Extracting Data Through Webmining

Mrs.Bhanu Bhardwaj
Asst Proff
DCE G.Noida

Abstract— This Paper comprises of upcoming evolutionary field of "Web Mining". This paper describes the extraction of useful information from internet. This paper present a clear view of what is Web Mining and How is it done. In this paper we encompass various Pros and Cons of Web Mining and also represent the similarities and differences between Data Mining and Web Mining. Here we describe some of the recently used Tools used in Web Mining.

Keywords- Web Mining, Data Extraction, Web Usage Mining, Web Content Mining, Web Structure Mining.

I. INTRODUCTION

Information has been a important part of humans but with evolution in technology handling of information changed from analog to digital, now we use huge computation to manage our Information and internet has now became a medium for collecting information. Upcoming ERA comprise totally of Internet, expansion of World Wide Web has increased the storing capacity of information online.

World Wide Web consist of different type of Data and storing huge heterogeneous Data online, reverts a serious issue of Extracting that information at the required time [1]. To get required information easily, efficiently and correctly we need a strong concept that easily mine the required information within fraction of seconds. This extraction of Information on Internet or World Wide Web is called "Web Mining".

We Describe Web mining as a technique of Mining Data on World Wide Web. In my view World Wide Web is a Mine of Huge Information's and Web Mining is a technique or rather an approach to extract useful information from that mine with ease of efficiency [2]. By Ease of efficiency we mean **"**Performing Extraction with minimal usage of Resources.**"**

II. DRAWBACKS EXPERIENCED IN EXITING SYSTEM

i. Explosive Growth of internet has rendered User to get effective information.*Maintaining the Integrity of the Specifications.*

ii. Users experience a heavy time loss.

iii. Knowledge discovery consumes a lot of System Resources.

iv. Caching Schemes fails in certain Conditions.

v. Network traffic gets over congested due to pre-fetching techniques.

III. WEB MINING

Data Mining can be referred as extraction of valuable knowledge patterns from huge bulks database [4] [3]. Data mining combines techniques including statistical analysis, visualization, induction, and neural networks to explore large amounts of data and discover relationships and patterns that shed light on business problems. In turn, companies can use these findings for more profitable, proactive decision making and competitive advantage [4].

The main aim of Data Mining is to extract knowledge from human understandable data set and presenting that knowledge in a well designed and user friendly GUI format so as to make knowledge more understandable and more desirable.

Web Mining can be defined as an "Application part of Data mining which is used for discovering various patterns on World Wide Web."

Web Mining is completely based on knowledge discovery from web. World Wide Web Consist of a huge amount of information in various forms such as Text, Graphics, various tags etc. so we need a powerful and efficient technique to extract useful and important information from the huge bulk of data present [2]. Web mining is based on data mining technique by using data mining technique discover the hidden data in web log. Thus, web mining, though considered to be a particular application of data mining, warrants a separate field of research.

IV. WEB MINING CATEGORIES

A. *Web Usage Mining*

It is the process of extracting useful information through server log file.To explains Web Usage Mining we would like to present a basic example: In any organization if we want to know which web site is most widely used then we can check for the server log files. It is just a process of finding what users a looking on World Wide Web. It includes http log files, app server log files, cookies, Meta data etc [5]. These log files are automatically generated when any user interacts with the server. Phases in Web Usage Mining:

- **Preprocessing** It is a process of preparing data so that it can be used for Pattern Discovery and Pattern

- **Analysis**. It Includes Cleaning of Server Log files accompanied Data Integration and Data Transformation.

- **Pattern Discovery** After the Data is Preprocessed, the this data is utilized for discovering Homogeneous Pattern.[8]

- **Pattern Analysis** once the patterns are discovered then these patterns is evaluated and analysis is performed on these patterns.

B. *Web Content Mining*

It is the process of extraction of Useful information from the Content of a web page. This content includes various text and images, audio, video, structures documents such as tables [3]. It is a process of mining, extraction, integration of useful information from web page content [7].

