Research Trends in Web Mining

[1] Rachana Parikh, [2] Hitul Marvaniya, [3] Rachit Adhvaryu
[1] Assistant Professor, [2] Assistant Professor, [3] Lecturer
Information Technology Department, V.V.P Engineering College, Rajkot, Gujarat, India.

***Abstract – Web Mining is one of the application of Data Mining*. It is a technique to extract information from the web which includes web documents, hyperlinks between the documents and web usage logs. Modern developments in digital media technologies have evolved a huge amount of data transmitting over the web and with this a huge data storage is required for easy and feasible access. Web mining plays an important role in the decision making in the corporate, education and research environment. In this paper we describe the detailed survey of web mining. The focus of the paper is to bring in light the importance of Web Mining. The paper describes the details of web mining like its types, techniques used to extract information from the web, its tools and its applications.**

***Keywords* – Web Mining, Web Content Mining, Web Structure Mining, Web Usage Mining**

I. INTRODUCTION

The World Wide Web (WWW) has a lots of information and this information is increasing in a large amount daily. It is a very complex task to identify or retrieve information from such information. A web or a website a collections of webpages containing images, videos, text or similar digital data. The need to understand large, complex, information-rich data sets is common to virtually all fields of business, science, and engineering[1]. The ability to extract useful knowledge hidden in these data and to act on that knowledge is becoming increasingly important in today's competitive world. Web Mining is the application of Data Mining techniques to extract information from the web which includes web documents, hyperlinks between the documents and web usage logs. Also The entire process of applying a computer-based methodology for discovering and extracting knowledge from web documents is a web mining[2]. As the web data is updated every second, it is not compulsory that every user will get the same data whenever it is retrieved. Web Mining is classified into 3 main mining techniques[2]. The taxonomy of web mining is as follows:

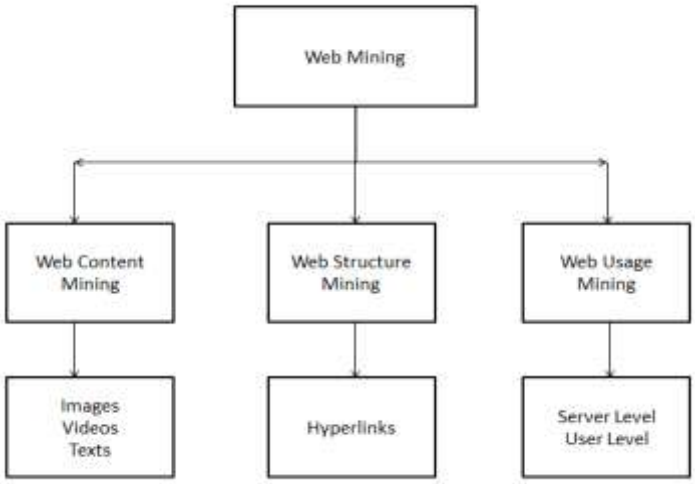

Fig1: Web Mining Taxonomy

The rest of the paper is organized as follows: Section II describes Web Content Mining, section III describes Web Structure Mining, section IV describes Web Usage Mining, section V describes tools of web mining, section VI describes applications of web mining and we conclude our study in chapter VII.

II. WEB CONTENT MINING

Web Content Mining is the way of extracting important and required information from the web documents. The information can be in the form of images, videos, audios, texts. Text mining and developing applications for the same is one of the favourite topic for the researchers. Research in web content mining consists resource discovery from the web, categorizing and clustering of documents and extraction of informations from the webpages[4]. Web content mining can be performed on images, videos, audio and texts. The mining techniques for each is different. Here is the brief details of each mining techniques.

Image Mining

Image Mining is the technique which is used to detect unusual patterns and extract important and useful information from the images stored on the web and large database. Thus image mining mainly deals with defining relationships between different

images from the web and large databases[5]. Image mining is used in various fields like medical diagnosis, space research, remote sensing, agriculture and industries. The images include maps, geological structures and even it is used in the field of education.

