CECS694: Web Mining for E-Commerce and Information Retrieval

Instructor: Olfa Nasraoui, Ph.D.
Objectives

• Fundamentals of knowledge discovery in semi-structured/unstructured data with an emphasis on the World Wide Web:
  – Web usage, content, and structure mining;
  – applications to e-commerce, personalization, information retrieval, adaptive Web sites,…, etc

• Self contained course, hence no heavy pre-requisites other than good grasp of math and statistics
Textbooks/Literature

• **Modeling the Internet and the Web - Probabilistic methods and algorithms**, by Baldi, Frasconi, and Smyth, Publisher: Wiley

• **Data Mining: Introductory and Advanced Topics** by Margaret Dunham, Publisher: Prentice Hall,

• **Research papers**: will be provided by the instructor.

• **Course web page**: https://www.louisville.edu/~o0nasr01/Websites/Web_mining_course.html
Course Structure

- **Overview of basic data mining techniques**: Frequent itemset and association rule discovery, classification, clustering.
- **Web usage mining and Web analytics**: This includes Web usage data preprocessing, techniques for discovery and analysis of Web usage patterns, user modeling and profiling, integrating semantics and domain knowledge in Web usage mining and analysis.
- **Web content and structure mining**: includes integration of Web content, usage, and structure data for Web mining, text mining such as classification and clustering of text documents, detecting emerging trends or topics in text.
- **Applications**: Recommendation and personalization systems, contextual information access and retrieval, information filtering systems, Web mining applications for business and competitive intelligence, and Web log analysis for security applications.
Web Mining

• **Web Mining**: Mining or discovering useful knowledge or information from **Web data**

• What kinds of data on the Web?

  • **Web content**: Actual content of the web pages: text, HTML tags, embedded images, … etc

  • **Web structure**: structure = Hyperlink linking page 1 to page 2

  • **Web usage**: Record of usage or access activity on a website
Common Data Abstractions

• Web Content:
  – Text passage = bag of words, each word $\Leftrightarrow$ one dimension or attribute $\Rightarrow$ vector space definition, ex:
    • binary: 1 if word is present, 0 otherwise;
    • real/frequency: # of occurrences of the word

• Web Structure:
  – Each page $\Leftrightarrow$ a node in a graph
  – Link from pg 1 to pg2 $\Leftrightarrow$ Edge from node 1 to node 2
  – $\Rightarrow$ Directed graph
Common Data Abstractions

• Web Usage:
  – A user session = set/bag of page views ➔ vector space (see text passage above)
  – A user session = sequence of page views (order matters)
  – Page view: set of web content units needed to form an entire web page, may include html, embedded graphics, banners, frames, …etc.
What Can Be Inferred by Mining Web Data?

- **Web Content:**
  - Discovering clusters of web pages based on pure content ➔ Topical organization of a web site
  - Topical organization can be contrasted against actual local web structure (see below) for effective re-evaluation of web site design
  - Topical organization ↔ Summarization of website contents that is easier to understand
  - Summarization can be exploited to accelerate search, ex: by creating an index of web pages according to content
  - Summarization can be used to create a virtual structure of website and for visualization
What Can Be Inferred by Mining Web Data

- **Web Structure:**
  - Mine web graph (on global or local level) to discover
    - **Hubs:** pages that link to many other pages
    - **Authorities:** pages linked to from many other pages
    - Analysis of web graph can reveal **web communities** (divide graph into several highly internally linked sub-graphs): it is likely that pages concerned w/ similar topics will be more interconnected
    - Web communities reveal **topics**, and how they are related, can help reveal **polysemous** words (a word with more than one meaning, such as “cricket”)
    - **Web communities** can reveal **conceptual classes/clusters** in analogy w/ language → **intelligent organization**
    - Web graph analysis might reflect **emerging topics or concepts**, as well as **drift** and **popularity**

CECS 694: Web Mining for E-Commerce and Information Retrieval
What Can Be Inferred by Mining Web Data

• **Web Usage:**
  – Web user *sessions* can be *segmented* into several *clusters*
  – Clusters summarize different access *patterns/trends ➔ profiles*
  – Profiles provide *higher level* data about website user *behavior* and its *evolution* compared to mere sessions or clicks
  – Profiles can be used to recognize *modes of usage* and *personalize/adapt website* to each user
  – Web usage can be contrasted against web structure (see above) for effective *website design re-evaluation*
Example of profiles description discovered using web usage mining, with corresponding interestingness measures

<table>
<thead>
<tr>
<th>$i$</th>
<th>$X_i$</th>
<th>$X_i^*$</th>
<th>$N_i^*$</th>
<th>description</th>
<th>$\sigma_i^*^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>219</td>
<td>132</td>
<td>140.5</td>
<td>main page, class list, course enquiries and people</td>
<td>0.16</td>
</tr>
<tr>
<td>2</td>
<td>119</td>
<td>73</td>
<td>77.0</td>
<td>main page, class list, course and undergraduate degree enquiries</td>
<td>0.27</td>
</tr>
<tr>
<td>3</td>
<td>140</td>
<td>85</td>
<td>91.6</td>
<td>main page and class list</td>
<td>0.13</td>
</tr>
<tr>
<td>4</td>
<td>129</td>
<td>71</td>
<td>80.7</td>
<td>main page, people, individual faculty, research and graduate degree pages</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Example Profiles for MU-CECS1 at $L = 2$

- General outside visitor: Profiles 1 and 3
- Prospective students: Profiles 2 and 4
- Insiders (students): Profiles 6, 7, …etc
A Tree of 92 User Profile discovered from a 1-day web log of a main campus website

- Y-axis: distance between profiles between 0 and 1
- X-axis: profile index
- Profiles connected/ joined toward the bottom: more similar
- Hierarchical organization based on distance/dissimilarity