Explorative Study of Web Data Mining Techniques and Tools: A Review

[1]Surbhi Sharma, [2]Dinesh Soni, [3]Dr. Arvind K Sharma
[1]Dept. of CSE, Rajasthan Technical University Kota, Rajasthan, India
[2]Rajasthan Technical University Kota, Rajasthan, India
[3]Dept. of CSI, University of Kota, Kota, Rajasthan, India

Abstract
Web data mining is usually a technique of data mining which is utilized for serving web-based applications by using web data over World Wide Web. It is a technique of retrieving information over the World Wide Web that contains web based documents, hyper documents, web links to various web pages and other resources over the Web. It evolves three main techniques such as structure mining, content mining, and usage mining. In this paper we have been presented that how web data mining is to be used, to be implemented, and to be obtained useful information from the Web. A survey of different web data mining techniques and tools has also been shown. Further, we have been tried to identify the research domain in web data mining where further future work can be continued.

Keywords
Web Data Mining, WWW, Web Mining, Techniques, Tools

I. Introduction

Today World Wide Web (WWW) has become a complex universe as it updates regularly. WWW is basically a source of huge amount of information that provides all the needful sources of data mining [1]. WWW is a vast resource of multiple types of information in various formats which is very useful in the analysis of business progress that is very much important to stand in the competition of business now days. WWW is an online system that contains interlinked files such as images, videos, audios and other form of multimedia data [2]. Web data mining has been frequently used all over the world from a small scale business to a large scale business. This technique of data mining is used for web based applications and is the major need of each and every field. Web data mining is a term used for a technique, through which various web resources are used for collecting the useful information that makes it easy for an individual or a company for utilizing these resources and information in their best interest. One of the important challenges is to mining the web data as the data available on the World Wide Web is increasing continuously, thus it is difficult to retrieve information without data mining. Data Mining, usually called Web mining when applied to the Internet, is a process of extracting hidden predictive information and discovering meaningful patterns, profiles, trends from huge databases. Data mining of the World Wide Web is mainly designed for the comfort of the developers and the users of web data system. As a major source of information the web serves as a resource provider for the researchers of web data mining domain. Out of the given information deriving only the required information of data is the main target of web mining. WWW contains massive information which can be utilized easily by anyone, anywhere and anytime.

The rest of paper is organized as follows: Section II presents an overview of web data mining and its taxonomy. Section III covers literature review. Section IV describes the complete proposed methodology. Section V explores several data mining tools. Section VI provides important research issues in web data mining. Section VII concludes the paper while references are mentioned in the last.

II. Web Data Mining–An Overview

In 1996, Etizoni [3] was the first person who has introduced the term Web mining. He initially started by making a hypothesis that information on web is sufficiently structured and outlines the subtasks of web mining. According to him web mining is the technique of extracting the required information from the World Wide Web documents and web services [3]. The World Wide Web has been serving as huge distributed global information service centre for news, advertisements, consumers, e-commerce, education, individual, company, etc. Also, the WWW has a rich and dynamic collection of hyperlinks, hyper documents, providing rich source for data mining and web mining. Extracting knowledge from the web is the main task of web mining.

A. Taxonomy of Web Mining

The various web data mining techniques can be classified into the following categories, each category is shown in fig. 1.

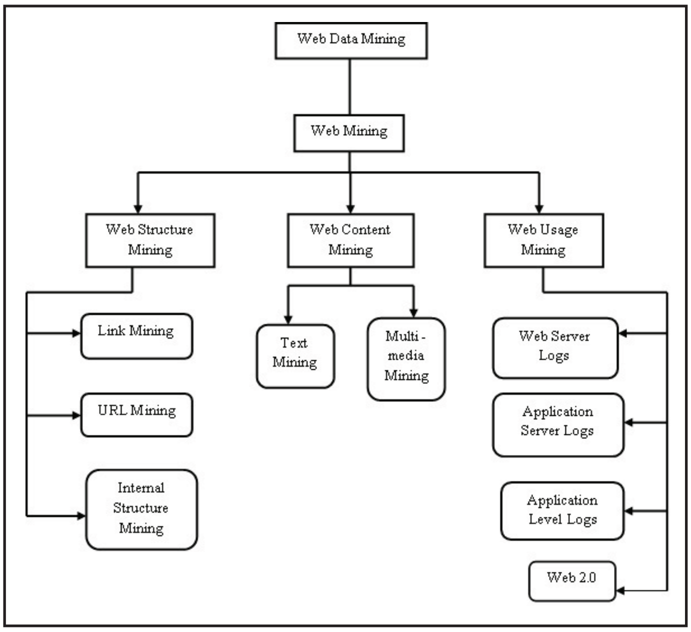

Fig. 1: Taxonomy of Web Mining

Furthermore, the web mining can be classified into three main categories which are as follows:

