 17.7%
  - Not very useful: 28.6%
  - Useless: 16.9%
Conclusion
Quality of Enhancements

Enhancements based on Semantic WUM More correlated with visitors’ real interests

Therefore, Enhancements of Contents and Organization are of higher quality

Are more useful for the final visitors
Processing Times

It was possible to obtain much better results in much lesser processing times using Concept based WUM

- We reduce 24~30 h to only 15~20 minutes
- But usefulness was much greater (from 17.7% to 41.1%)
Make easier the analysis phase

Using semantic WUM is possible to discover more hidden patterns therefore to perform more enhancements (of high quality) on the web site
MUCHAS GRACIAS
ご静聴ありがとうございます - THANK YOU VERY MUCH