WEB MINING TASKS AND TYPES: A SURVEY

Chintandeep Kaur*

Rinkle Rani Aggarwal*

ABSTRACT

In recent years the growth of the World Wide Web has exceeded all expectations. Today there are several billions of HTML documents, pictures and other multimedia files available via internet and the number is still rising.As a large and dynamic information source that is structurally complex and ever growing, the World Wide Web is fertile ground for data-mining principles, or Web mining. In 1996 it's Etzioni who first coined the term web mining. Etzioni starts by making a hypothesis that information on web is sufficiently structured and outlines the subtasks of web mining. Web mining is a very hot research topic which combines two of the activated research areas: Data Mining and World Wide Web. The Web mining research relates to several research communities such as Database, Information Retrieval and Artificial Intelligence. Web mining basically can be divided into three categories: web content mining, web structure mining and web usage mining, these three categories deal with different features of a web page, web content mining deals with discovering useful information or knowledge from web page contents, web structure mining deals with discovering and modelling the link structure of web, web usage mining is used to discover interesting usage patterns from web data. This paper is a survey paper which explains in detail the concepts of web mining focusing on tasks and types of web mining.

*Department of Computer Science & Engineering, Thapar University, Patiala.

1. INTRODUCTION

Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web[18]. Just as data mining aims at discovering valuable information that is hidden in conventional databases, the emerging field of web mining aims at finding and extracting relevant information that is hidden in Web-related data, in particular hyper-text documents published on the Web [2]. Web-mining is a multi-disciplinary effort that draws techniques from fields like in-formation retrieval, statistics, machine learning, natural language processing, and others. Web mining has new character compared with the traditional data mining. First, the objects of Web mining are a large number of Web documents which are heterogeneously distributed and each data source are heterogeneous; second, the Web document itself is semi-structured or un-structured and lack the semantics the machine can understand. This area of research is so huge today due to the tremendous growth of information sources available on the web and the recent interest in e-commerce[15].

2. WEB MINING SUBTASKS

Web mining is used to understand customer behavior, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign [3]. Web mining can be decomposed into the subtasks, namely:

a) Resource finding: The task of retrieving intended Web documents. By resource finding we mean the process of retrieving the data that is either online or offline from the text sources available on the web such as electronic newsletters, electronic newswire, the text contents of HTML documents obtained by removing HTML tags, and also the manual selection of Web resources.

b) Information selection and pre-processing: Automatically selecting and pre-processing specific information from retrieved Web resources. It is a kind of transformation processes of the original data retrieved in the IR process. These transformations could be either a kind of pre-processing that are mentioned above such as stop words, stemming, etc. or a pre-processing aimed at obtaining the desired representation such as finding phrases in the training corpus, transforming the representation to relational or first order logic form, etc.

c) Generalization: It automatically discovers general patterns at individual Web sites as well as across multiple sites. Machine learning or data mining techniques are typically used in the process of generalization. Humans play an important role in the information or knowledge discovery process on the Web since the Web is an interactive medium.

d) Analysis:Validating and/or interpretation of the mined patterns [15].

3. WEB MINING CATEGORIES

Kosala and Blockeel [14] who perform research in the area of web mining and suggest the three web mining categories depending on which kind of data to be mined that is mining for information or mining the web link structure or mining for user navigation patterns as shown in figure 1.

