Web Mining Overview, Techniques, Tools and Applications: A Survey

Anurag Kumar[1], Ravi Kumar Singh[2]

[1] Assistant Professor, Dept. of Computer Science & Engineering, Dr. APJ Abdul Kalam UIT Jhabua, M.P., India
[2] Assistant Professor, Dept. of Computer Science & Engineering, Prestige Institute of Engineering Management & Research, Indore, M.P., India

---***---

Abstract - *Web Mining is moving the World Wide Web towards a more useful environment in which users can quickly and easily find the information they need. Large amount of text documents, multimedia files and images are available in the web and it is still increasing. Data mining is the form of extracting data's available in the internet. Web mining is a part of data mining. Web mining is used to discover and extract information from Web-related data sources such as Web documents, Web content, hyperlinks and server logs. The term Web mining has been used in three distinct ways. The first, called Web content mining is the process of information discovery from sources across the World Wide Web. The second, called Web structure mining is the process of analyzing the relationship between Web pages linked by information or direct link connection through the use of graph theory. The third, called Web usage mining is the process of extracting patterns and information from server logs to gain insight on user activity. In this paper, we are trying to give a brief idea regarding web mining concerned with its techniques, tools and applications.*

Key Words: **Web mining, Web Content Mining, Web Usage Mining, Web Structure Mining, Mining tools**

1. INTRODUCTION

The World Wide Web (WWW) is a huge resource of multiple types of information in various formats which is very useful for the analysis of business progress, which is very important now a days to stand in the competition of business. The basic idea of web mining is to assist users or site owners in finding something useful/relevant information. Data mining, often called Web mining when applied to the Internet, is the process of using data mining techniques and algorithms to extract information directly from the Web by extracting it from Web documents, Web content, hyperlinks and server logs. The main goal of Web mining is to look for useful patterns in Web data by collecting and analyzing information in order to gain insight into trends, the industry and users. Web mining is an iterative process of discovering knowledge and is proving to be a valuable strategy for understanding consumer and business activity on the Web. Web mining has two views in general. Web mining with the User-centric view allows to Discovery of documents on a subject, Discovery of semantically related documents or document segments, Extraction of relevant knowledge about a subject from multiple sources, Knowledge/information filtering. Web mining with the owner-centric view allows getting Increasing contact / conversion efficiency (Web marketing), Targeted promotion of services, products, ads; Measuring the effectiveness of site content / structure, Providing dynamic personalized services or content. In the field of Customer analysis, it includes customer profitability, modeling customer behavior and reactions, customer satisfaction etc. Web mining in this field helps us to find strategy that should be used to get number of customers with quality as discussed in [1]. It is used to understand customer behavior, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign [2, 5].Basically there are three sub categories for mining web information. These sub categories are

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

1.1 Web Content Mining

Web Content Mining is the process of mining useful information from the contents of Web pages and Web documents, which are mostly text, images and audio/video files. The mining of 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 [2]. It includes extraction of structured data from web pages, similarity and integration of data's with similar meaning, view extraction from online sources, and concept hierarchy, knowledge incorporation [1].

Some of the prominent web content mining techniques are:-

- Unstructured text mining,
- Structured mining,
- Semi structured text mining, and
- Multimedia mining.

1. Unstructured Text Data Mining:

Most of the web pages are in the form of text. Content mining requires application of data mining and text mining techniques [4]. The data mining techniques to unstructured text is known as Knowledge Discovery in Texts (KDT), or text data mining, or text mining. Some of the techniques used in text mining are

- Information Extraction,
- Topic Tracking
- Summarization
- Categorization
- Clustering
- Information Visualization

2. Structured Data Mining
The Structured data on the Web represents their host pages. Structured data is easily extracted compared to unstructured texts. The techniques used for mining structured data are

- Web Crawler
- Wrapper Generation,
- Page content Mining.

3. Semi-Structured Data Mining
Semi-structured data evolving from rigidly structured relational tables with numbers and strings to enable the natural representation of complex real world objects without sending the application writer into contortions. HTML is a special case of such intra-document structure [12]. The techniques used for semi structured data mining are

- Object Exchange Model (OEM),
- Top down Extraction
- Web Data Extraction language

4. Multimedia Data Mining
Multimedia data mining can be defined as the process of finding interesting patterns from media data such as audio, video, image and text that are not ordinarily accessible by basic queries and associated results. The aim of doing Multimedia data mining is to use the discovered patterns to improve decision making. Comparison of Multimedia data mining techniques with state of the art video processing, audio processing and image processing techniques is also provided [13]. The techniques of Multimedia data mining are:

- SKICAT
- Color Histogram Matching
- Multimedia Miner
- Shot Boundary Detection.