1. Web Structure Mining

This is mainly used for describing the structure of the contents of the website. It can be defined in terms of graph, where web pages are its nodes and its hyperlinks are its edges [4]. It shows the web links from one web page to another therefore we can say it shows the relationship among the web and users. The focus of this is structural summary of web pages and web sites. Web structure mining mainly works on- Link mining, Internal structure mining and URL mining [3].

2. Web Content Mining

This method is used for extracting the required content from the various web pages available and its contents such as, image, audio, video, text, etc. The primary resources of web that are mined are individual web pages. Web content mining is mainly associated with text mining because most of the web content is in the form of text [2]. Thus Web content mining needs its own applications of text mining and many other distinct approaches. It mainly focuses on- Web Text mining and Web Multimedia Mining [3].

3. Web Usage Mining

This technique is used to define the mode through which users can interact with the servers or can access the available web pages. It includes information generated by client server transaction from one or more web localities [4]. Its main objective is to finding the usage patterns from applications based on web. It consists of three phases: preprocessing, discovery of usage pattern and analysis of the pattern. This technique of mining is used by server logs and it aimed at getting useful users. Who can access information in the form of web logs.

III. Literature Review

This section provides a summary of several researches carried out by the many authors on the basis of past literatures and articles in the domain of web data mining. Some of the research works are discussed.

In 2016, Swapnil S. Patil and H.P. Khandagale [1], implemented web usage mining techniques for enhancing web navigation usability .This paper provided a standard way for the developers for recognizing the actual behavior of usage and the developed system was more useful for the developers as well as the users.

In 2012, Arvind Kumar Sharma and P.C. Gupta [2], present a survey of web content mining tool to improve the techniques of web data mining, in that several tools for web content mining are discussed with their merits and demerits.

In 2010, Brijendra Singh and Hemant Kumar Singh [3], given a paper which gives the survey and comparison of various web data mining methods and also provides some important research issues.

In 2014, K. Mohammad Mujahid, et al. [4], presented a paper in which a study is done on web mining and it presented facts on how to extracting needful information from the web also describes the past present and future of web mining.

In 2013, Abdelhakim Herrouz, et al. [5], wrote a paper which includes the overview of web content mining tools with a comparative table of these tools based on some criteria.

In 2012, Romil. V. Patel et al.[8], presented a paper on web mining with artificial neural network which is also a technique of web mining. It mainly focuses on web usage mining.

In 2001, Khaled M. Hammouda[9], presented a review paper on web mining which covers the most representative approaches of clustering.

In 2016, Lourdu Caroline. A, et al. [10], a survey the web data mining techniques and its applications which have used in cloud computing technologies. According to this paper the information stored over the cloud for any business or knowledge based applications, could be easily used through web mining techniques.

In 2016, Bibu Skaria, et al. [11], provided a paper in which a brief introduction of web mining was given which also included overview of different web usage mining techniques in brief.

In 2015, Qingyu Zhang and Richard S. Segall [12], presented a survey of techniques and softwares for web data mining by dividing the process into five subtasks for which comparison was done of the softwares.

In 2014, Akshay A. Adsod, et al.[13], given a paper which mainly concentrates on a diagram of web mining procedures and its importance in related territories.

In 2012, C.J. Carmona [14], et al., wrote a paper in which a technique of web data mining i.e. web usage mining is used for improving the design of e-commerce website: OrOliveSur.com

In 2011, Mikalai Tsytsarau and Themis Palpanas [15], presented a paper in which a survey is done on mining subjective data on the web. In this, authors reviewed the development of opinion mining and sentiment analysis from the past years and also give directions for future.

In 2015, S. R. Kalaiselvi, et al.[16], provided a paper on web mining its concepts and applications but it mainly concentrate on only web usage mining and its phases and some of its application areas in the field of education, health and social media.

In 2013, Ahmad Tasnim Siddiqui, et al.[17], presented a paper of using web mining techniques in e-commerce applications, the main focus of this paper was business over internet.

In 2013, Monika Yadav and Pradeep Mittal [18], provided a paper in which they given overall introduction of web mining and contribution of computer science in the field of web mining.