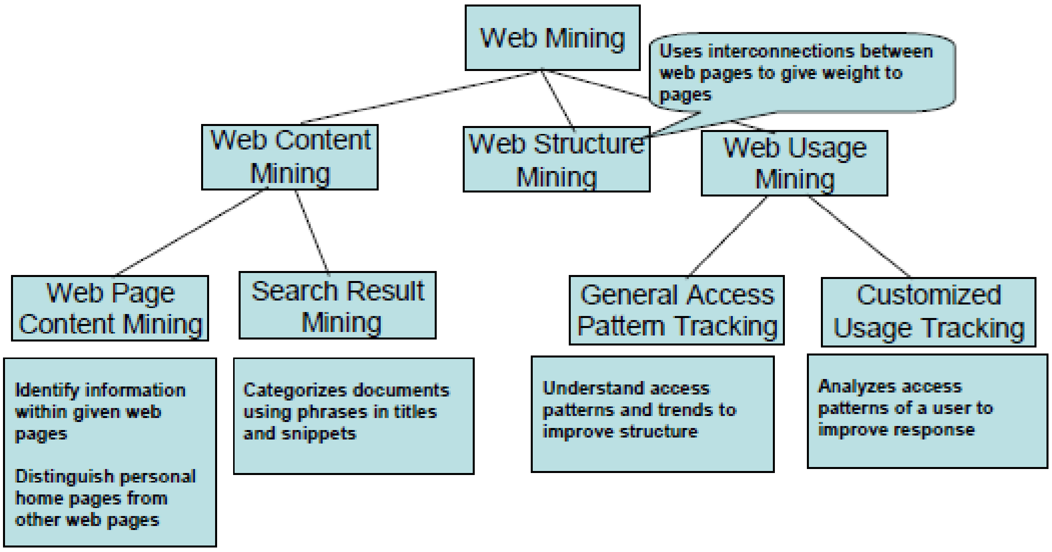
Web Mining
Uses interconnections between web pages to give weight to pages
Web Content Mining
Web Structure Mining
Web Usage Mining
Web Page Content Mining
Search Result Mining
General Access Pattern Tracking
Customized Usage Tracking
Identify information within given web pages
Distinguish personal home pages from other web pages
Categorizes documents using phrases in titles and snippets
Understand access patterns and trends to improve structure
Analyzes access patterns of a user to improve response

Figure 1 : Web mining taxonomy[16]

3.1 **Content Mining** : Mining for information focuses on the development of techniques for assisting a user in finding documents that meet a certain criterion that is web content mining. Web content mining refers to the discovery of useful information from web contents, including text, image, audio, video, etc mining the link structure aims at developing techniques to take advantage of the collective judgment of web page quality which is available in the form of hyperlinks that is web structure mining [5].

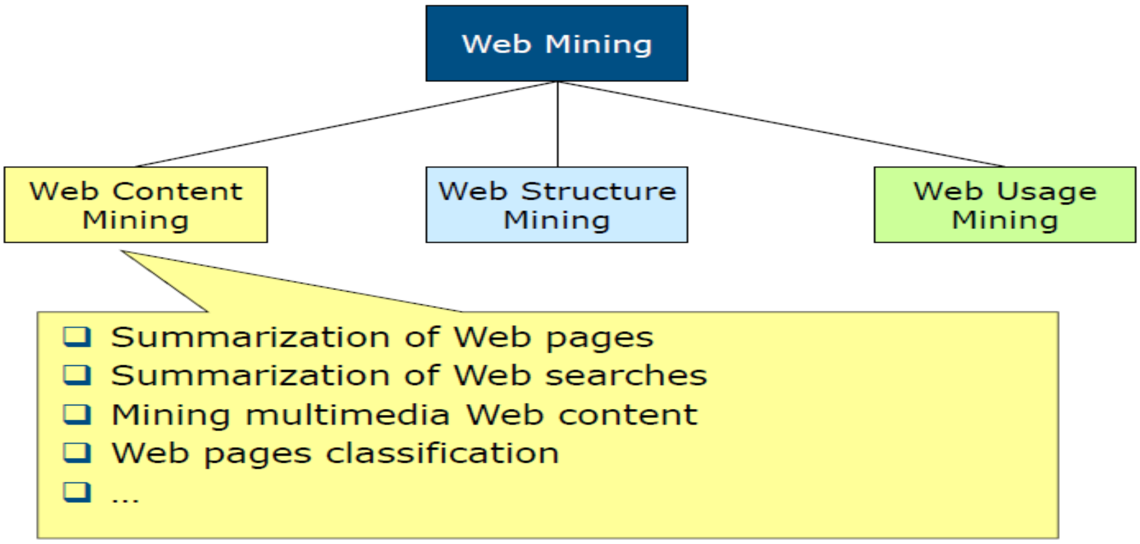
Web Mining
Web Content Mining
Web Structure Mining
Web Usage Mining
❑ Summarization of Web pages
❑ Summarization of Web searches
❑ Mining multimedia Web content
❑ Web pages classification
❑ ...

