Volume 3, Issue 3, March 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Web Mining: An Introduction

Monika Yadav
M Tech Student
Department Of Computer Science and Applications
Kurukshetra University, Kurukshetra
Haryana, India

Mr. Pradeep Mittal
Assistant Professor
Department Of Computer Science and Applications
Kurukshetra University, Kurukshetra
Haryana, India

Abstract* --From its very beginning, the potential of extracting valuable knowledge from the Web has been quite evident. Web mining – i.e. the application of data mining techniques to extract knowledge from Web content, structure, and usage is the collection of technologies to fulfill this potential. Web mining is the application of data mining techniques to extract knowledge from Web data, where at least one of structure (hyperlink) or usage (Web log) data is used in the mining process (with or without other types of Web data). Interest in Web mining has grown rapidly in its short existence, both in the research and practitioner communities. The present paper deals with a preliminary discussion of WEB mining, few key computer science contributions in the field of web mining , the prominent successful applications and outlines some promising areas of future research.

Keywords*-- Web mining, Web usage mining, Web structure mining, Web content mining.

I. INTRODUCTION

Web mining - is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, us-age logs of web sites, etc.

Internet has became an indispensable part of our lives now a days so the techniques which are helpful in extracting data present on the web is an interesting area of research. These techniques helps to extract knowledge from Web data, in which at least one of structure or usage (Web log) data is used in the mining process (with or without other types of Web). According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining.

