A Review on Web Mining

Sushma Gupta[1] Rakesh Patel[2] Chanda Patel[3]
[1,3]B.E. Student [2]Lecturer
[1,2,3]Department of Information Technology
[1,2,3]Kirodimal Institute of Technology, Raigarh(C.G.),India

Abstract— Nowadays, the Web has become one of the most widespread platforms for information change and retrieval. As it becomes easier to publish documents, as the number of users, and thus publishers, increases and as the number of documents grows, searching for information is turning into a cumbersome and time-consuming operation. The Web mining research relates to several research communities such as Database, Information Retrieval and Artificial Intelligence. 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).

Key words: Web Mining, data mining, Web data

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).Web mining should be decomposed into these subtasks:-

1) Resource finding: The task of retrieving the intended information from Web.
2) Information Extraction: Automatically selecting and pre-processing specific information from the retrieved Web resources.
3) Generalization: Automatically discovers general patters at the both individual Web sites and across multiple sites d.
4) Analysis: Analyzing the mined pattern.

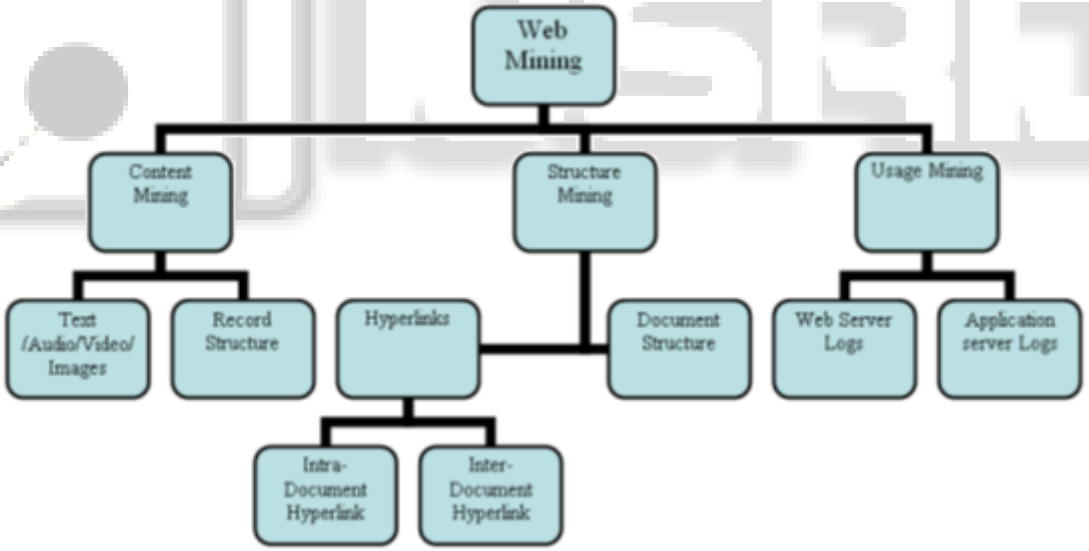

Fig. 1: Web Mining Taxonomy

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.