Figure 2: Web-content mining[9]

Web content mining targets the knowledge discovery, in which the main objects are the traditional collections of text documents and, more recently, also the collections of multimedia documents such as images, videos, audios, which are embedded in or linked to the Web pages as shown in figure 2. Web content mining could be differentiated from two points of view: the agent-based approach or the database approach. The first approach aims on improving the information finding and filtering and could be placed into the following three categories [4]:

a) *Intelligent Search Agents*. These agents search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information.

b) *Information Filtering/ Categorization*. These agents use information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them.

c) *Personalized Web Agents*. These agents learn user preferences and discover Web information based on these preferences, and preferences of other users with similar interest.

The second approach aims on modelling the data on the Web into more structured form in order to apply standard database querying mechanism and data mining applications to analyze it. The two main categories are *Multilevel databases* and *Web query systems[1,15,17]*.

Following are the problems in web content mining.

a) Data/information extraction: Our focus will be on extraction of structured data from Web pages, such as products and search results. Extracting such data allows one to provide services. Two main types of techniques, machine learning and automatic extraction are covered.

b) 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. How to identify or match semantically similar data is a very important problem with many practical applications. Some existing techniques and problems are examined.

c) 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. We will introduce a few tasks and techniques to mine such sources.

d) 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.

e) 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 [1].

3.2 **Structure Mining** :

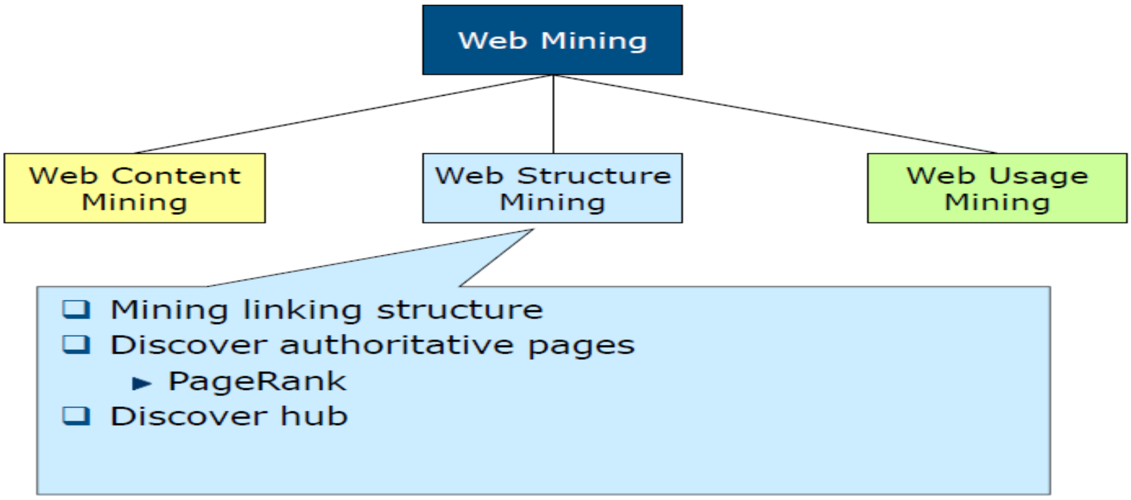

Figure 3 : Web structure mining [9]

Web structure mining tries to discover the model underlying the link structures of the web. Model is based on the topology of hyperlinks with or without description of links. Markov chain model can be used to categorize web pages and is useful to generate information such as similarity and relationship between different websites. Figure 3 shows that web structure mining focuses on the hyperlink structure of the Web[8]. The different objects are linked in some way. 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 [12]. Two algorithms that have been proposed to lead with those potential correlations: HITS [11] and PageRank [13].