C. *Web Structure Mining*

It makes the use of Graph theory to analyze the node and connections between various structures of a website [2]. In this various web pages are represented as a node and hyperlinks are represented as an edge [3]. In web structure mining the mining is performed in two steps:

i. Mining the document structure to get a analyzed tree like structure described by HTML or XML tag usage.

ii. Extraction of information through hyperlinks.

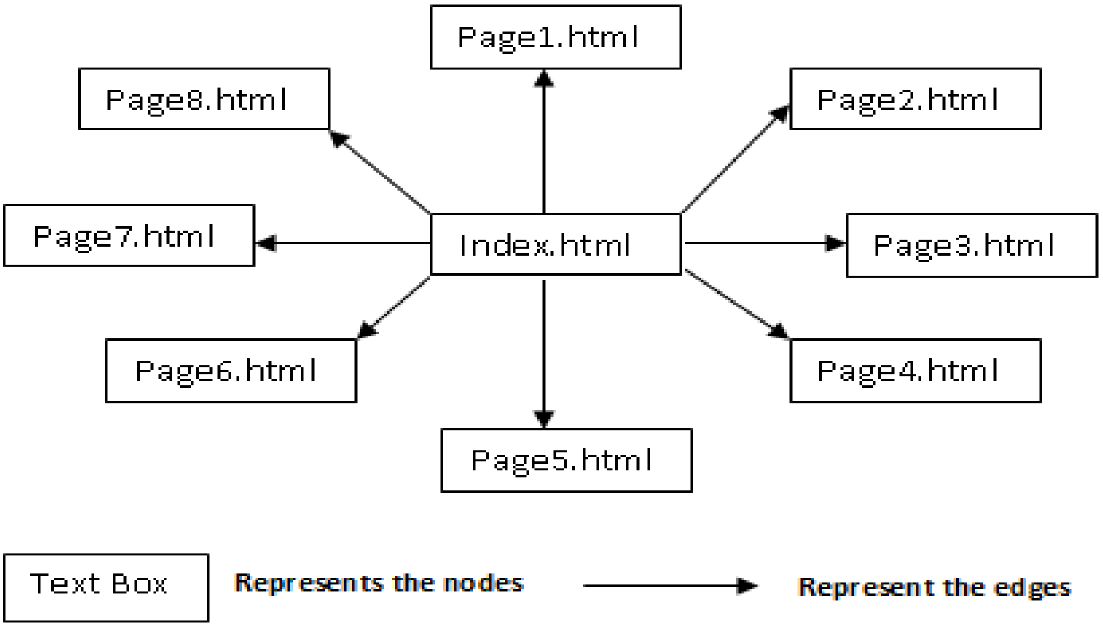

Figure No 2.