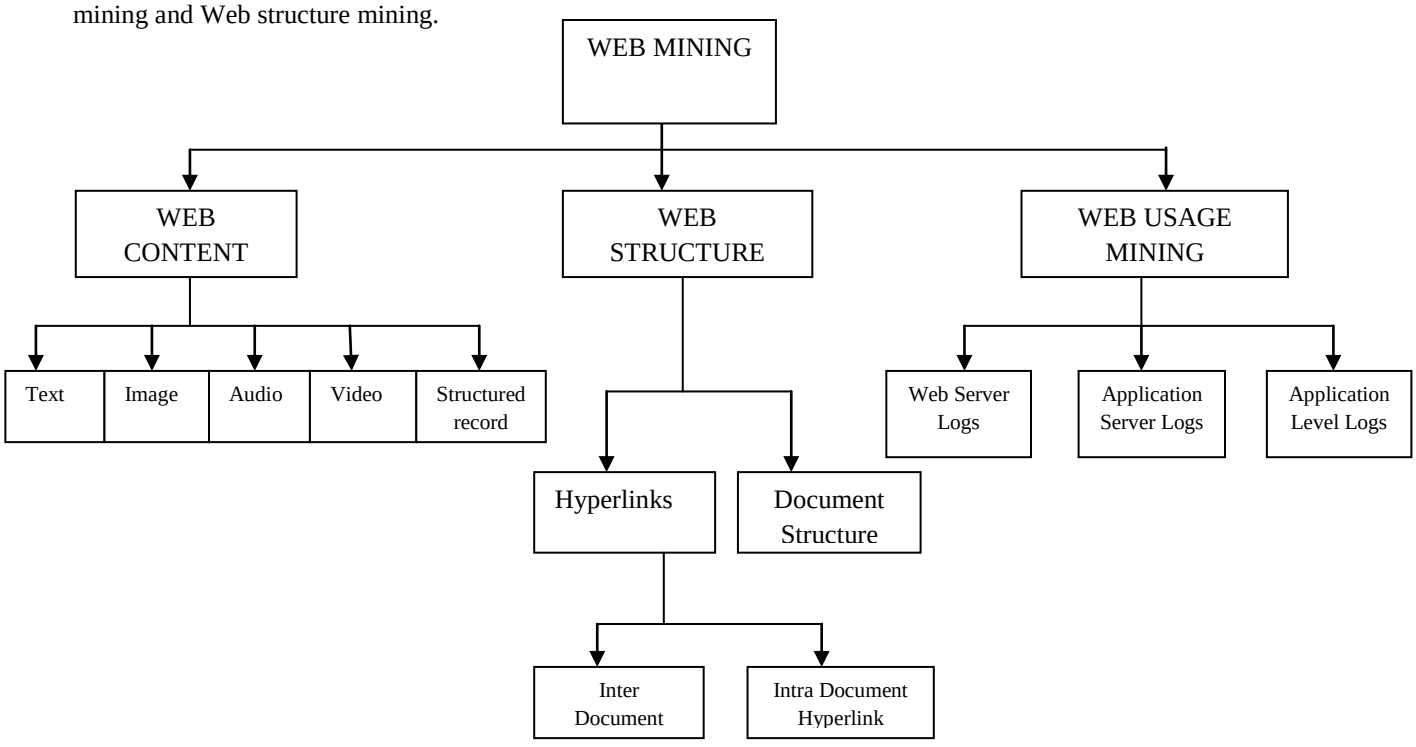

Fig 1. Web Mining Taxonomy

With the explosive growth of information sources available on the World Wide Web and the rapidly increasing pace of adoption to Internet commerce, the Internet has evolved into a gold mine that contains or dynamically generates information that is beneficial to E-businesses[1]. A web site is the most direct link a company has to its current and potential customers. The companies can study visitor's activities through web analysis, and find the patterns in the visitor's behavior. These rich results yielded by web analysis, when coupled with company data warehouses, offer great opportunities for the near future.

II. TYPES OF WEB MINING

There are three types of web mining which are discussed below:

A. Web Usage Mining

Web usage mining is the process of extracting useful information from server logs i.e. users history. Web usage mining is the process of finding out what users are looking for on Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. This technology is basically concentrated upon the use of the web technologies which could help for betterment.

Web usage mining process involves the log time of pages. The world's largest portal like yahoo, msn etc., needs a lot of insights from the behaviour of their users' web visits. Without this usage reports, it will be difficult to structure their monetization efforts. Usage mining has direct impact on businesses[3].

This is the activity that involves the automatic discovery of user access patterns from one or more Web servers. As more organizations rely on the Internet and the World Wide Web to conduct business, the traditional strategies and techniques for market analysis need to be revisited in this context. Organizations often generate and collect large volumes of data in their daily operations. Most of this information is usually generated automatically by Web servers and collected in server access logs. Other sources of user information include referrer logs which contains information about the referring pages for each page reference, and user registration or survey data gathered via tools such as CGI scripts[6].

Analyzing such data can help these organizations to determine the life time value of customers, cross marketing strategies across products, and effectiveness of promotional campaigns, among other things. Analysis of server access logs and user registration data can also provide valuable information on how to better structure a Web site in order to create a more effective presence for the organization. In organizations using intranet technologies, such analysis can shed light on more effective management of workgroup communication and organizational infrastructure. Finally, for organizations that sell advertising on the World Wide Web, analyzing user access patterns helps in targeting ads to specific groups of users[4].

- **Web Server Data**

User logs are collected by the web server and typically include IP address, page reference and access time.

- **Application Server Data**

Commercial application servers such as Weblogic, StoryServer, have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

- **Application Level Data**

New kinds of events can be defined in an application, and logging can be turned on for them — generating histories of these events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the above the categories[7].

A. Web Structure Mining

Web structure mining, one of three categories of web mining for data, is a tool used to identify the relationship between Web pages linked by information or direct link connection. This structure data is discoverable by the provision of web structure schema through database techniques for Web pages. This connection allows a search engine to pull data relating to a search query directly to the linking Web page from the Web site the content rests upon. This completion takes place through use of spiders scanning the Web sites, retrieving the home page, then, linking the information through reference links to bring forth the specific page containing the desired information.[7]

Structure mining uses minimize two main problems of the World Wide Web due to its vast amount of information. The first of these problems is irrelevant search results. Relevance of search information become misconstrued due to the problem that search engines often only allow for low precision criteria. The second of these problems is the inability to index the vast amount if information provided on the Web. This causes a low amount of recall with content mining. This minimization comes in part with the function of discovering the model underlying the Web hyperlink structure provided by Web structure mining.

