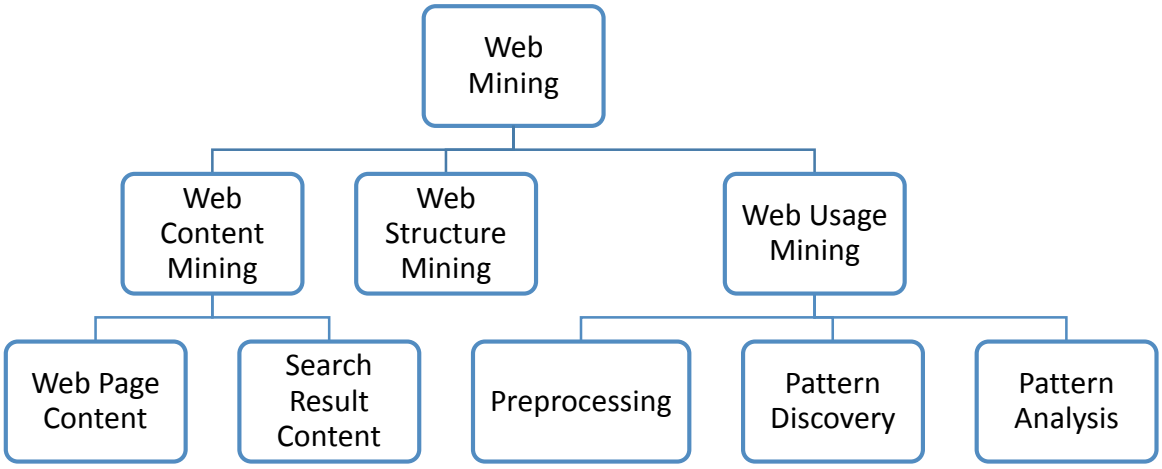
Advanced Applications

7.1 Web Mining

In recent years the growth of the World Wide Web exceeded all expectations. Today there are several trillions of HTML documents, pictures and other multimedia files available via internet and the number is still rising. The WWW is huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, hyperlink information, access and usage information. But considering the impressive variety of the web, retrieving interesting content has become a very difficult task.

Web mining research overlaps substantially with other areas, including data mining, text mining, information retrieval, and web retrieval.



One possible categorization of Web mining is based on which part of the Web is mined. There are three main areas of Web mining: *Web content mining*, *Web structure mining*, and *Web usage mining*. Each area is classified by the type of data used in the mining process. Web content mining uses Web page content as the data source for the mining process. This could include text, images, videos, or any other type of content on Web pages. Web structure mining focuses on the link structure of Web pages. Web usage mining does not use data from the Web itself but takes as input data recorded from the interaction of users using the Internet.

i. Web Content Mining

This is the process of mining useful information from the contents of Web pages and Web documents, which are mostly text, images and audio/video files. Techniques used in this discipline have been heavily drawn from natural language processing (NLP) and information retrieval.

The most common use of Web content mining is in the process of searching. There are many different solutions that take as input Web page text or images with the intent of helping users find information that is of interest to them. For example, crawlers are currently used by search engines to extract web content into the indices that allow immediate feedback from searches. The same crawlers can be altered in such a way that rather than seeking to download all reachable content on the Internet, they can be focused on a particular topic or area of interest.

Web content mining can also be seen directly in the search process. All major search engines currently use a list like structure to display search results. The list is ordered by a ranking algorithm behind the scenes. An alternative view of search results that has been attempted is to provide the users with clusters of Web pages as results rather than individual Web pages. Often a hierarchical clustering that will give multiple topic levels is performed.

ii. Web Structure Mining

Web structure mining considers the relationships between Web pages. Most Web pages include one or more hyperlinks. These hyperlinks are assumed in structure mining to provide an endorsement by the linking page of the page linked. There are two things that can be obtained from this: the structure of a website in terms of how it is connected to other sites and the document structure of the website itself, as to how each page is connected. This type of structure mining can be used to reveal the structure (schema) of Web pages; this would be good for navigation purpose and make it possible to compare/integrate Web page schemes.

