DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4294

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

DATA MINING TECHNIQUES WITH WEB LOG: A REVIEW

Karuna Nidhi Pandagre
Research Scholar
Aisect University
Bhopal, India

Dr S.Veenadhari
Associate Professor, CSE Department
Aisect University
Bhopal, India

Abstract: Web mining is the use of data mining techniques to automatically discover and extract information from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. Since large volumes of data is stored on web majority of the information retrieved through search engines may not be of much use to the web users. The information/data available on the web are extremely dynamic. The uploaded information can be used to gain more knowledge and based on the findings and analysis of the information make predictions as to what would be the best choice and the right approach to move toward on a particular issue. Application of data mining techniques would help to achieve the same. Web data mining is not only focused to gain business information but also used by various organizational departments to make the right predictions and decisions on business development, work flow, production processes and more by going through the business models derived from the data mining. In this paper a comprehensive review of different web mining techniques is being presented with their limitations.

Keywords: WWW, Pattern Discovery, Pre-processing, Data Mining Techniques.

INTRODUCTION

With the introduction of World Wide Web, explosive growth of information has been observed along with the increase in internet commerce. This necessitated the application of data mining techniques to infer the available data to a meaningful conclusions and this process is called web mining.

Data mining is the techniques to knowledge discover patterns in huge volumes of raw data and it is the extraction of implicit information from a large dataset. Web mining a huge, widely-distributed, highly heterogeneous, semi structured, interconnected, evolving, and hypertext/hypermedia information repository. Web is a collection of billions of data .Web mining can be referred as the revolution of the data mining techniques to web data. Web mining has three parts – content, structure and usage mining of web data.

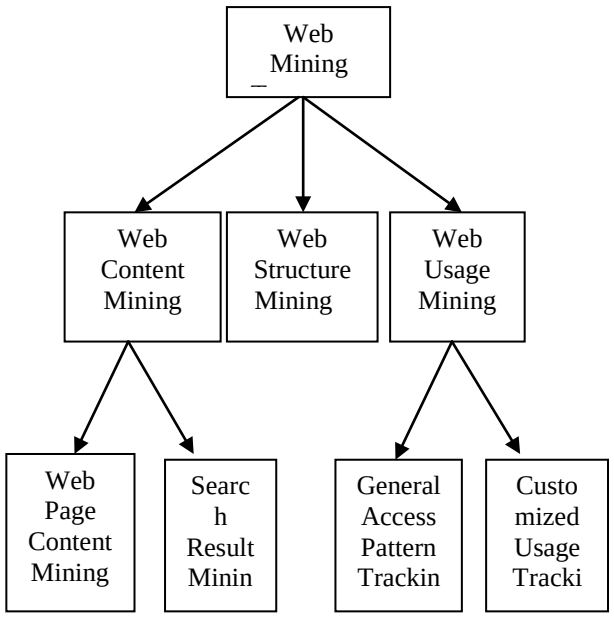

Fig 1: Web Mining Taxonomy

By Dr.Osmar R. Zaiane, 2001 in Web Mining: From Concepts to Practical Systems [12]

Web Content mining:

Web Content mining refers to the discovery of useful information from the contents of the webpage using text mining techniques. Webpage can be in traditional text form or in the form of multimedia document containing table, form, image, video and audio. Web content mining identifies the useful information from the Web contents.

Web structure mining:

The goal of web structure mining is to generate structural summary about web pages and web sites. It shows the

relationship between the user and the web. It discovers the link structure of hyperlinks at the inter document level.

Web usage mining:
It is also known as Web log mining, is used to analyze the behavior of website users. It tries to discover useful information secondary data derived from the interaction of users while surfing web. Web usage mining collects the data from Web log records to determine user access patterns of Web pages.
Web Mining enables one to discover web pages, text documents, multimedia files, images and other types or resources from web with different approaches of algorithm.
Data mining process consist of Learning the application domain, Gathering and integrating of data, Cleaning and preprocessing data, Reducing and projecting data, Choosing functions of data mining, Choosing the mining algorithm(s), Data mining: search for patterns of interest, Evaluating results, Interpretation: analysis of results, Use of discovered knowledge.

The layout of this paper for the upcoming sections will be as section 2 will give an overview of different Data Mining techniques. In Section 3 is about Conclusion. In section 4 References.

SURVEY ON WEB LOG MINING

Oren Etzioni [3] first proposed the term of web mining in his paper in the year 1996 as a sequence of tasks and Colley et.al. [2] defined web mining in terms of the types of web data that was being used in the mining process. This definition of web mining is being widely adopted by the researchers **(Madria et.al**[10]; **Borges** and **Levene** [1]; **Kosala and Blockeel** [4]).