3.3**Usage Mining**:

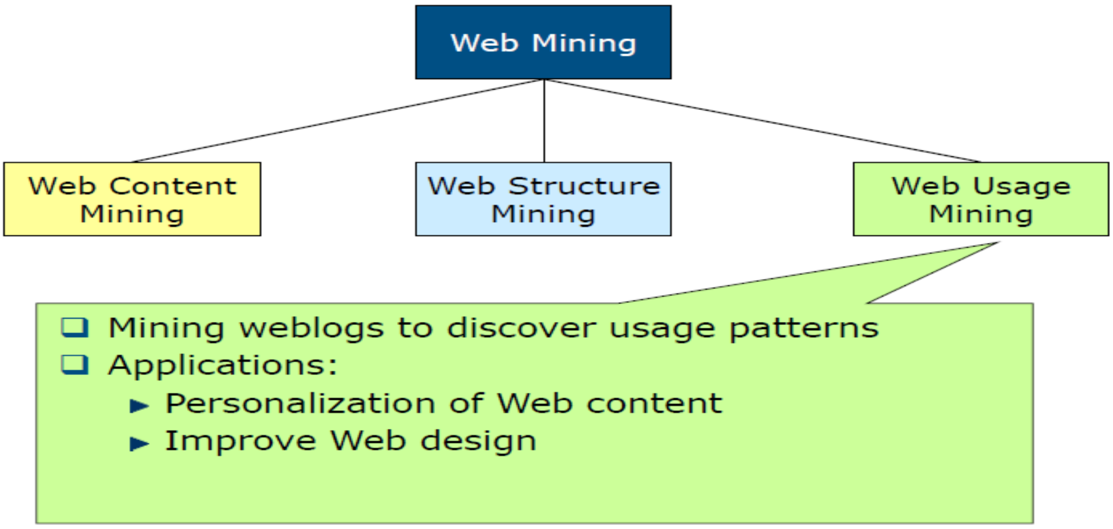

Figure 4 : Web Usage Mining [9]

Finally, mining for user navigation patterns focuses on techniques which study the user behaviour when navigating the web that is web usages mining as shown in figure 4. Web usage mining refers discovery of user access patterns from Web servers. Web usages data include data from web server access logs, proxy server logs, browser logs, user profiles, registration data, user session or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls or any other data as result of interaction [4]. Web Usage Mining Web server's record and accumulate data about user interactions whenever requests for resources are received. Analyzing the web access logs of different web sites can help understand the user behavior and the web structure, thereby improving the design of this colossal collection of resources. There are two main tendencies in Web Usage Mining driven by the applications of the discoveries: General Access Pattern Tracking and Customized Usage Tracking.[7] The general access pattern tracking analyzes the web logs to understand access patterns and trends. These analyses can shed light on better structure and grouping of resource providers. Customized usage tracking analyzes individual trends. Its purpose is to customize web sites to users. The information displayed the depth of the site structure and the format of the resources can all be dynamically customized for each user over time based on their access patterns.

4. WEB DATA MINING

With more than two billion pages created by millions of Web page authors and organizations, the World Wide Web is a tremendously rich knowledge base. The knowledge comes not only from the content of the pages themselves, but also from the unique characteristics of the Web, such as its hyperlink structure and its diversity of content and languages. Analysis of these characteristics often reveals interesting patterns and new knowledge. Such knowledge can be used to improve users' efficiency and effectiveness in searching for information on the Web, and also for applications unrelated to the Web, such as support for decision making or business management. The Web's size and its unstructured and dynamic content, as well as its multilingual nature, make the extraction of useful knowledge a challenging research problem. Furthermore, the Web generates a large amount of data in other formats that contain valuable information. For example, Web server logs' information about user access patterns can be used for information personalization or improving Web page design. Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the Worldwide Web. Two different approaches were taken in initially defining web mining [10]:

i. Process-centric View – Web mining as a sequence of tasks

ii. Data-centric view – web mining as a web data that was being used in the mining process.