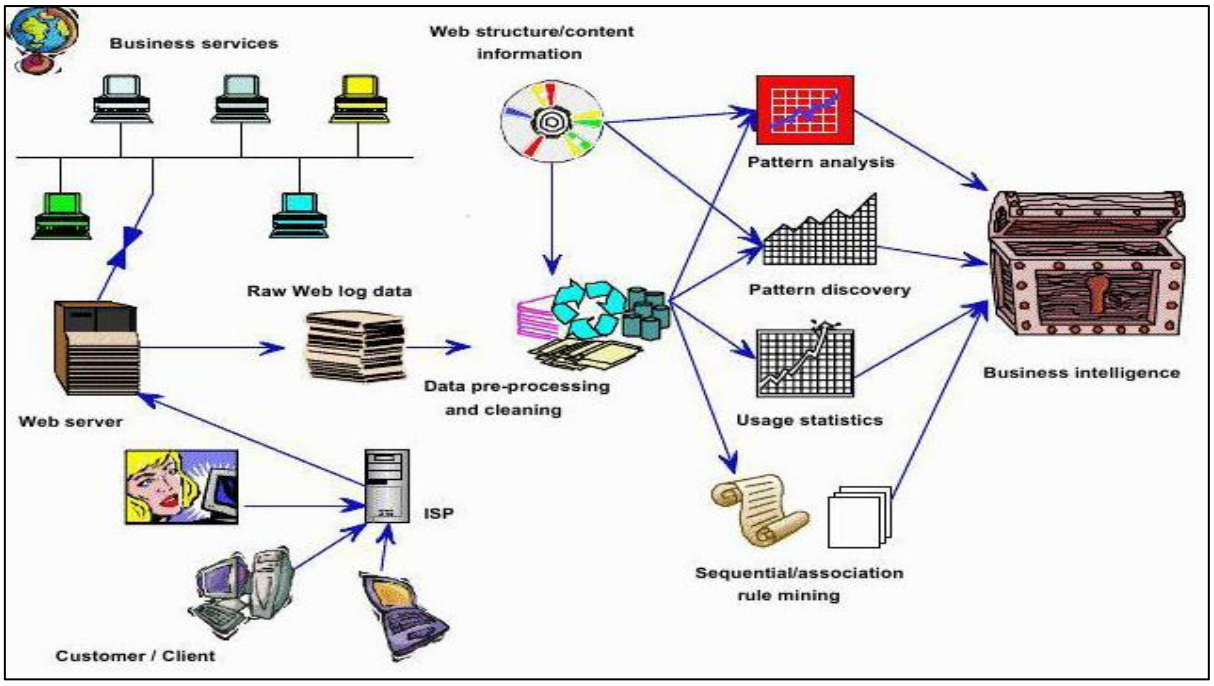

Fig. 2: Web Usage Mining

1) Web Server Data

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

2) 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.

3) 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].

B. 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, and linking the information through reference links to bring forth the specific page containing the desired information

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. This enables clustering of connected Web pages to establish the relationship of these pages.

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

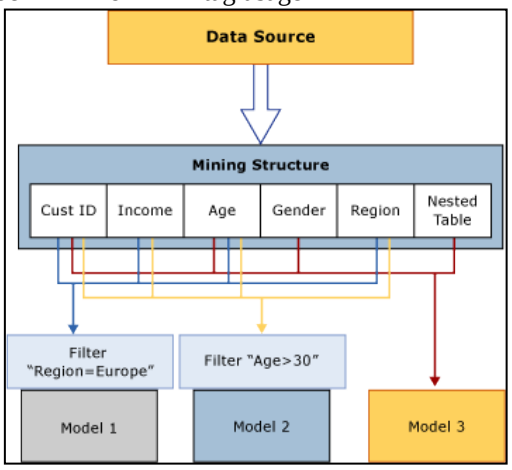

Fig. 2: Web Structure Mining

C. 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. The web content mining is differentiated from two different points of view : Information Retrieval View and Database View. The below figure shows 3 types of Web Content Mining Techniques:-

