Web Mining

Student Projects

Prof. Dr. Christian Bizer
Cäcilia Zirn
Robert Meusel
Oliver Lehmberg

FFS 2014
Student Projects

■ Goals
  ■ Gain more practical experience with
    - Web Usage Mining,
    - Web Structure Mining or
    - Web Content Mining
  ■ Get to know additional problem-specific
    - preprocessing methods
    - Web mining methods

■ Expectation
  ■ Select an interesting Web mining problem of your choice
  ■ Solve the problem using
    - the Web mining methods that we have learned so far plus some advanced problem-specific data pre-processing
    - other Web mining methods which might be helpful for solving the problem and build on what we learned in class
Procedure

- **Teams of two students**
  1. realize a Web mining project
  2. write 10 to 12 page summary of the project and the methods employed in the project
  3. present the project results to the other students (12 minutes presentation + 8 minutes discussion)

- **Final mark for the course**
  - 50 % written exam
  - 30 % written summary about the project
  - 20 % project presentation
## Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Tuesday</th>
<th>Thursday</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4.2014</td>
<td>Introduction to Student Projects</td>
<td>Preparation of Project Outlines</td>
</tr>
<tr>
<td>8.4.2014, 23:59</td>
<td>Submission of Project Outlines</td>
<td></td>
</tr>
<tr>
<td>10.4.2014</td>
<td>Feedback about Project Outlines</td>
<td>Project work</td>
</tr>
<tr>
<td></td>
<td>Easter break</td>
<td></td>
</tr>
<tr>
<td>1.5.2014</td>
<td>Holiday</td>
<td>Coaching</td>
</tr>
<tr>
<td>8.5.2014</td>
<td>Project work</td>
<td>Coaching</td>
</tr>
<tr>
<td>15.5.2014</td>
<td>Project work</td>
<td>Coaching</td>
</tr>
<tr>
<td>19.5.2014, 23:59</td>
<td>Submission of project summaries</td>
<td>Presentation of Project Results</td>
</tr>
<tr>
<td>22.5.2014</td>
<td>Presentation of Project Results</td>
<td>Presentation of Project Results</td>
</tr>
<tr>
<td>11.6.2014</td>
<td>Written Exam</td>
<td></td>
</tr>
</tbody>
</table>
Where to find Web Usage Data Sets?

- **MovieLens**
  - Movie ratings collected via MovieLens website
  - 10M Dataset: 71,000 users, 10,600 movies, 10 million ratings

- **Netflix**
  - Provided by commercial movie rental website for Netflix competition ($1,000,000 for 10% better RMSE)
  - 480,000 users rated 18,000 movies, 100M ratings

- **Amazon Public Datasets**
  - Wikipedia Traffic Statistics

- **Stanford Large Network Dataset Collection**
  - Amazon product co-purchasing data

- **Yahoo Webscope**
  - Music and Movie rating data

- **Web 2.0 Platforms offer plenty of additional rating data**
  - E.g. LastFM, delicious
Where to find Web Structure Data Sets?

- **Stanford Large Network Dataset Collection**
  - Social networks: Facebook, Google+
  - Citation networks: Arxiv, US Patents
  - Product co-purchasing network: Amazon

- **Common Crawl Hyperlink Networks**
  - aggregated by Website (40 million vertices)
  - [http://webdatacommons.org/hyperlinkgraph/](http://webdatacommons.org/hyperlinkgraph/)

- **Billion Triples Challenge Dataset**
  - Linked Data crawled from the public Web of Linked Data in May/June 2012
  - 1.4 billion triples (17 GB gzipped)

- **Archive.org**
  - Friendster social network (2011)

- **Infochimps**
  - CiteULike ([http://www.citeulike.org/](http://www.citeulike.org/))
Where to find Web Content Data Sets?

- **Multiple-aspect Restaurant Reviews**
  - [http://people.csail.mit.edu/bsnyder/naacl07](http://people.csail.mit.edu/bsnyder/naacl07)
  - Reviews taken from we8there, ratings for 5 main aspects of restaurants

- **The J.D. Power and Associates Sentiment Corpus**
  - [http://verbs.colorado.edu/jdpacorpus/](http://verbs.colorado.edu/jdpacorpus/)
  - Blog entries about cars and cameras, manually labeled with product features (including labeled part-of relations), opinion phrases and opinion targets

- **Stanford Large Network Dataset Collection**
  - Amazon product metadata and review information about 548,552 different products

- **Yahoo Webscope**
  - Topics: Web search, advertising, review data

- **Programmable Web**
  - Website giving an overview about 4100 public Web APIs

- **academictorrents.com**
  - Various large data sets
  - e.g. Enron Email Bag of Words, Arizona State University Twitter Data Set
Some Project Ideas (not binding)

- **Web Usage Log Mining**
  1. Learn a classifier for categorizing the visitors of your website.
  2. Identify common navigation paths, drop-out pages

- **Recommender Systems**
  1. Defend a recommender system by identifying fake ratings
  2. Experiment with hybrid recommenders in specific application domain

- **Network Analysis**
  1. Common Crawl Link Network (analyze by country or topical domain)
  2. Linked Data Cloud (analyze by country, topical domain)

- **Sentiment Analysis**
  1. Extracting product features and opinions using advanced methods
  2. Generating opinion summaries (aggregating extracted information)
Project Outlines

- 2-3 pages (sharp!) without title or toc pages, DWS master thesis layout
- due 8.4.2014, 23:59
- send by eMail to Chris, Cäcilia, Oliver, and Robert
- answer the following questions:

1. What is the problem you are solving?
2. What data will you use?
   - Where will you get it?
   - How will you gather it?
3. How will you solve the problem?
   - What preprocessing steps will be required?
   - Which algorithms you plan to use?
   - Be as specific as you can!
4. How will you evaluate, measure success?
Coaching Sessions

■ We will give you tipps and answer questions concerning your project.

■ Please send us an email that you want to attend a coaching session.
  ■ until Tuesday night
  ■ including the questions that you like to discuss

■ We will assign you a time slot afterwards and inform you about the slot via email.
Project Summaries

- 10-12 pages (sharp!) without title or toc pages
- due 19.5.2014, 23:59
- send by email to Chris, Cäcilia, Oliver, and Robert
- describe your solution including the steps to get there:
  1. Application domain and goals
  2. Structure and size of the data set
  3. Preprocessing
  4. Actual Web Mining
  5. Evaluation
  6. Results

- Requirements
  - You must use the DWS master thesis layout.
  - Please cite sources properly. Preferred citation style [Author, year].
  - Also submit your Rapid Miner processes and (a subset) of your data.
Final Exam

- **Date:** 11.6.2014
- **Duration:** 60 minutes
- **Structure:** 5 - 6 open questions that
  - check whether you have understood the content of the lecture
  - require you to describe the ideas behind algorithms and methods
  - might require you to do some simple calculations