Web Data Mining
Trends & Techniques
Authors: Ujwala Manoj Patil & J.B. Patil
Publication: August 2012

Team Members:
Vishma Shah
Pooja Vora
Background & Problem Definition

Three types of Mining:

- Data Mining
- Text Mining
- Web Mining

Web Data Mining: Deals with extraction of interesting or hidden knowledge from the World Wide Web.

Problem Definition: Web Data Mining Techniques
Technical Highlights of Problem Solving

In **Web Data Mining**, data can be collected from Web servers, client sites, and proxy server or obtained from organization’s database.

**Types of Data that can be used in Web Data Mining:**

- **Web Content:** Data on web pages e.g. HTML, audio, images
- **Web Structure:** Organization of web pages e.g. hyperlinks
- **Web Usage:** Data related to usage of Web e.g. web logs
- **Web User Profile:** User’s information e.g. registration data, profile data

**Categories Of Web Data Mining:**

- Web Content Mining
- Web Usage Mining
- Web Structure Mining
Illustration of the Methods Introduced

Web Content Mining:

- Discovery of useful info from Web Content
- Algorithms based on unstructured data (e.g. text, video, audio) or semi-structured HTML data
- Goal:
  - From Database view - model the data on the web and to integrate it -> more sophisticated queries can be performed
  - What about Informational Retrieval View..??
- Mining results can be used to build Web Warehouse & Web DB -> apply warehouse & database techniques on the data.
- Few Applications: Classifying Web documents, Comparing website contents, content based image retrieval
Web Structure Mining:

- Discover useful knowledge based on hyperlink structure of the website
- used to classify Web pages into authority and hub pages -> generate similarities or the difference between different Web sites.
- Application Example: algorithm that extract the structure of Web site based on hyperlink analysis -> identifies and filters noise hyperlinks.
Web Usage Mining:

- Understanding user behavior when user interacts with the Web -> from behavior. Websites can be reorganized according to the user requirements.
- Use web log sources i.e. cookies, client browser history, proxy log.
- Application Example: Prediction of user’s future navigation intentions based on the navigation pattern.
- One more Example: Analyze the proxy log to insure that all users get fair share of the bandwidth to control the Internet usage.
Pros & Cons / Opinion

Pros:
- Very useful for personalized marketing
- Identifying Criminal Activities

Cons:
- Most critical issue-> invasion of privacy

Opinion:
References

- http://www.web-datamining.net/
Data Warehouse & Data Mining
Project Proposal
Data warehouse Project:
Title: OLAP operations on Company Dataset.

Dataset:
Company data set contains information about employee and their work records. Like department, products, working hours.

Purpose:
- Analyze employees work schedule
- Calculate employees productivity
Data Mining Project :
Title : Data Mining on Consumer Complains Dataset

Dataset:
• Data set contains information related to Consumer issues and Company responses.
• Issues like: Using a debit or ATM card, Incorrect information on credit report, False statements or representation on Debt Collection, etc.
• Around 14 attributes & above 70000 rows (data from Nov 2011 - Nov 2014)

Purpose:
• Predict type of recurring issues
• Predict company’s customer support
Methods

• Analyze the Data sets
• Preprocessing the Data – Data Cleaning
• Reduce the dataset
• For DW Project - Apply OLAP Operations
• For DM Project - Apply Algorithms, using a tool
  • Initial Thought - one classification algorithm & one clustering algorithm.
Schedule

Schedule:

- Week1: Data Preprocessing->Data cleaning
- Week2: Data mart with OLAP operations, Fact Table & Star Schema generation
- Week3: Use tool(WEKA/Rapidminer) for Data mining Project and Apply algorithm and Analyze the results
- Week4: Final Testing & Final Report
References

- DM dataset http://www.data.gov/consumer/
- OLAP Courseware http://athena.ecs.csus.edu/~olap/olap/introduction.php
- Data Mining Courseware http://athena.ecs.csus.edu/~datamini/