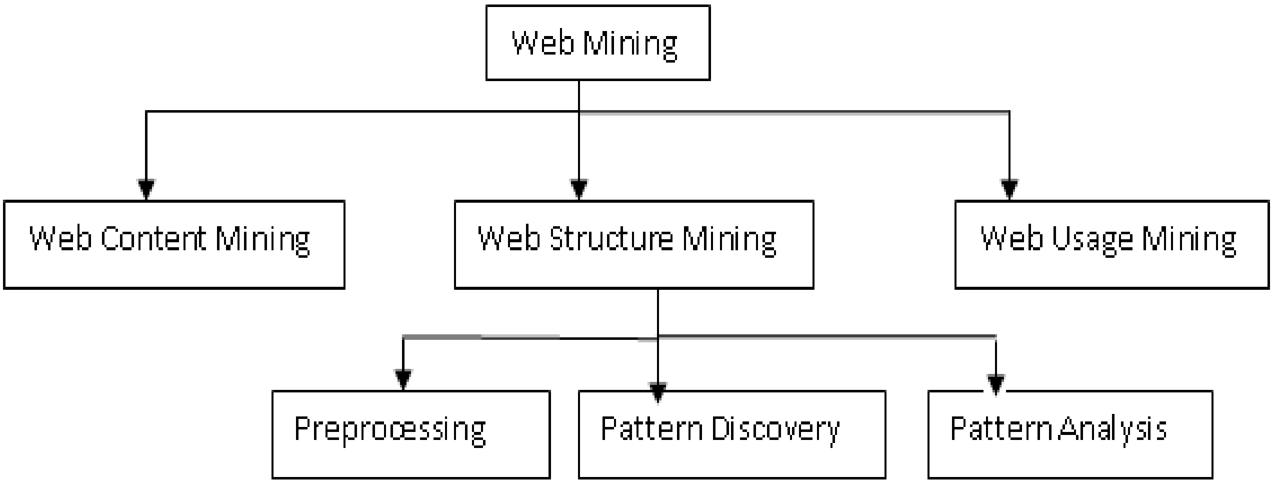

Figure No 3.(Components Web Mining)

V. STRUCTURE OF WEB MINING

Structure of Web Mining represents actually How Web Mining take place. Patterns are evaluated by three techniques of Web Mining i.e Web Content Mining, Web Usage Mining, Web Structure Mining. These techniques evaluate the needed patterns and then these patterns are analyzed to get a user desired output. Desired output is feed into the user understandable GUI [6].

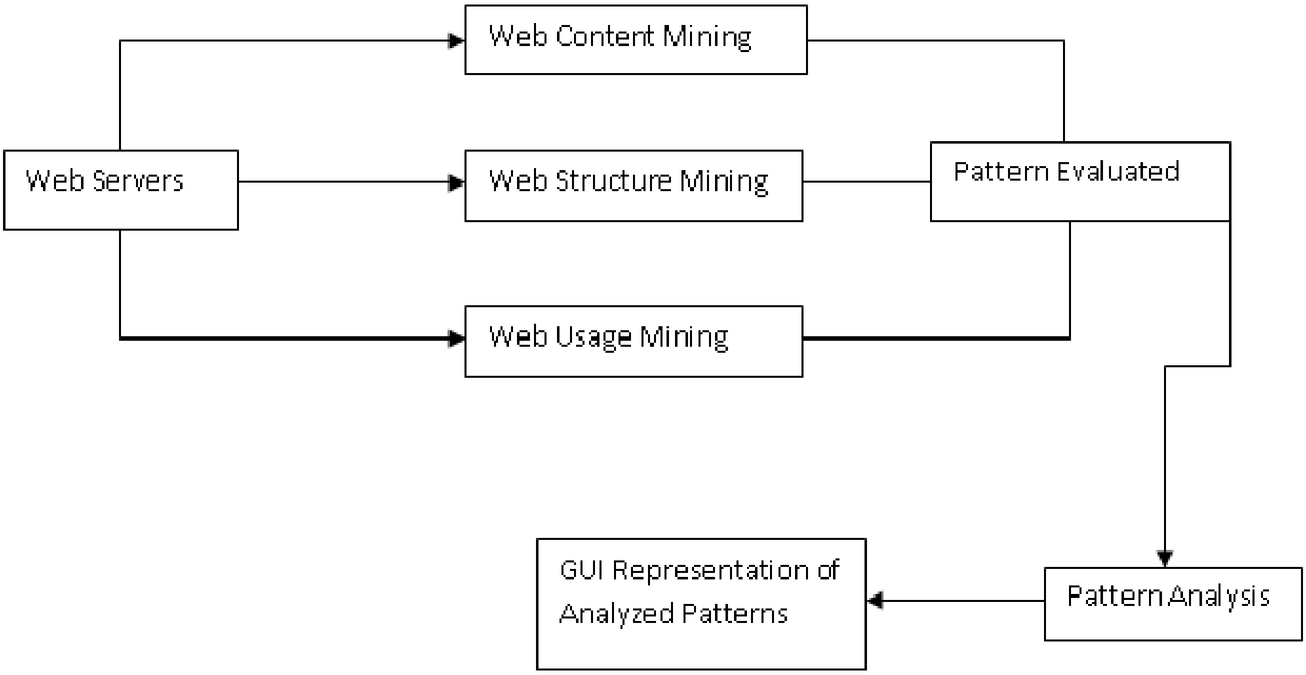

Figure No 4.(Stucture of Web mining)

