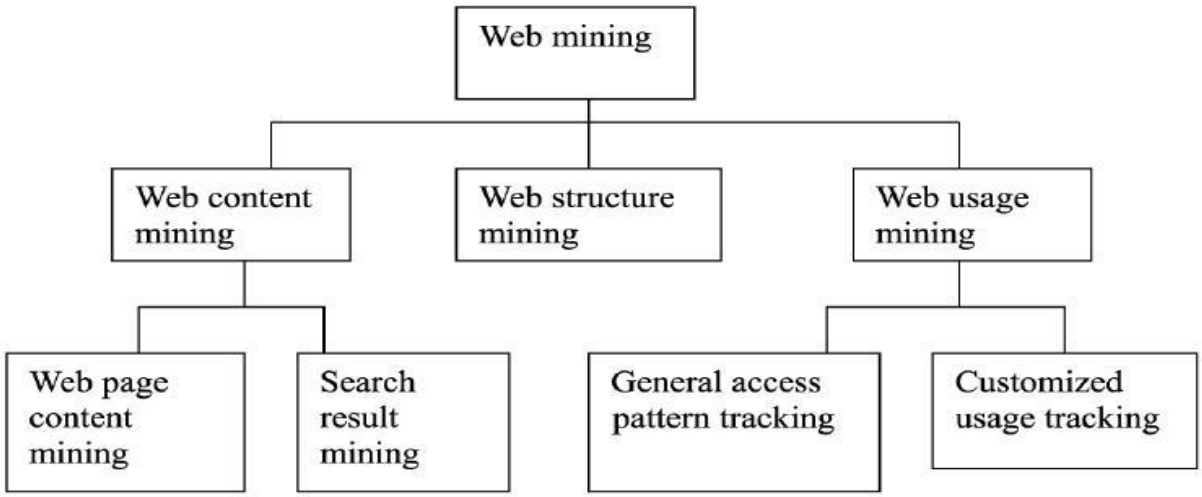
Unit 7: Advanced Application

Web mining:

Web mining is the application of data mining techniques to find interesting and potentially useful knowledge from web data. In other words, it is the use of the data mining techniques to automatically discover and extract information from web documents/services. Web mining can be **broadly divided into three categories**:



1. **Web Content Mining**
2. **Web Structure Mining**
3. **Web Usage Mining.**

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.

In the past few years, there was a rapid expansion of activities in the Web content mining area. This is not surprising because of the phenomenal growth of the Web contents and significant economic benefit of such mining. However, due to the heterogeneity and the lack of structure of Web data, automated discovery of targeted or unexpected knowledge information still present many challenging research problems.

Web Content Mining Problems/Challenges

- Data/Information Extraction: Extraction of structured data from Web pages, such as products and search results is a difficult task. Extracting

such data allows one to provide services. Two main types of techniques, machine learning and automatic extraction are used to solve this problem.

- Web Information Integration and Schema Matching: Although the Web contains a huge amount of data, each web site (or even page) represents similar information differently. Identifying or matching semantically similar data is a very important problem with many practical applications.
- Opinion extraction from online sources: There are many online opinion sources, e.g., customer reviews of products, forums, blogs and chat rooms. Mining opinions (especially consumer opinions) is of great importance for marketing intelligence and product benchmarking.
- Knowledge synthesis: Concept hierarchies or ontology are useful in many applications. However, generating them manually is very time consuming. A few existing methods that explores the information redundancy of the Web will be presented. The main application is to synthesize and organize the pieces of information on the Web to give the user a coherent picture of the topic domain.
- Segmenting Web pages and detecting noise: In many Web applications, one only wants the main content of the Web page without advertisements, navigation links, copyright notices. Automatically segmenting Web page to extract the main content of the pages is interesting problem.

All these tasks present major research challenges and their solutions.

Web Structure Mining

Web Structure Mining focuses on analysis of the link structure of the web and one of its purposes is to identify more preferable documents. The different objects are linked in some way. 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.

Simply applying the traditional processes and assuming that the events are independent can lead to wrong conclusions. However, the appropriate handling of the links could lead to potential correlations, and then improve the predictive accuracy of the learned models.

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 can also have another direction – discovering the structure of Web document itself. 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. This type of structure mining will facilitate introducing database techniques for accessing information in Web pages by providing a reference schema.

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 collects the data from Web log records to discover user access patterns of web pages. There are several available research projects and commercial tools that analyze those patterns for different purposes. The insight knowledge could be utilized in personalization, system improvement, site modification, business intelligence and usage characterization.

The only information left behind by many users visiting a Web site is the path through the pages they have accessed. Most of the Web information retrieval tools only use the textual information, while they ignore the link information that could be very valuable.

In general, there are mainly four kinds of data mining techniques applied to the web mining domain to discover the user navigation pattern: Association Rule mining

Sequential pattern

Clustering

Classification

Applications of Web Mining

With the rapid growth of World Wide Web, Web mining becomes a very hot and popular topic in Web research. E-commerce and E-services are claimed to be the killer applications

for Web mining, and Web mining now also plays an important role for E-commerce website and E-services to understand how their websites and services are used and to provide better services for their customers and users.

A few applications are:

E-commerce Customer Behavior Analysis

E-commerce Transaction Analysis

E-commerce Website Design

E-banking

M-commerce

Web Advertisement

Search Engine

Online Auction.

Time-series data mining

A 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, inventory studies, 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.

With the growing deployment of a large number of sensors, telemetry devices, and other on-line data collection tools, the amount of time-series data is increasing rapidly, often in the order of gigabytes per day (such as in stock trading) or even per minute (such as from NASA space programs).

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.

"How can we study time-series data?" In general, there are two goals in time-series analysis:

(1) modeling time series (i.e., to gain insight into the mechanisms or underlying forces that generate the time series), and

(2) 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:

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. For example, the trend curve of Figure 8.4 is indicated by a dashed curve. Typical methods for determining a trend curve or trend line include the weighted moving average method and the least squares method, discussed later.

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.

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.

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.