The goal of Web structure mining is to generate structural summary about the Web site and Web page. Technically, Web content mining mainly focuses on the structure of inner document, while Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different Web sites.

Web structure mining is also used to aid in Web content mining processes. Often, classification tasks will consider features from the content of the Web page and may consider the structure of the Web pages. One of the more common features in Web - mining tasks taken from structure mining is the use of anchor text. Anchor text refers to the text displayed to users on an HTML hyperlink. Oftentimes the anchor text provides summary keywords not found on the original Web page. The anchor text is often as brief as search engine queries.

iii. Web Usage Mining

Web Usage Mining focuses on techniques that could predict the behavior of users while they are interacting with the WWW. Web usage mining, discover user navigation patterns from web data, tries to discover the useful information from the secondary data derived from the interactions of the users while surfing on the Web.

Web usage mining takes advantage of many of the data mining approaches available. Classification may be used to identify characteristics unique to users that make large purchases. Clustering may be used to segment the Web user population. For example, one may identify three types of behavior occurring on a university class Web site. These three behavior patterns could be described as users studying for a test, users working on projects, and users consistently downloading lecture notes from home for study. Association mining may identify two or more pages often viewed together during the same session, but that are not directly linked on a Web site. Sequence analysis may offer opportunities to predict user navigation patterns and therefore allow for within site, targeted advertisements.

7.2 Time Series Data Mining

Time-series database consists of sequences of values or events obtained over repeated measurements of time. The values are typically measured at equal time intervals (e.g., hourly, daily, weekly). Time-series databases are popular in many applications, such as stock market analysis, economic and sales forecasting, financial analysis, utility studies, inventory studies, revenue projections, workload projections, process and quality control, observation of natural phenomena (such as atmosphere, temperature, wind, earthquake), scientific and engineering experiments, and medical treatments. A time-series database is also a sequence database. However, a sequence database is any database that consists of sequences of ordered events, with or without

concrete notions of time. For example, Web page traversal sequences and customer shopping transaction sequences are sequence data, but they may not be time-series data.

How can we find correlation relationships within time-series data? How can we analyze such huge numbers of time series to find similar or regular patterns, trends, bursts (such as sudden sharp changes), and outliers, with fast or even on-line real-time response? This has become an increasingly important and challenging problem.

Trend Analysis

In general, there are two goals in time-series analysis:

- **Modeling time series** (i.e., to gain insight into the mechanisms or underlying forces that generate the time series), and
- **Forecasting time series** (i.e., to predict the future values of the time-series variables).

Trend analysis consists of the following four major components or movements for characterizing time-series data:

i. Trend or long term movements

These indicate the general direction in which a time series graph is moving over a long interval of time. This movement is displayed by a trend curve, or a trend line.

ii. Cyclic movements or cyclic variations:

These refer to the cycles, that is, the long-term oscillations about a trend line or curve, which may or may not be periodic. That is, the cycles need not necessarily follow exactly similar patterns after equal intervals of time.

iii. Seasonal movements or seasonal variations:

These are systematic or calendar related. Examples include events that recur annually, such as the sudden increase in sales of chocolates and flowers before Valentine's Day or of department store items before Christmas. The observed increase in water consumption in summer due to warm weather is another example. In these examples, seasonal movements are the identical or nearly identical patterns that a time series appears to follow during corresponding months of successive years.

iv. **Irregular or random movements:** These characterize the sporadic motion of time series due to random or chance events, such as labor disputes, floods, or announced personnel changes within companies.

Note that regression analysis has been a popular tool for modeling time series, finding trends and outliers in such data sets.

References

[1] J. Han and K. Micheline, Data Mining: Concepts and Techniques, San Francisco: Elsevier Inc., 2006.

[2] M. Kantardzic, DATA MINING Concepts, Models, Methods, and Algorithms, New Jersey: John Wiley & Sons, 2011.