# WEB MINING OVERVIEW

**Michael Kunc**
Doctoral Degree Programme (1), FIT BUT
E-mail: kunc@fit.vutbr.cz

Supervised by: Jaroslav Zendulka
E-mail: zendulka@fit.vutbr.cz

## ABSTRACT

Methods of Web data mining can be divided into three categories according to a type of mined information and goals that particular categories set: Web content mining, Web structure mining and Web usage mining. Web content mining is the process of extracting useful information from the content of Web documents. Web structure mining uses the hyperlink structure of the Web to yield useful information, including definitive pages specification, hyperlinked communities identification, Web pages categorization and Web site completeness evaluation. Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. This paper is an overview of these techniques.

## 1 INTRODUCTION

The World Wide Web (Web) is popular and interactive medium to disseminate information today. The Web is huge, diverse, dynamic, widely distributed global information service center. Users could encounter following problems when interacting with the Web:

*a) Finding relevant information*

Most people use some search service when they want to find specific information on the Web. A user usually inputs a simple keyword query and a result is a list of ranked pages. This ranking is based on their similarity to the query. Today's search tools have some problems: Low precision and low recall, mainly because of wrong or incomplete keyword query. This leads to irrelevance of many search results.

*b) Creating new knowledge*

This problem is data-triggered process that presumes that we have a collection of Web data and we want to extract potentially useful knowledge from these data.

*c) Personalisation of information*

People differ in the contents and presentations they prefer while interacting with the Web.

*d) Learning about consumers or individual users*

This is a group of sub-problems such as mass customizing information to intended consumers, problems related to effective Web site design and management, problems related to marketing and others.

Web mining techniques could be used to solve information overload problems above.

## 2 WEB MINING

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. Web mining should be decomposed into these subtasks:

1. Resource finding: The task of retrieving intended Web documents.
2. Information selection and preprocessing: Automatically selecting and preprocessing specific information from retrieved Web resources.
3. Generalization: Automatically discovers general patterns at individual Web sites as well as across multiple sites.
4. Analysis: Validation and/or interpretation of the mined patterns.

Resource finding is the process of retrieving data from text sources available on the Web such as electronic magazines and newsletters or text contents of HTML documents. Information selection and preprocessing step is transformation process retrieved in information retrieval (IR) process from original data. These transformations cover removing stop words, finding phrases in the training corpus, transforming the representation to relational or first order logic form, etc. Data mining techniques and machine learning are often used for generalization. In information and knowledge discovery process, people play very important role. This is important for validation and/or interpretation in last step.

### 2.1 WEB MINING CATEGORIES

Web mining is categorized into three areas of interest based on part of Web to mine:

1. Web content mining
    - describes discovery of useful information from contents, data and documents
    - two different points of view: IR view and DB view
2. Web structure mining
    - model of link structures, topology of hyperlinks
    - categorizing of web pages
3. Web usage mining
    - mines secondary data derived from user interactions

Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web structure mining is the process of inferring knowledge from the Web organization and links between references and referents in the Web. Finally, Web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in Web access logs.

## 3 WEB CONTENT MINING

Web content mining is the process of extracting useful information from the content of Web documents. Logical structure, semantic content and layout are contained in semi-structured Web page text.

Topic discovery, extracting association patterns, clustering of Web documents and classification of Web pages are some of research issues in text mining. These activities use techniques from other disciplines – IR, IE (information extraction), NLP (natural language processing) and others.

Automatic extraction of semantic relations and structures from Web is a growing application of Web content mining. In this area, several algorithms are used: Hierarchical clustering algorithms on terms in order to create concept hierarchies, formal concept analysis and association rule mining to learn generalized conceptual relations and automatic extraction of structured data records from semi-structured HTML pages. Primary goal of each algorithm is to create a set of formally defined domain ontologies that represent Web site content. Common representation approaches are vector-space models, descriptive logics, first order logic, relational models and probabilistic relational models.

Structured data extraction is one of most widely studied research topics of Web content mining. Structured data on the Web are often very important as they represent their host pages' essential information. Extracting such data allows one to provide value added services, e.g. shopping and meta-search. In contrast to unstructured texts, structured data is also easier to extract. This problem has been studied by researchers in AI and database and data mining.

## 4 WEB STRUCTURE MINING

Web structure mining uses the hyperlink structure of the Web to yield useful information, including definitive pages specification, hyperlinked communities identification, Web pages categorization and Web site completeness evaluation. Web structure mining can be divided into two categories based on the kind of structured data used:

1. Web graph mining: The Web provides additional information about how different documents are connected to each other via hyperlinks. The Web can be viewed as a (directed) graph whose nodes are Web pages and whose edges are hyperlinks between them.

2. Deep Web mining: Web also contains a vast amount of noncrawlable content. This hidden part of the Web is referred to as the *deep Web* or the *hidden Web*. Compared to the static surface Web, the deep Web contains a much larger amount of high-quality structured information.

Most of mining algorithms, that are improving the performance of Web search, are based on two assumptions. (a) Hyperlinks convey human endorsement. If there exists a link from page A to page B, and these two pages are authored by different people, then the first author found the second page valuable. Thus the importance of a page can be propagated to those pages it links to. (b) Pages that are co-cited by a certain page are likely related to the same topic. The popularity or importance of a page is correlated to the number of incoming links to some extendt,

and related pages tend to be clustered together through dense linkages among them.

Web information extraction has the goal of pulling out information from a collection of Web pages and converting it to a homogeneous form that is more readily digested and analyzed for both humans and machines. The result of IE could be used to improve the indexing process, because IE removes irrelevant information in Web pages and facilitates other advanced search functions due to the structured nature of data.

