What is Web Mining ?

1 Overview

Data Mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. (from D.Hand, H. Mannila, P. Smyth, *Principles of Data Mining*)

When observational dataset come from the World Wide Web, their analysis is called **Web Mining**:

The textbook states:

> **Web mining** is the discovery of useful information or knowledge from the **Web hyperlink structure**, **page content** and **usage data**.

Categories of web mining tasks:

- **Web structure mining.**
- **Web content mining.**
- **Web usage mining.**
- **User modeling/Opinion mining.**

Web Structure Mining

Web structure mining is discovery of information from the organizational structure of web pages.

Data used. The key dataset is the *World Wide Web structure* represented as a *directed graph.*

Problems studied:

- Link analysis
- Community discovery
- Page importance analysis (e.g., PageRank)
- Co-citation analysis
- Social network analysis

Web content mining

Web content mining is discovery of information from the content (as well as possibly hyperlinks) of the web pages.

Data used. Page contents (usually indexed in some form).

Problems studied:

- Information retrieval (web search engines)
- Topic discovery/classification
- Information extraction
- Opinion mining (see below)

Web usage mining

Web usage mining is discovery of information about the human use of the Web.

Data used. Web server logs.

Problems studied:

- Preprocessing of clickstream data
- Session detection
- Visitor analysis/classification
- Behavior prediction

User modeling/Opinion mining.

User modeling is discovery of information about the users of the Web from observation of their behavior on the World Wide Web. User modeling combines elements of **web content mining** (discovery of human opinions expressed on existing web pages) and **web usage mining** (analysis of web traversal behavior).

Data used. (any or all of) Web page content, web server logs, web structure graph.

Problems studied.

- Collaborative filtering
- Opinion search
- Opinion classification
- User Clustering (discovery of groups of users with similar behavior)
- Predictive modeling