1.2 Web Structure Mining

Web structure mining is based on the link structures with or without the 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. The goal of web structure mining is to generate structured summary about websites and web pages. It uses treelike structure to analyze and describe HTML or XML. Some algorithms have been proposed to model the Web topology such as HITS, PageRank and improvements of HITS by adding content information to the links structure and by using outlier filtering. These models are mainly applied as a method to calculate the quality rank or relevancy of each Web page. 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.

1.3 Web Usage Mining

The Web usage mining is also known as Web Log mining, which is used to analyze the behavior of website users. This focuses on technique that can be used to predict the user behavior while user interacts with the web. Web usage mining allows the collection of Web access information for Web pages. This information is often gathered automatically into access logs via the Web server. Mainly there are four types of data sources present in which usage data is recorded at different levels they are: client level collection, browser level collection, server level collection and proxy level collection. It contains four processing stages including

- Data collection
- Preprocessing
- Pattern discovery and Analysis

2. WEB MINING TOOLS

As we have seen web mining having sub categories as Web Content Mining, Web Structure Mining, Web Usage Mining. Different types of tools used in all these mining categories. We will see tools of these different categories one by one.

2.1 Web Content Mining Tool [4]

(i) Web Info Extractor
This tool is helpful in mining extract structure or unstructured data from web page, extracting web content, and monitoring content update.

(ii) Mozenda
To extract web data easily and to manage it affordably Mozenda is useful. Mozenda supports logins, paging throughout lists of results, AJAX, frames, with other difficult web sites. Mined data can be accessed online, exported, as well as used throughout an API.

(iii) Screen-Scraper[8]
Screen-scraper allows mining the content from the web, like searching a database, SQL server or SQL database, which interfaces with the software, to achieve the content mining requirements. Latest screen scrapers provide the information in HTML, thus it be able to access with a browser.

(iv) Automation Anywhere7
Automation anywhere is a web data extraction tool used for retrieving web data effortlessly, screen scrape from web pages or use it for web mining.

Commonalities and Differences between the Above Tools

Commonalities
All the tools automate the business task and retrieve the web data in an efficient way.

Differences

- Screen-scrapper needs prior knowledge of proxy server and some knowledge of HTML and HTTP where as other tools do not require any such knowledge and it need Internet connection to run.
- Automation-Anywhere 7 allows recording of actions this facility is not provided in the other tools.
- Though we have setup file, Mozenda will not allow us to install without Internet connection, other tools can be install offline.

2.2 Web Structure Mining [6]

It is a process to discover the relationship between web pages linked by information or direct link connection. There are some possible tasks of link mining:

1. Link-Based Classification
Is the most recent upgrade of a classic data mining task to linked Domains. The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the web page.

2. Link-Based Cluster Analysis
The goal in cluster analysis is to find naturally occurring sub-classes. The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups. Different than the previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data.

3. Link Type
There are a wide range of tasks concerning the prediction of the existence of links, such as predicting the type of link between two entities, or predicting the purpose of a link.

4. Link Strength
Links could be associated with weights.

5. Link Cardinality
The main task here is to predict the number of links between objects. There are some uses of web structure mining like it is:

- Used to rank the user's query
- Deciding what page will be added to the collection
- Page categorization
- Finding related pages
- Finding duplicated web sites
- To find out similarity between them Many authors have proposed web structure mining algorithms like: Page rank algorithm, weighted page rank algorithm, Hyper Induced Topic search algorithm, weighted Topic sensitive page rank algorithm. In next section we will explain these algorithms in detail.

2.3 Web Usage Mining [3, 14]

There are different task to be carry out in Web Usage Mining that are given below and different tools are used for that work.

(i) Data Preprocessing
The first step of Web Usage Mining is preprocessing of data stored in web logs as it is noisy in nature. Preprocessing consist of converting the usage, content and structure information contained in various available data sources into the data abstractions necessary for pattern discovery.