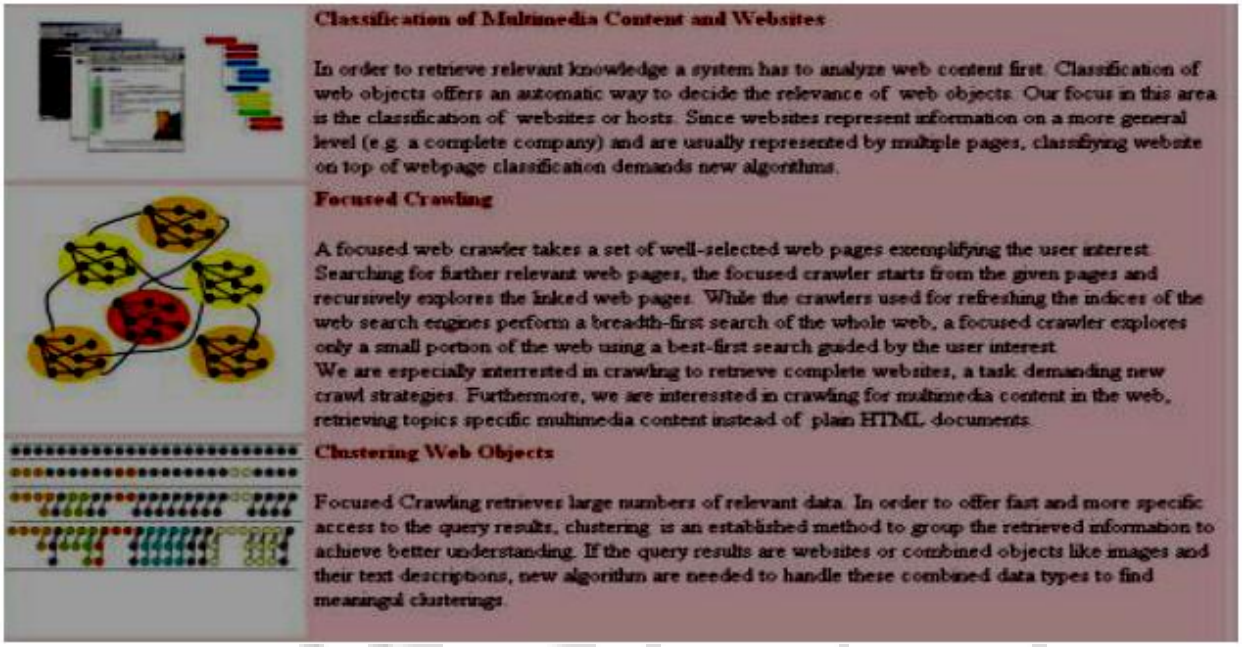

Fig. 3: Web Content Mining

III. Web Mining Application

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) Search engine Google provides advanced and efficient searching capabilities

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.[1] 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.[2] The web continues to increase in size and complexity with time hence making it difficult to extract relevant information.[3]

References

[1] Web Mining: An Introduction Monika Yadav Mr. Pradeep Mittal M Tech Student Assistant Professor Department Of Computer Science and Applications Department Of Computer Science and Applications Kurukshetra University, Kurukshetra Kurukshetra University, Kurukshetra Haryana, India Haryana, India

[2] Web Content Mining: Its Techniques and Uses Govind Murari Upadhyay, Kanika Dhingra (Assistant Professor) IITM, Janakpuri, New Delhi, India
[3] Web Mining and Knowledge Discovery of Usage Patterns
[4] Overview Of Web Content Mining Tools 1Abdelhakim Herrouz, 2Chabane Khentout Mahieddine Djoudi Department of Computer Science, University Kasdi Merbah of Ouargla, Algeria Laboratoire des Réseaux et des Systèmes Distribués, University Ferhat Abbas of Sétif, Algeria Department XLIM-SIC UMR CNRS 7252 & TechNE Research Group, University of Poitiers, Teleport 2, Boulevard Marie et Pierre Curie, B.P 30179, 86960 Futuroscope Cedex, France
[5] 5.Herrouz, A., Khentout, C., Djoudi, M. Overview of Visualization Tools for Web Browser History Data, IJCSI International Journal of Computer Science Issues, Vol.9, Issue 6, No3, November 2012, pp. 92-98, (2012).
[6] Lieu, B., Web Information Systems Engenineering-WISE 2005, 6th International Conference on Web Information Systems Engineering, New York, NY, USA, November 20-22, 2005, Proceedings. Volume 3806 of Lecture Notes in Computer Science, pages 763, Springer (2005).
[7] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava , "Web Mining: information and Pattern Discovery on the WWW"
[8] Mary Garvin , "Data Mining and the Web: What They Can Do Together"
[9] Faustina Johnson and Santosh Kumar Gupta Web Content Mining Techniques: A Survey. International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012
[10] B. Berendt. Web usage mining, site semantics, and the support of navigation