It is usually difficult or even impossible to directly obtain the structures of the Web sites' back-end databases without cooperation from the sites. Instead, the sites present two other distinguishing structures: Interface schema and result schema. The *interface schema* is the schema of the query interface, which exposes attributes that can be queried in the backend database. The *result schema* is the schema of the query results, which exposes attributes that are shown to users.

## 5 WEB USAGE MINING

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Capturing, modeling and analyzing of behavioral patterns of users is the goal of this web mining category. Web usage mining process can be divided into three independent tasks: Preprocessing, pattern discovery and pattern analysis. The Figure 1 shows this process.

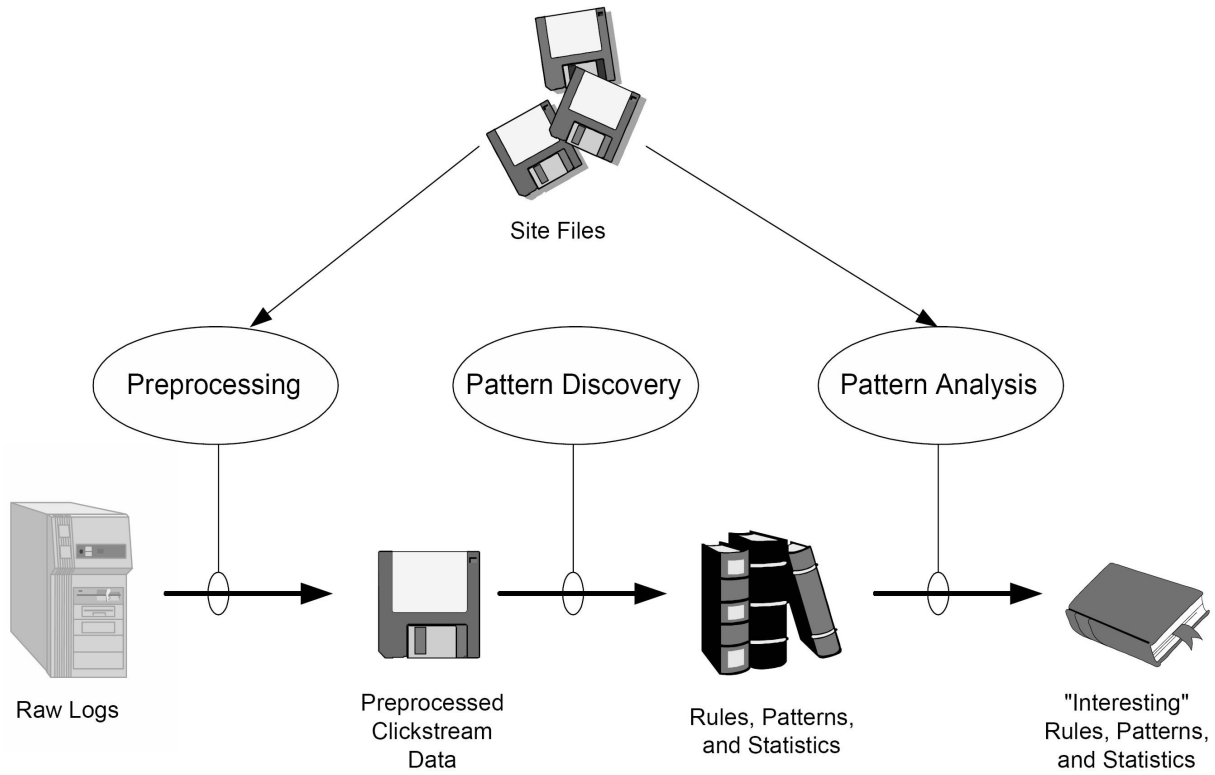


Figure 1: The process of Web usage mining (taken over from [3])

Preprocessing is first phase of Web mining process. Usage, content and structure information contained in the various available data sources are converted for next step that is pattern discovery. Pattern discovery is based on methods and algorithms developed from several areas such as data mining, machine learning and pattern recognition. This is used for understanding how users use some Web site. Pattern analysis is the final step in Web usage mining process. In this phase, uninteresting rules or patterns from the set found in pattern discovery are filtered. It turns discovered patterns, rules and statistics into knowledges. Knowledge query mechanism such as SQL is a form of pattern analysis. Loading usage data into a data cube in order to perform OLAP operations is another method. Highlighting overall patterns or trends in the data is usually done by some visualization technique, such as graphing patterns or assigning colors to different values.

## 6 CONCLUSION

This paper describes Web mining. It is an application of data mining techniques to extract knowledge from the content, structure, and usage of Web data sources. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web structure mining is the process of inferring knowledge from the Web organization and links between references and referents in the Web. Finally, Web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in Web access logs. There are many applications of these techniques, for example search engines, Web analysis, Web agents, personalization services etc. New modifications and extensions of these techniques should be next topics in this area of research.

## REFERENCES


[1] Han, J., Kamber, M.: Data Mining: Concepts and Techniques, Second edition, p. 628-648. Morgan Kaufmann Publishers, 2006.

[2] Liu, B., Chang, K. Ch.: Editorial: Special Issue on Web Content Mining. SIGKDD Explorations, 2004. Paper available on WWW: `http://delivery.acm.org/10.1145/1050000/1046457/p1-liu.pdf` (January 2007).

[3] Srivastava, J., Cooley, R., Deshpande, M., Tan, P.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, 2000. Paper available on WWW: `http://www.acm.org/sigs/sigkdd/explorations/issue1-2/srivastava.pdf` (January 2007).

[4] Vakali, A., Pallis, G.: Web Data Management Practices: Emerging Techniques and Technologies. Idea Group Publishing, 2007.