(ii) Usage Preprocessing
This is considered as most difficult task of web usage mining because of presence of incomplete and inconsistent data in server log. Unless a client side tracking mechanism is used, only IP address, agent and server side click stream are available to identify users and server sessions. Some of the encountered problems are: single IP address/multiple server sessions, multiple IP address/single server session, multiple IP address/single user and multiple agent/single user. Usage preprocessing also encountered the problem of inferring cached page references.

(iii) Content Preprocessing
Content preprocessing consist of transforming unstructured and semi structured documents like text, images, scripts into the forms that are suitable for web usage mining.

(iv) Pattern Discovery
It focuses on to uncover patterns from the abstractions produced as a result of preprocessing phase. Pattern discovery drawn upon various methods and techniques developed from several fields such as data mining, machine learning, statistics and pattern recognition. Discovery of desired patterns and to extract understandable knowledge from them is a challenging task. This phase explains some of algorithms.

(v) Pattern Analysis
Pattern analysis is last step in the overall web usage mining process. The motivation behind this phase is to separates the interesting and uninteresting patterns from the overall patterns discovered during pattern discovery phase.
Different types of tools used in all the three stages of web usage mining are described in table 1.

Table -1: Tools Used in Various Stages of Web Usage Mining

Tools	Features
Data Pre-Processing Tools	
Data Preparator	Performs cleaning, extraction and transformation of data before pattern discovery
Sumatra TT	Platform independent data transformation tool. Based on Sumatra script and support Rapid application Development
Lisp Miner	Performs data pre-processing by analyzing the click stream and data collected.
Speed Tracer	Mines web server logs and reconstruct the user navigational path for session identification
Pattern Discovery Tools	
Sewebar- Cms	Provides interaction between data analyst and domain expert to perform discovery of patterns. Helps in selection of rules among various rules in association rule mining [34].
i-Miner	Discover data cluster by using fuzzy clustering algorithm and fuzzy inference system for pattern discovery and analysis
Argunaut	Develop the patterns of useful data by using sequence of various rules.
MiDas(Mining Internet Data for Associative Sequences)	Discover marketing based navigational pattern from log files. It applies more features to traditional sequential method.
Pattern Analysis Tools	
Webalizer	GNU GPL license based and produces web pages after analyzing patterns.
Naviz	Visualization tool that combines 2-D graph of visitor access and grouping of related pages. It describes the pattern of user navigation on the web.
WebViz	Analyze the patterns and provides them in the form of graphical patterns
Web Miner	Mines the useful patterns and provides the user specific information
Stratdyn	Enhances WUM and provides visualization of patterns

3. WEB MINING APPLICATIONS [15]

In past few years web applications are being developed at a much faster rate in the industry and also research in web related technologies. Many of these are based on the use of web mining concepts, even though the organizations that developed these applications. We describe some of the most successful applications in this section.

(i) Web Search--Google

Google is one of the most popular and widely used search engines. It provides users access to information from over 2 billion web pages that it has indexed on its server. The quality and quickness of the search facility makes it the most successful search engine. Earlier search engines concentrated on web content alone to return the relevant pages to a query. Google was the first to introduce the importance of the link structure in mining information from the web. PageRank, which measures the importance of a page, is the underlying technology in all Google search products, and uses structural information of the web graph to return high quality results.

The Google toolbar is another service provided by Google that seeks to make search easier and informative by providing additional features such as highlighting the query words on the returned web pages. Google's web directory provides a fast and easy way to search within a certain topic or related topics. The advertising program introduced by Google targets users by providing advertisements that are relevant to a search query. One of the latest services offered by Google is Google News. It integrates news from the online versions of all newspapers and organizes them categorically to make it easier for users to read "the most relevant news." It seeks to provide latest information by constantly retrieving pages from news site worldwide that are being updated on a regular basis.

(ii) Web-Wide Tracking

"Web-wide tracking," is an individual across all sites he visits, is an intriguing and controversial technology. It can provide an understanding of an individual's lifestyle and habits to a level that is unprecedented, which is clearly of tremendous interest to
marketers. Example- DoubleClick Inc.

(iii) Understanding Web Communities-AOL

It is One of the biggest successes of America Online (AOL) has been its sizeable and loyal customer base. A large portion of this customer base participates in various AOL communities, which are collections of users with similar interests. AOL provides them with useful information and services. Over time these communities have grown to be well-visited waterholes for AOL users with shared interests. Applying web mining to the data collected from community interactions provides AOL with a very good understanding of its communities, which it has used for targeted marketing through advertisements and e-mail solicitation. Recently, it has started the concept of "community sponsorship," whereby an organization, say Nike, may sponsor a community called "Young Athletic Twenty Somethings."