There is no agreed definition of Web Data Mining but one simple definition is:

"*Web Data Mining is the application of data mining techniques to find interesting and potentially useful knowledge from web data. It is normally expected that either the hyperlink structure of the web or the web log data or both have been used in the mining process.*" [11]

The Web offers various opportunities and challenges to data mining[6]:

- The amount of information on the Web is huge, and easily accessible.

- The coverage of Web information is very wide and diverse. One can find information about almost anything.
- Information/data of almost all types exist on the Web, e.g., structured tables, texts, multimedia data, etc.
- Much of the Web information is semi-structured due to the nested structure of HTML code.
- Much of the Web information is linked. There are hyperlinks among pages within a site, and across different sites.
- Much of the Web information is redundant. The same piece of information or its variants may appear in many pages.
- The Web is noisy. A Web page typically contains a mixture of many kinds of information, e.g., main contents, advertisements, navigation panels, copyright notices,
- The Web consists of surface Web and deep Web.
  In surface Web pages that can be browsed using a browser.
  In deep Web databases that can only be accessed through parameterized query interfaces.
- The Web is also about services. Many Web sites and pages enable people to perform operations with input parameters, i.e., they provide services.
- The Web is dynamic. Information on the Web changes constantly. Keeping up with the changes and monitoring the changes are important issues.
- Above all, the Web is a virtual society. It is not only about data, information and services, but also about interactions among people, organizations and automatic systems.

The important data mining techniques applied in the web domain include Association Rule, Sequential pattern discovery, clustering, path analysis, classification and outlier discovery [4].

i. Association Rule Mining:

Predict the association and correlation among set of items “where the presence of one set of items in a transaction implies (with a certain degree of confidence) the presence of other items [2].

1) Discovers the correlations between pages that are most often referenced together in a single server session/user session.
2) Provide the information:

a) What are the set of pages frequently accessed together by web users?

b) What page will be fetched next?

c) What are paths frequently accessed by web users?

3) Associations and correlations:

a) Page association from usage data – user sessions, user transactions

b) Page associations from content data – similarity based on content analysis.

c) Page associations based on structure – link connectivity between pages.

Advantages:

a) Guide for web site restructuring – by adding links that interconnect pages often viewed together.

b) Improve the system performance by pre-fetching web data.

ii. Sequential pattern discovery:

It is applied to web access server transaction logs. The purpose is to discover sequential patterns that indicate user visit patterns over a certain period. That is, the order in which URLs tend to be accessed.

Advantages:

a) Useful user trends can be discovered

b) Predictions concerning visit pattern can be made

c) To improve website navigation

d) Personalize advertisements

e) Dynamically reorganize link structure and adopt web site contents to individual client requirements or to provide clients with automatic recommendations that best suit customer profiles.

iii. Clustering:

It groups together items (users, pages, etc.,) that have similar characteristics.

a) Page clusters:

It consists of groups of pages that seem to be conceptually related according to users' perception.

b) User Cluster:

It consists of groups or users that seem to be behave similarly when navigating through a web site.

iv. Classification:

It maps a data item into one of several predetermined classes. Example: describing each user's category using profiles. Classification algorithms are decision tree, naïve Bayesian classifier, neural networks.

v. Path Analysis:

A technique that involves the generation of some form of graph that represents relation[s] defined on web pages. This can be the physical layout of a web site in which the web pages are nodes and links between these pages are directed edges. Most graphs are involved in determining frequent traversal patterns more frequently visited paths in a web site.

To use data mining on our web site, we have to establish and record visitor and item characteristics, and visitor interactions.

Visitor characteristics include:

i. Demographics – are tangible attributes such as home address, income, property, etc.

ii. Psychographics – are personality types such as early technology interest, buying tendencies.

iii. Techno graphics – are attributes of visitor's system, such as operating system, browser, and modem speed.

Item characteristics include:

i. Web content information – media type, content category, URL.

ii. Product information - product category, color, size, price

Visitor interactions include:

i. Visitor-item interactions include purchase history, advertising history, and preference information…

ii. Visitor-site statistics are per session characteristics, such as total time, pages viewed, and so on.