The main purpose for structure mining is to extract previously unknown relationships between Web pages. This structure data mining provides use for a business to link the information of its own Web site to enable navigation and cluster information into site maps. This allows its users the ability to access the desired information through keyword association and content mining. Hyperlink hierarchy is also determined to path the related information within the sites to the relationship of competitor links and connection through search engines and third party co-links[5]. This enables clustering of connected Web pages to establish the relationship of these pages.

On the WWW, the use of structure mining enables the determination of similar structure of Web pages by clustering through the identification of underlying structure. This information can be used to project the similarities of web content. The known similarities then provide ability to maintain or improve the information of a site to enable access of web spiders in a higher ratio. The larger the amount of Web crawlers, the more beneficial to the site because of related content to searches.

In the business world, structure mining can be quite useful in determining the connection between two or more business Web sites. The determined connection brings forth a useful tool for mapping competing companies through third party links such as resellers and customers. This cluster map allows for the content of the business pages placing upon the search engine results through connection of keywords and co-links throughout the relationship of the Web pages. This determined information will provide the proper path through structure mining to improve navigation of these pages through their relationships and link hierarchy of the Web sites.

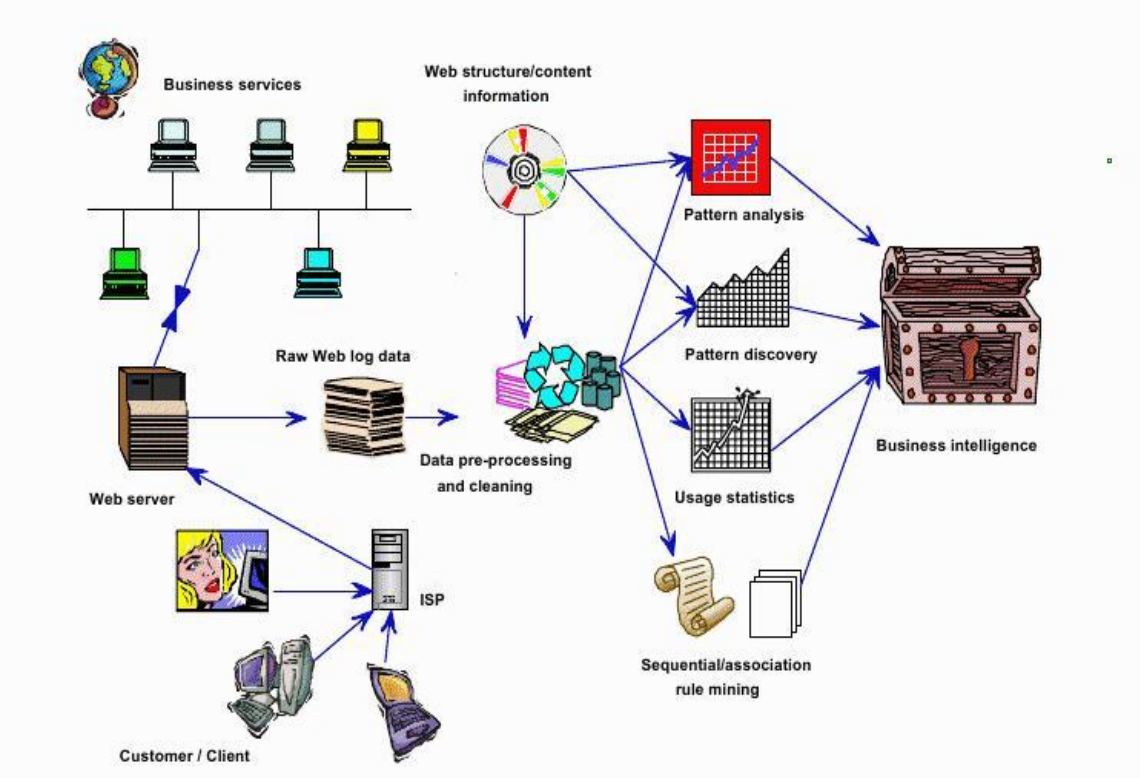

Fig 2. Web Usage Mining

With improved navigation of Web pages on business Web sites, connecting the requested information to a search engine becomes more effective. This stronger connection allows generating traffic to a business site to provide results that are more productive. The more links provided within the relationship of the web pages enable the navigation to yield the link hierarchy allowing navigation ease. This improved navigation attracts the spiders to the correct locations providing the requested information, proving more beneficial in clicks to a particular site.

Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided a into two kinds:

1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.
2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage[9].

B. Web Content Mining

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page contents. Content mining is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query. This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query. With the massive amount of information that is available on the World Wide Web, content mining provides the results lists to search engines in order of highest relevance to the keywords in the query. [8]

The web content mining is differentiated from two different points of view : Information Retrieval View and Database View. R. Kosala et al. summarized the research works done for unstructured data and semi-structured data from information

retrieval view. It shows that most of the researches use bag of words, which is based on the statistics about single words in isolation, to represent unstructured text and take single word found in the training corpus as features. For the semi-structured data, all the works utilize the HTML structures inside the documents and some utilized the hyperlink structure between the documents for document representation. As for the database view, in order to have the better information management and querying on the web, the mining always tries to infer the structure of the web site to transform a web site to become a database. This type of mining uses the ideas and principles of data mining and knowledge discovery to screen more specific data. The use of the Web as a provider of information is unfortunately more complex than working with static databases. Because of its very dynamic nature and its vast number of documents, there is a need for new solutions that are not depending on accessing the complete data on the outset[10]. Another important aspect is the presentation of query results. Due to its enormous size, a web query can retrieve thousands of resulting webpages. Thus meaningful methods for presenting these large results are necessary to help a user to select the most interesting content.

The below figure shows 3 types of Web Content Mining Techniques:-

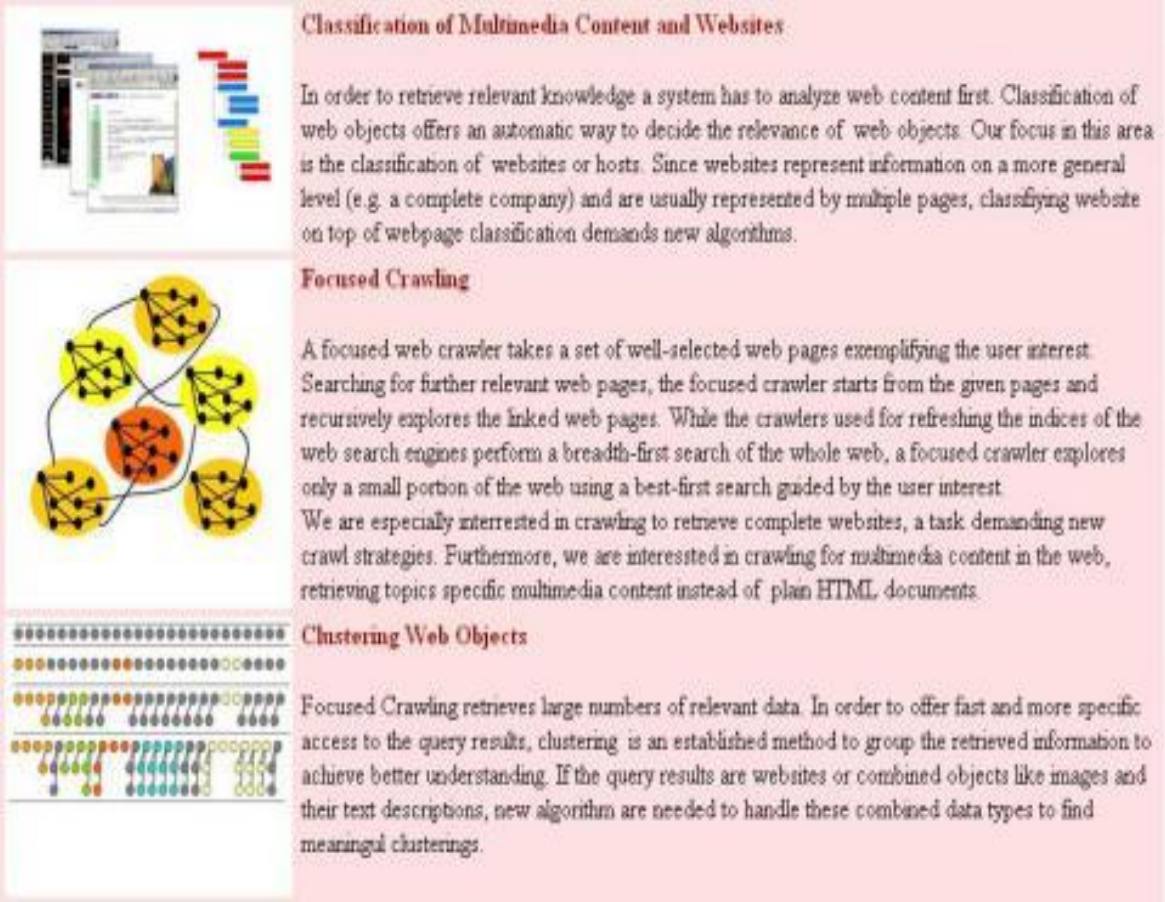

Fig 3. Web Content Mining

III.Web Mining Applications

Web mining extends analysis much further by combining other corporate information with Web traffic data. Practical applications of Web mining technology are abundant, and are by no means the limit to this technology. Web mining tools can be extended and programmed to answer almost any question.

It can be applied in following areas:

1. Web mining can provide companies managerial insight into visitor profiles, which help top management take strategic actions accordingly[2].