(iv) EBay

The genius of eBay's founders was to create an infrastructure that gave this urge a global reach, with the convenience of doing it from one's home PC. E-bay has detailed data on bid history, participant rating, bid data, usage data. In addition, it popularized auctions as a product selling and buying mechanism and provides the thrill of gambling without the trouble of having to go to Las Vegas. All of this has made eBay as one of the most successful businesses of the internet era. eBay is now using web mining techniques to analyze bidding behaviour to determine if a bid is fraudulent .Recent efforts are geared towards understanding participants' bidding behaviours/patterns to create a more efficient auction market.

(v) Personalized Portal for the Web—MyYahoo

Yahoo is an one of the search engine.Yahoo was the first to introduce the concept of a "personalized portal," i.e. a web site designed to have the look-and-feel and content personalized to the needs of an individual end-user. Mining MyYahoo usage logs provides Yahoo valuable insight into an individual's web usage habits, enabling Yahoo to provide personalized content, which in turn has led to the tremendous popularity of the Yahoo web site.

(vi) V-TAG Web Mining Server-Cannotate Technologies

The web mining server supports information agents that monitor, extract and summarize information from web sources. It is easily to set up graphical user interface. Automation of tracking and summarizing helps businesses and enterprises to analyse the various processes easily

4. CONCLUSIONS

In this paper we described so many tools available to work on web mining some prominent tools/techniques for Web Content Mining, Web Structure Mining and Web Usage Mining. We analyzed their strengths and limitations and provide comparison among them. Finally applications are discussed which specifies a fields where actually web mining is used. So we can say that this paper may be used as a reference by researchers when deciding which tool/techniques are suitable.

REFERENCES

[1] D. Sridevi, Dr. A. Pandurangan, Dr. S. Gunasekaran, "Survey on Latest Trends in Web Mining", International Journal of Research in Advent Technology, Vol. 2, No.3, March 2014.

[2] R. Agrawal, T. Imielinski and A. Swami, database mining: a performance perspective, IEEE Transactions on Knowledge and Data Engineering, 1993.

[3] Kamika Chaudhary, Santosh Kumar Gupta, Web Usage Mining Tools & Techniques: A Survey in International Journal of Scientific & Engineering Research, Volume 4, Issue 6,June-2013 1762 ISSN 2229-5518.

[4] V. Bharanipriya & V. Kamakshi Prasad, Web Content Mining tools: A Comparative Study in International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 211-215.

[5] Bodyan G.C, Shestakov T.V, "Web Mining in Technology Management", Engineering Universe for Scientific Research and Management, Vol 1 Issue 2, April 2009.

[6] Preeti Chopra, Md. Ataullah, a Survey on Improving the Efficiency of Different Web Structure Mining Algorithms in International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-3, February 2013.

[7] http://en.wikipedia.org/wiki/Web_mining

[8] Screen-scraper, http://www.screen-scraper.com Viewed 19 February 2013.

[9] Zhang, Q., Segall, R.S., Web Mining: A Survey of Current Research, Techniques, and Software, International Journal of Information Technology & Decision Making. Vol.7, No. 4, pp. 683-720. World Scientific Publishing Company (2008).

[10] Darshna Navadiya, Roshni Patel, Web Content Mining Techniques-A Comprehensive Survey, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 10, December- 2012 ISSN: 2278-0181.

[11] Mozenda, at: http://www.mozenda. com/web-mining-software Viewed 18 February 2013.

[12] Web Mining https://www. techopedia.com/definition/ 15634/ web-mining

[13] Chidansh Amitkumar Bhatt, Mohan S. Kankanhalli Multimedia data mining: state of the art and challenge. Journal Multimedia Tools and Applications archive Volume 51 Issue 1, January 2011.

[14] J. Srivastava, R. Cooley, M. Deshpande and P. Tan., "Web Usage Mining: Discovery and Applications of Usage Patterns from Web data", Department of Computer Science and Engineering, University of Minnesota. SIGKDD Explorations, 1(2):12, January 1999.

[15] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, "Web Mining — Concepts, Applications, and Research Directions", AHPCRC technical report 2003-110,July 2003

[16] J. Srivastava, R. Cooley, M. Deshpande, and P.N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, vol. I, no. 2, pp. 12-23, 2000.