In 2013, R. Malarvizhi and K. Saraswathi [19], presented a paper which mainly concentrates on web content mining and its techniques and tools.

In 2013, B. Lalithadevi, et al.[20], proposed new approaches for improving WWW techniques in data mining. It mainly focused on web usage mining.

IV. Proposed Methodology

To facilitate web data mining, there are various techniques of web mining which can be applied to find patterns and trends from the data collected from the World Wide Web. This section provides a brief discussion about more popular web data mining techniques available nowadays. Some of the popular web data mining techniques are as follows:

A. Classification

It is one of the most commonly used data mining technique. It consists of a set of predefined examples for developing a new model, which can easily classify massive amount of data records. As its name suggests it is basically a group of items that belong to a particular category on the basis of their common features [6]. The primary aim of this technique is to assign an accurate class to the previously unseen records.

B. Association Rule Mining

It is a basic technique of web data mining that is used for associating relationships among a set of variables and its data items. It consists of two parts antecedent and consequent, an antecedent is the data item and consequent is data item found in combination of antecedent [7]. It is a technique of analyzing data for if/then patterns.

C. Artificial Neural Network

Artificial Neural Network is one of the data mining techniques that is based on the works perform by the brain or a particular task perform by the brain [8]. It is the interconnected group of nodes with a vast network of neurons in a brain. This technique is used in web data mining for gathering information from the web in the form of neural networks which may be linear or non-linear and utilizing this required information for one or other purpose of the end user.

D. Clustering

Clustering is also one of the popular techniques of data mining which is based on concept of hierarchy model which groups together those items which are having similar features. It is believed that making group of similar items into a cluster is very helpful for retrieving the relevant information easily and quickly and allows the users to focus their search in the right direction [9]. The cluster of similar items makes it more appropriate data gathering system.

V. Web Data Mining Tools

There are various Web data mining tools as open source softwares which are freely available for mining of web data. These tools have been used to gather correct and perfect information by using weblog data. In this section, some of the useful and popular web data mining tools are explored and discussed here.

A. WEKA

Written in Java, WEKA (Waikato Environment for Knowledge Analysis) is a well-known tool of machine learning software [26,2]. Weka supports several typical data mining tasks, particularly data preprocessing, clustering, classification, regression, visualization, and feature selection. Its techniques are based on the hypothesis that the data is available as a single flat file or relation, where each data point is labeled by a fixed number of attributes. WEKA provides access to SQL databases utilizing Java Database Connectivity (JDBC) and can process the result returned by a database query. Its main user interface is the Explorer, but the same functionality can be accessed from the command line or through the component-based Knowledge Flow interface.

B. Tanagra

Tanagra is free Data Mining software for academic and research purposes [22-23]. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. It runs under almost Windows Systems, in any case it has been tested under Windows 98, 2000, XP, Vista and Windows 7/8.1.

C. Orange

It is a component-based data mining and machine learning software tool that features friendly yet powerful, fast and versatile visual programming front-end for explorative data analysis and visualization and Python bindings and libraries for scripting. It contains complete set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques. It is written in C++ and Python [21, 23], and its graphical user interface is based on cross-platform of framework.

D. Rapid Miner

It is formerly called as YALE (Yet Another Learning Environment) is an environment for machine learning and data mining experiments that is utilized for both research and real-world data mining tasks [24 , 23]. It enables experiments to be made up of a huge number of arbitrarily nestable operators, which are detailed in XML files and are made with the graphical user interface of Rapid Miner. Rapid Miner provides more than 500 operators for all main machine learning procedures, and it also combines learning schemes and attribute evaluators of the Weka learning environment. It is available as a stand-alone tool for data analysis and as a data-mining engine that can be integrated into your own products.

E. KNIME

KNIME (Konstanz Information Miner) is a user friendly, intelligible and comprehensive open-source data integration, processing, analysis, and exploration platform [23, 25]. It gives users the ability to visually create data flows or pipelines, selectively execute some or all analysis steps, and later studies the results, models, and interactive views. KNIME is written in Java, and it is based on Eclipse and makes use of its extension method to support plugins thus providing additional functionality. Through plugins, users can add modules for text, image, and time series processing and the integration of various other open source projects, such as R programming language, Weka, and LibSVM etc.

F. Screen-Scraper

This is a tool used for extracting data from websites and uses that information in other contexts similar to databases it allows mining of data through World Wide Web. It includes mining of web data consisting of searching of databases which interacts with the available software to achieve the requirements. One of the most regular usages of this software and services is to mine data on products and download them to a spreadsheet [5].