We have a lot of information about web visitors and content, but we probably are not making the best use of it. The existing OLAP systems can report only on directly observed and easily correlated information. They rely on users to discover patterns and decide what to do with them. The information is even too complex for humans to discover these patterns using an OLAP system. To solve these problems, data mining techniques are utilized.

5. CONCLUSIONS

As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract all manner of useful knowledge from it. The past five years have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it Web data mining is a fast rising research area today [17]. As the web data and its usage will rise in future. It will prolong to generate more content, structure and usage data. So the importance of web data continues increasing. Web data is mainly semi-structured and unstructured. 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. Most of the knowledge represented in HTML Web documents, there are numerous other file formats that are publicly accessible on the Internet. Also, if both the actual Web Documents and corresponding Back Link Documents were mainly composed of multimedia information (e.g. graphics, audio, etc.), SVD will not be particularly effective in revealing more textual information. It would be worthwhile to research new techniques to include these file formats and multimedia information for knowledge representation. Web Data Mining is perhaps still in its infancy and much research is being carried out in the area.

REFERENCES

[1] A. A. Barfourosh, H.R. Motahary Nezhad, M. L. Anderson, D. Perlis, “Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition”, 2002.

[2] About Web, available:http://www.technicalsymposium.com/web_mining_notes.html

[3] B. Singh, H.K. Singh, “Web data Mining Research”, in the Proceedings of 2010 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-10, Dec. 2010

[4] Cooley, R.; Mobasher, B.; Srivastava, J.; “Web mining: information and pattern discovery on the World Wide Web”. *In Proceedings of Ninth IEEE* International Conference. pp. 558 – 567, 3-8 Nov. 1997.

[5] G. Srivastava, K. Sharma, V. Kumar," Web Mining: Today and Tomorrow", in the Proceedings of 2011 3rd International Conference on Electronics Computer Technology (ICECT), pp.399-403, April 2011

[6] Han, J., Kamber, M. Kamber. “Data mining: concepts and techniques”. Morgan Kaufmann Publishers, 2000..

[7] J. Srivastava, R. Cooley, M. Deshpande, Pag-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data" in proceedings of *ACM SIGKDD Explorations Newsletter*, Vol. 1 Issue 2,January 2000.

[8]J. Borges and M. Levene. "Mining Association Rules in Hypertext Databases". In *Knowledge Discovery and Data Mining*, pp 149–153, 1998

[9] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques", 3rd edition,USA.

[10] J. Srivastava, "Web Mining: Accomplishments & Future Directions", National Science Foundation Workshop on Next Generation Data Mining NGDM02, pp. 1-148, 2002

[11] Kleinberg, J.M., "Authoritative sources in a hyperlinked environment", *In Proceedings of ACM-SIAM Symposium on Discrete Algorithms,* pp. 668-677 – 1998.

[12] L. Getoor, "Link Mining: A New Data Mining Challenge". *SIGKDD Explorations*, vol. 4, issue 2, 2003.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank citation ranking: Bring order to the web. *Technical report*, Stanford University, 1998

[14] O. Etzioni, "The world wide Web: Quagmire or gold mine." Communications of the ACM, Vol. 39 No. 11, pp. 65-68, Nov. 1996.

[15] R. Kosala, H. Blockeel "Web mining research: A survey," ACM SIGKDD Explorations, Vol. 2 No. 1, pp. 1-15, June 2000

[16] R.S. Segall , Q. Zhang ,"Teaching web mining in the classroom: with an overview of web usage mining" in proceedings of 2008 SWDSI meeting.

[17] W. Jicheng, H. Yuan, W. Gangshan, Z. Fuyan." Web mining: knowledge discovery on the Web. Systems, Man, and Cybernetics" in Proceedings *of 1999* IEEE International Conference- on Vol. 2, pp.137 - 141 ,12-15 Oct. 1999.

[18] Web mining definition, available: http://en.wikipedia.org/wiki/Web_mining.