VI. Pros and Cons of Web Mining

Pros

- Web Mining has been a new face of e-commerce due to an ease in personalizing marketing which provides eventually higher volumes of trade.
- It also helps in countering the Cyber Terrorism.
- Web Mining is a powerful tool for Cloud Computing.

Cons

- Web Mining has been a new face of e-commerce due to an ease in personalizing marketing which provides eventually higher volumes of trade.
- It also helps in countering the Cyber Terrorism.
- Web Mining is a powerful tool for Cloud Computing.

VII. Comparison Between Data Mining and Web Mining

Table 1 Data Mining v/s Web Mining

Parameter	Data Mining	Web Mining
Time Stamp	Crawling process take heavy time due to large sizes of databases	Crawling process takes pretty less time due to small sizes of database
Confidentiality	Confidentiality is not rendered because of limited databases in corporate Sector	Confidentiality is render because there are large no. of database online

VIII. Tools For Web Mining

Some of useful tools used for Web Mining are [5].

i. **QL2 Software**, Specializes in web data harvesting and extraction using SQL-like query language (WebQL).

ii. **Screen Scraper**, Products and services for web site data extraction. Flagship product, screen-scraper, provides a GUI to define links to follow and information to extract, and works with several programming languages and platforms.

iii. **SAS Enterprise Miner**, an integrated suite which provides a user-friendly GUI front-end to the SEMMA (Sample, Explore, Modify, Model, Assess) process.

iv. **SPAD**, provides powerful exploratory analyses and data mining tools, including PCA, clustering, interactive decision trees, discriminate analyses, neural networks, text mining and more, all via user-friendly GUI.

v. **Website Parser**, A web site parser tool is a program that will allow you to gather information from many web sites and web pages throughout the Internet. The tool goes through the targeted sites and is able to grab large amounts of data, through the parameters that you have created. This data can be used in XLS, CSV, XML, and TSV files for later use. Being able to gather huge amounts of information quickly and easily is an invaluable tool for any business owner or retail site.

vi. **Web Extractor Software**, Web extractor software may be one of the smartest software tools to invest in. The cutting edge technology may be used in a variety of settings. It has been effectively utilized by law enforcement, researchers, and several businesses by extracting vital information from specific websites. Data extraction, screen scraping, and web crawling may only be a few of the features available.

vii. **Mozenda-Mozenda**, is a Software as a Service (SaaS) company that enables users of all types to easily and affordably extract and manage web data. With Mozenda, users can set up agents that routinely extract data, store data, and publish data to multiple destinations. Once information is in the Mozenda systems users can format, repurpose, and mashup the data to be used in other online/offline applications or as intelligence.

viii. **WizSoft Software**, Develops software based on mathematical algorithms, mainly for the business sector in the fields of data mining, data auditing, concept-based text search engines, knowledge management, computational linguistics, accounting and inventory management, and operations research.

IX. Conclusion

This Paper provides a complete and clear cut view of every aspects of Web Mining: Extracting Data Online. We tried to explain Web mining in a simpler way just to make the mass aware of this upcoming technology. This paper clearly depicts the future potential of Web Mining and also provides basic information about Web Mining which is essential for a beginner. The paper also consists of various Diagrams so as to help beginners in understanding the Framework of Web Mining.

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] 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.

[3] G. Shrivastava, 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

[4] 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

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

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

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