2. The company can obtain some subjective measurements through Web Mining on the effectiveness of their marketing campaign or marketing research, which will help the business to improve and align their marketing strategies timely.
3. In the business world, structure mining can be quite useful in determining the connection between two or more business Web sites.
4. This allows accounting, customer profile, inventory, and demographic information to be correlated with Web browsing

5. The company can identify the strength and weakness of its web marketing campaign through Web Mining, and then make strategic adjustments, obtain the feedback from Web Mining again to see the improvement.
6. Search engine Google provides advanced and efficient searching capabilities[11].

IV.CONCLUSION

It is a revolution that the Internet has grown from a simple search tool to a gold mine. Companies find a new and better way to do business: E-commerce through the Internet. However, E-business cannot just build a web site and then sit back and reap the benefits, which, in most cases, is fruitless. Companies have to implement Web mining systems to understand their customers' profiles, and to identify their own strength and weakness of their E-marketing efforts on the web through continuous improvements. Internet is a gold mine, but only for those companies who realize the importance of Web mining and adopt a Web mining strategy now.

References:

[1] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava , *"Web Mining: information and Pattern Discovery on the WWW"*

[2] Mary Garvin , *"Data Mining and the Web: What They Can Do Together*"

[3] B. Masand, M. Spiliopoulou, J. Srivastava, O. Zaiane, ed. Proceedings of "*WebKDD2002 –Web Mining for Usage Patterns and User Profiles*", Edmonton, CA, 2002.

[4] M. Spiliopoulou, "Data Mining for the Web", *Proceedings of the Symposium on Principles of Knowledge Discovery in Databases* (PKDD), 1999.

[5] J. Srivastava, B. Mobasher, Panel discussion on "Web Mining: Hype or Reality?" at *the 9th IEEE International Conference on Tools With Artificial Intelligence* (ICTAI '97), Newport Beach, CA, 1997.

[6] R. Kohavi, "*Mining E-Commerce Data*: The Good, the Bad, the Ugly", Invited Industrial presentation at the ACM SIGKDD Conference, San Francisco, CA, 2001.

[7] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, *"Web Mining Concepts ,Applications and Research Directions"*

[8] R. Kosala, H. Blockeel, "*Web Mining Research: A Survey*", in SIGKDD Explorations 2(1), ACM, July 2000.

[9] J. Srivastava, R. Cooley, M.Deshpande, P-N. Tan. " *Web Usage Mining: Discovery and Applications of usage patterns from Web Data", SIGKDD Explorations, Vol1, Issue 2, 2000*

[10] T. Berners-Lee, R. Cailliau, A. Loutonen, H. Nielsen, and A. Secret. The World- Wide Web. *Communications of the ACM*, 37(8):76- 82, 1994.

[11] R. Cooley, B. Mobasher, J. Srivastava, "*Data Preparation for Mining World Wide Web Browsing Patterns*", Knowledge and Information Systems, 1(1), 1999.