G. Web Info Extractor

This tool is helpful in mining web data, extracting web content, and monitoring content update. Thorny template rules are not required to be defined. For mining web data and for content retrieval it is a very powerful tool. Some of the features [5] are as follows:

- No need to learn boring and complex template rules and it is easy to define extract tool.
- Extract tabular as well as unstructured data to file or database.
- Monitor Web pages and extract new content when update.
- Can deal with text, image and other link file.
- Can deal with Web page in all language.
- Running multi-task at the same time.
- Support recursive task definition.

H. Automation Anywhere

Automation anywhere is a tool used for data extraction used for retrieving web data, screen scrape from Web pages. It is also used for Web mining. Its main features[5] are as follows:

- Unique SMART Automation Technology for fast automation of complex tasks.
- Record keyboard and mouse or use point and click wizards to create automated tasks quickly.

I. Web Content Extractor

It is a powerful and easy to use data extraction tool for web scraping, data mining or data extraction from the internet [2]. Some of the features are:

- It helps to collect the market figures, product pricing data, or real estate data.
- It helps users to extract the information about books, including their titles, authors, descriptions, ISBNs, images, and prices, from online book sellers.
- It assists users in automate extraction of auction information from auction sites.
- It assists to Journalists extract news and articles from news sites.
- It helps people seeking a job extract job postings from online job websites. Find a new job faster and with minimum inconveniences
- It Extracts the online information about vacation and holiday places, including their names, addresses, descriptions, images, and prices, from web sites[5].

J. Web Log Expert Tool

Web Log Expert is a fast and powerful Web log Analyzer Web mining tool [27]. This software tool helps to reveal important statistics regarding a Web site's usage such as: activity of visitors, access statistics, and paths through the website, visitors' browsers, and much more.

K. Absolute Log Analyzer Tool

Absolute Log Analyzer is a client-based log file analysis software tool it is designed for Web traffic analysis [27]. Firstly, log files need to be added to the analysis and the results are then displayed. Apart from the graphical user interface (GUI), Absolute Log Analyzer also has a Command Line Interface (CLI).

VI. Research Issues

Web data mining is a young and hot research area today. We have been presented some vital issues of web data mining in this section. Some of the issues are as follows:

- Developing intelligent tools for information retrieval
- Extracting statistical information and discover interesting user patterns
- Clustering the user into groups according to their navigational behaviour.
- Discovering the potential correlations between web pages and user groups
- Identifying potential customers for e-commerce market
- Enhancing the quality and delivery of Internet information services to the end users
- Improving web server system performance and web site design.
- Facilitating web personalization
- Ordering documents matching a user query (ranking)
- Deciding what pages to add to a collection page categorization
- Finding related pages
- Evaluating duplicate web sites and also to find out similarity between them
- Extracting keywords and key phrases
- Identifying the system errors and users behaviour
- Discovering grammatical rules collections
- Hypertext classification/ categorization
- Extracting key phrases from text documents
- Hierarchical clustering
- Predicting relationships
- Learning extraction rules

VII. Conclusion

Web data mining is a fertile area of research where data mining may be applied. It has a large and computational research domain. The aim of this paper is to explore various web data mining techniques and tools used to mine or extract useful information from the World Wide Web. This paper discusses about the vital research issues in the web data mining and covers the basic concepts of web data mining techniques, tools, and their taxonomy. This paper opens a new door to the researchers who wish to pursue their research in the area of web data mining.

References

[1] Swapnil S. Patil, et al., "Enhancing Web Navigation Usability Using Web Usage Mining Techniques", International Research Journal of Engineering and Technology (IRJET), Vol. 04 , Issue 06, June 2016.

[2] Arvind Kumar Sharma, P.C. Gupta,"Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining", International Journal of Advanced Research in Computer Engineering & Technology, Vol. 1, Issue 8,