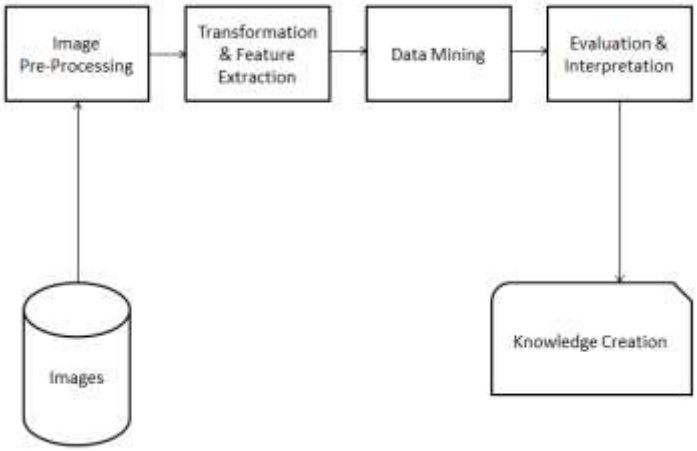

Fig2: Image Mining Process

The main problem in image mining is to find out a technique to convert a low level pixel images to a high level pixel images. Ji Zhang, Wynne Hsu and Mong Li Lee[5] proposed an efficient imformation-driven framework for image mining. In that four levels of information were defined: Pixel Level, Object Level, Semantic Concept Level and Pattern and Knowledge Level. They highlighted the requirement of image mining and also defined unique features of image databases.

Video Mining

Mining video is more complicated than mining image data. Video is the collection of moving images like animation. There are 3 types of videos: 1) the produced (movies, news videos etc), 2) the raw (traffic, surveillance etc) and 3) the medical video (x-ray, cardiogram etc). The information from the video can be a) detecting trigger events (movement of vehicle and people), b) determining typical patterns of activity, generating person or object centric views of activity, and c) classifying activities into named categories (walking, sleeping etc), clustering and determination of interactions between two entities[6]. Video mining can also be classified into pixel based, statistics based, feature based and histogram based.

Audio Mining

As audio is the continuous media like video, the techniques and tools for audio processing and mining information is similar to video mining. Audio data can be in the form of radio, speech or spoken languages. Also television news has audio which are integrated with the videos. Mining audio data requires conversion of audio to speech for better processing. Audio data can be directly mined by using audio information processing methods and extracting selected audio data. Very few work has been carried out in the field of audio mining.

Text Mining

The most trending research in the field of web content mining is text mining. The text mining refers to the text representation, classification, clustering, information extraction and search for hidden patterns. Text mining is the process of extracting useful information from the text and converting to automated discovery of knowledge[8]. It is natural extension of data mining or applying data mining techniques on a specific domain. Text categorization is one of the technique of text mining which classifies the texts in a particular domain[9]. It works on keywords extraction from the document. A lot of work has been carried out in the field of text mining.

III. WEB STRUCURE MINING

This type of mining focuses on the data which describes the structure of the content of the web page. It is classified into two types: 1) intra-page structure: existence of links within the page, 2) inter-page structure: the connection of one web page with other web pages[2]. This can be classified into two types based on structure of information:

Hyperlinks

A hyperlink is a structural unit that connects one web page with other web page either within same location or different location. A hyperlink connecting web pages in the same location is called intra-document hyperlink and a hyperlink connecting web pages at different locations is called inter-document hyperlink[2].

Document Structure

The content within a web page can also be organized in a tree structure based on HTML and XML Tags used to create a web page. Mining can be done to identify document object model (DOM) structures automatically from the documents[2].

IV. WEB USAGE MINING

Web usage mining is one techniques of data mining to extract interesting and useful patterns from the web usage logs[2]. Usage logs stores the identity or the origin of web users along with the browsing behavior on the web site. Web mining can be grouped based on the type of usage logs:

Web Server Data

User logs are collected by the web server which includes IP address, page references and the time accessed by the user[2].

Application Server Data

Application servers are used to track various types of business events which can be used to improve the performance for any business firms[2]. For e.g. E-commerce websites uses such servers to know the events, business policies developed by their competitors.

User Level Data

User level data is the software developed using the information available from the web server and application server data[2]. It is an end user application which is used by different users for different purposes.