Based on the kinds of data to be mined web mining can be broadly divided into three categories i.e., web content mining, web structure mining and web usage mining. Web content mining is the process in which useful information is extracted from the contents of web documents. The contents of web document may consists of text, images, videos, audio, tables. Web structure mining is the process of discovering structure information such as nodes, hyperlinks from the web. In web usage mining data mining techniques are applied to discover interesting usage patterns such as user's browsing behaviour, user logs etc from web usage data, in order to understand and better serve the needs of web-based applications (**Srivastava, Cooley, Deshpande,** and **Tan 2000**). **M. Spiliopoulou** [6] categorized the Web mining into Web usage mining,

Web text mining and user modeling mining; while today the most recognized categories of the Web data mining are Web content mining, Web structure mining, and Web usage mining. It is clear that the classification is based on what type of Web data to mine.

Madria et.al., [5] described web content mining as automatic search of information resource available online, and involves mining web data contents. In the Web mining domain, Web content mining essentially is an analog of data mining techniques for relational databases, since it is possible to find similar types of knowledge from the unstructured data residing in Web documents.

Kosala and Blockeel [4] summarized the research works done for unstructured data and semi-structured data from information retrieval view. It shows that most of the researches use bag of words, which is based on the statistics about single words in isolation, to represent unstructured text and take single word found in the training corpus as features. For the semi-structured data, all the works utilize the HTML structures inside the documents and some utilized the hyperlink structure between the documents for document representation. How to discover interesting and informative facts describing the connectivity in the Web subset, based on the given collection of interconnected web documents has been described by **Madria et al**. [5].
The research on the application of the techniques from machine learning, statistical pattern recognition, and data mining to analyzing hypertext. All the recent advances in content mining research have been reviewed in this paper. **Monika and Mittal** [7] in their paper presented a preliminary discussion of WEB mining, key contributions of computer science in the field of web mining. They also focussed on the prominent successful applications and outline of some promising areas of future research.

Raju and Babu [8]. Outlined three different modes of web mining, namely web content mining, web structure mining and web usage mining. They also presented the significance of introducing the web mining techniques in the area of web personalization which helps in development of business requirements once the goals of business are known.

The analysis presented by **Sakthipriya et.al**., [9] mainly aimed to analyze the web mining categories and its web applications. **Shipra and Pandey** [10] explained in detail on the studies conducted as well as analysis of each web content mining techniques. They concluded that the future scope of web content mining is to predict the user needs to improve the usability and scalability. **Asha and Remya** [11] presented a survey on web mining strategies used for mining different forms of existence of data on Web. They also described emerging techniques used in data extraction from web pages called top-k web pages that describes top-k instances of any particular topic of interest which is very useful in search or fact answering systems.

CONCLUSIONS

Use of data mining techniques for meaningful interpretation of data available on the web documents, hyperlinks web sites, etc., is commonly referred as web mining. Various researchers have indicated the significance and importance of application of data mining techniques in web mining. In this paper a comprehensive review of different web mining techniques, strategies adopted by different researchers in extracting the information from web data and emerging techniques used in data extraction from web pages are being discussed.

REFERENCES

[1] Borges, J and Levene M.1999. Data mining of user navigation patterns. In Proceedings of the WEBKDD'99 Workshop on Web Usage Analysis and User Profiling, August 15, 1999, San Diego, CA, USA, pages 31-39.
[2] Cooley, R., B. Mobasher, and J. Srivastava.1997 Web mining: Information and pattern discovery on the

world wide Web. In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997

[3] Oren Etzioni. 1996.The world wide Web: Quagmire or gold mine. Communications of the ACM, 39(11):65-68, 1996

[4] Kosala, R. H. Blockeel.2000. Web mining Research: A Survey in SIGKDD explorations. Volume 2 (1): 1-15.

[5] Madria,S.K., Bhowmick, S.S., W.K.Ng, and E.P.Lim.1999.Research issues in Web data mining. In Proceedings of Data Warehousing and Knowledge Discovery, First International Conference, DaWaK '99, pages 303-312, 1999

[6] Spiliopoulou. M.1999. Data mining for the Web. In Proceedings of Principles of Data Mining and Knowledge Discovery, Third European conference, PKDD'99, P 588- 589

[7] Monika Yadav and Pradeep Mittal. 2013. Web Mining: An Introduction. International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 3, 683-687

[8] Raju, Y and Suresh Babu. 2015. A Novel Approaches in Web Mining Techniques in case of Web Personalization. International Journal of Research in Computer Applications and Robotics. Volume 3. Issues 2, 6-12.

[9] Sakthipriya, C. , Srinaganya, G,. J. G. R. Sathiaseelan 2015, An Analysis of Recent Trends and Challenges in Web Usage Mining Applications. International Journal of Computer Science and Mobile Computing Vol. 4, Issue. 4, April pg.41 – 48 1

[10] Shipra Saini and Hari Mohan Pandey. 2015. Review on Web Content Mining Techniques. International Journal of Computer Applications Volume 118 – No. 18,33-36.

[11] Asha Joy and R.Remya. 2015. Techniques for Web Mining of Various Forms of Existence of Data on Web: A Review. International Journal of Advance Research in Computer Science and Management Studies. Volume 3, Issue 1, 279-281.

[12] Dr.Osmar R. Zaiane, 2001 in Web Mining: From Concepts to Practical Systems.