October 2012.
[3] Brijendra Singh, Hemant Kumar Singh,"Web Data Mining Research: A Survey", IEEE, 2010
[4] K. Mohammad Mujahid, et al. , "Web Mining: Day-Today", International Journal of Emerging Trends and Technology in Computer Science, Vol. 3, Issue 5, Sept-Oct, 2014
[5] Abdelhakim Herrouz, et al.,"Overview of Web Content Mining Tools", The International Journal of Engineering and Science (IJES),Vol. 2, Issue 6, 2013.
[6] [Online] Available: https://sites.google.com/site/assignmentssolved/mca/semester6/mc0088/12
[7] [Online] Available: [Online]http://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining
[8] Romil. V. Patel, et al., "Introduction to Integrating Web Mining With Neural Network", International Journal of Computer Science and Information Technology & Security, Vol. 2, No. 6, December 2012
[9] Hammouda, K.,"Web Mining: Clustering Web Documents a Preliminary Review", Vectors-2 (2001): 1-13.
[10] Lourdu Caroline. A, et al., "Implementation of Different Techniques of Web Data Mining through Cloud Computing Technologies", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 6, Issue 6, June 2016.
[11] Bibu Skaria, et al.,"Literature Review on Web Mining", Bonfring International Journal of Data Mining, Vol. 6, No. 1, January 2016.
[12] Qingyu Zhang, Richard S. Segall,"Web Mining: A Survey of Current Research, Techniques, and Software", International Journal of Information Technology & Decision Making , Vol. 7, No. 4, 2008
[13] Akshay A. Adsod et al.,"A Review on: Web Mining Techniques", International Journal of Engineering Trends and Technology, Vol. 10, No. 3, April, 2014.
[14] Carmona, Cristóbal J., et al.,"Web Usage Mining to Improve the Design of an e-commerce Website: OrOliveSur.com", Expert Systems with Applications, 39.12 (2012): pp. 11243-11249.
[15] Tsytsarau, Mikalai, Themis Palpanas,"Survey on mining subjective data on the Web", Data Mining and Knowledge Discovery 24.3 (2012): pp. 478-514.
[16] Kalaiselvi, SR, S. Maheshwari, V. Shobana,"Web Mining–Data Mining Concepts, Applications, and Research Directions", Vol. 4, Issue 11, November 2015.
[17] Siddiqui, Ahmad Tasnim, Sultan Aljahdali,"Web mining Techniques in e-commerce applications", International Journal of Computer Applications, Vol. 69, No. 8, May, 2013.
[18] Monika Yadav, Pradeep Mittal,"Web Mining: An Introduction", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 3, March 2013.
[19] Malarvizhi, R., K. Saraswathi,"Web Content Mining Techniques, Tools & Algorithms–A Comprehensive Study", International Journal of Computer Trends and Technology (IJCTT), Vol. 4, 2013.
[20] Lalithadevi, B., A. Merry Ida, W. Ancy Breen,"A New Approach for Improving World Wide Web Techniques in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 1, January 2013.
[21] [Online] Available: http://orange.biolab.si/
[22] [Online] Available: http://eric.univlyon2.fr/~ricco/tanagra/en/tanagra.html
[23] Sharma Arvind K., et al., "Evaluating WEKA over the Open Source Web Data Mining Tools", International Journal of Engineering 8.1: 2015.
[24] [Online] Available: https://rapidminer.com/
[25] [Online] Available: http://toolkit.snd.org/tools/other/knime/
[26] [Online] Available: http://www.cs.waikato.ac.nz/ml/weka/
[27] Sharma Arvind K.,"A Comparative Study between Web Mining Tools over some WUM Algorithms to Analyze Web Access Logs", International Journal of Innovative Technology and Exploring Engineering, Vol. 1, Issue1, June 2012.

Surbhi Sharma has received her B.Tech degree in Computer Science and Engineering from Rajasthan College of Engineering for Women, affiliated to Rajasthan Technical University, Kota, India in 2013 and presently pursuing her M.Tech Computer Science and Engineering from Rajasthan Technical University, Kota, Rajasthan, India. Her research interests include: Web Data Mining, Machine Learning Tools and Web Applications.

Dinesh Soni received his B.E. degree in Computer Engineering from Rajasthan University, Jaipur, India, in 2007 and M.Tech degree in Computer Technology from Indian Institute of Technology, Delhi, India in 2015. He is working as Assistant Professor in Department of Computer Science and Engineering, Rajasthan Technical University, Kota since 2008. His research interests include: Machine learning and Computer Vision.

Dr. Arvind K Sharma has received his Ph.D degree in Computer Science in the year 2013. He has more than 13 years of work experience in academic field. He has published more than 40 Papers in many National, International Journals and Conference Proceedings. He is a Managing Editor of International Journal of Computer Science and Technology. He is also Editorial Board Member and Reviewer of several National and International Journals and Conferences. He is a Member of numerous academic and professional bodies such as IEEE, WASET, IEDRC, IAENG Hong Kong, IACSIT Singapore, UACEE UK, ACM, New York. His area of interest includes- Web Usage Mining, Web Engineering, Opinion Mining, Data Analytics and Machine Learning Tools.