V. TOOLS FOR WEB MINING

Web mining tools are the softwares or the applications which helps the users to download essential information from the web. It collects the exact and required information for the user which can be helpful in mining. The different tools are:

Automation Anywhere

It is a tool which is used to extract web data very easily, screen shots of web data which can be used in web mining. It is unique Intelligent and Smart Automation Application used for quick automation of any complex tasks[10].

Web Info Extractor

This tools is used to collect web content, constantly updating data and analyzing data. This tool can be used to extract information like images, vidoes and texts[11].

Screen-Scraper

It is a tool to extract information from the websites. It can be used in searching databases and document structure. It provides a graphical interface allowing the user to navigate through URLs, data elements and hyperlinks and extract useful information from it[12].

Mozenda

This tool is used to extract and manage web data. User is allowed to setup tools at diiferent places whch can store and publish data at a regular interval of time[13]. This data can be used with other applications for mining puroses.

Web Content Extractor

It is a tool for extracting web scraping, data mining and information retrieval from the internet. This tool is used to extract information from various websites like online auctions, online shopping, business directories, financial sites etc. The data can be represented in the form of excel, HTML, XML or any other script[14].

VI. APPLICATIONS OF WEB MINING

There are many applications of web mining. Most dominating applications of web mining are as follows:

Customer Experience on B2C E-commerce

In the world of online shopping, it is very important to know the customer behavior and experience with the website. The feedback of the experience given by the user helps the website owner to improve their content in an efficient way. The main target of the website owner is that once a customer is visiting the website for a purchase, the user should not move to the other websites.

Web Search

Google is the best and widely used search engines. It provides the users to access information from over billiions of web pages indexed on its server. The quickness and quality of information provided by the search engines make them the most successful search engines. Web mining helps the search engines to know the behavior of the user, the keywords they search and based on this they give a PageRank. PageRank provides the importance of a web page. Thus web mining can be very useful application for search engines.

VII. CONCLUSION

As the web data and its usage increases day by day, it is very important to analyze the web data and extract the information. Thus web mining and its techniques play an important role in in information extraction from the web. In this paper, we provided survey of web mining techniques, tools and applications. By the techniques and the tools described in the paper, one can use for his

research and new techniques can be developed for more effective, efficient and faster results. We hope that this survey is the beginning for a fruitful discussions in future.

REFERENCES

[1] Arvind Kumar Sharma, P.C. Gupta, ―Exploration of efficient methodologies for the improvement in web mining techniques-A survey, International Journal of Research in IT & Management (ISSN 2231-4334) Vol.1, Issue 3, July 2011.
[2] R. Kosala, H. Blockeel, ―Web Mining Research: A Survey, SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.
[3] 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.
[4] S. K. Madria, S. S. Bhowmick, W. K. Ng,and E. P. Lim, "Research Issues in Web Mining", in proceding of data mining and knowledge discovery, 1st International conference, DaWk 99, pp 303-312, 1999.
[5] Ji Zhang, Wynne Hsu and Mong Li Lee "An Information-Driven Framework for Image Mining" Database and Expert Systems Applications in Computer Science, 2001, Volume 2113/2001, 232-242, DOI: 10.1007/3-540-44759-8_24
[6] Boresczky J. S. and L. A. Rowe, "A Comparison of Video Shot Boundary Detection Techniques", Storage & Retrieval for Image and Video Databases IV, Proc. SPIE 2670, 1996, pp.170-179.
[7] A. Czyzewski, "Mining Knowledge in Noisy Audio Data", in Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, pages 220-225, 1996.
[8] Chakrabarti S. " Mining the web: Analysis of Hypertext and Semi Structured Data", morgan Kaufmann, San Francisco, CA
[9] Hearst, M. A, "What is Text Mining?", http://www.sims.berkely.edu/~hearst/text-mining.html.
[10] Automation Anywhere Manual. AA, http://www.automationanywhere.com
[11] 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)
[12] Mozendo, http://www.Mozendo.com
[13] www.screen-scraper.com
[14] Web Content Extractor help. WCE, http://www.newprosoft/web-content-extractor.htm