ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Review of web usage of data mining in web mining

Arjun Sidana
M.Tech Scholar
University College of Engineering
Punjabi University, Patiala

Dr. Himanshu Aggarwal
Professor
University College of Engineering,
Punjabi University, Patiala

***Abstract*—**The WWW (World Wide Web) contain a huge amount of data that is rising in both dimension and volume day by day. Data mining process has been in use in almost every field of business. Nowadays, various data mining processes use web mining techniques for discovering the valid, novel, understandable and useful data. Web Mining can be classified into three major categories including the web content mining, web structure mining and web usage mining. Web usage mining is an effective approach for discovering the relevant and useful information through data preprocessing, pattern discovery and pattern analysis. There are various web mining techniques available but suffer from many privacy issues. In this paper, we will explore the various web usage mining algorithm used in data mining. The review of web mining research will help for the further research in the same field.

***Index Terms*—**Data mining, Web mining, Web Usage Mining, pre-processing, Pattern discovery, Pattern analysis.

I. INTRODUCTION

A tremendous revolution has been experienced in the information availability and exchange through the internet in the last decade. It is apparent from businesses and organizations collecting data and information related to their operations via the internet [1]. While more efficient and effective means of storing, extracting and manipulating was the main focus of the database technologists, the machine learningcommunity concentrated on creating techniques for acquiring and learning knowledge from the data [2].

Web Navigation is the process in which network of information resources i.e. the WWW (World Wide Web) is navigated. The information in the WWW is organized as hypermedia or hypertext. In this way, every website developed and implemented for the users have some form of navigation [1]. Unfortunately, navigation for every website isn't much good. Web designers who develop websites are likely to develop an amazing website but lack to build awebsite from user's point of view [3]. Thus there is huge requirement to check the navigation-related issues to make the user's experience satisfactory. Web usage mining is an excellent approach to resolving the issues of website navigation.

1.1. DATA MINING

Data mining is an interactive and iterative discovery process [6]. The main purpose of the data mining process is to extract changes, patterns, anomalies, associations, statistically important structures from themassive amount of data. Along with this, the mined outcome must be useful, valid, understandable and novel [4]. These qualities mentioned here are very significant as described below:

Valid: It is critical that the rules, information, patterns and models which are discovered through the data mining process are valid. All these elements must not only valid in the data samples that are already examined, but in the future new data Samples as well. If the above condition is achieved, only then the models and rules can be considered meaningful [4].

Novel: It is necessary that all the data, models and rules generated through the data mining process are new and experts have only a little or no knowledge about them. Otherwise, they would develop little or no new understanding of the problems and data samples.

Useful: It is vital that models, pattern, and rules developed are helpful to take some significant action. In simple words, they should be useful to direct to some action which will yield positive results in the future [8].

Understandable: Understanding the discovered patterns, rules and models are highly essential. Therefore, they should be easy enough that anyone can understand and analyze them for solving the problem [15].

The figure 1. Shows the data mining process that clearly demonstrates that how data is selected. After the data selection, pre-processing and cleaning of data help to obtain the target data [5]. Target data is extracted for the feature selection and turned into the processes data. Then some data mining algorithms are used to convert it into the transformed data. The last phase of the data mining process helps to detect the patterns through the interpretation evaluation [4].

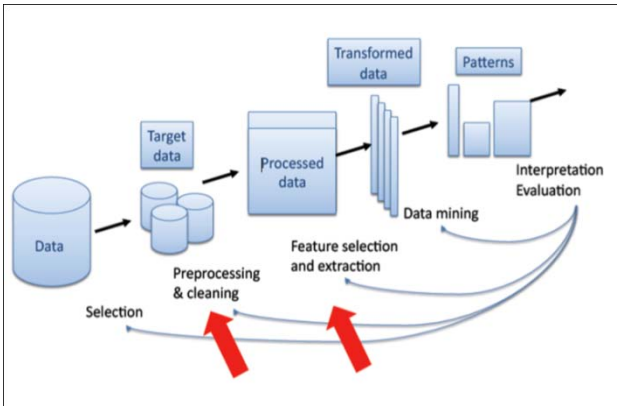

Figure 1: Data Mining Process

But there are some challenges to the data mining process. These includes:

1. Understandability of the discovered pattern.
2. High dimensionality and ahuge amount of data sets.
3. Incomplete non-standard data and data integration is difficult.
4. The issue in assessing the statistical significance of the discovered data.
5. A Large amount of redundant data as an outcome of thesearch. [14].

At present, most of the data mining applications make use of the web mining. Web mining is a technique which is rising with high speed. It is an application of the data mining techniques. In the next section, the details about the web mining and how it is used in the data miningis explained [6].

1.2. WEB MINING

Web mining is one of the parts of the Data mining that applies mining techniques on the data generated and residing on the web to discover unknown useful and interesting patterns. Web usage mining providing the support for the website design, business making decision, personalizing the server and much more [5]. Web mining applies data mining, chart technology and artificial intelligence for serving the users in a better way. It instantly becomes one of the most important areas in the information and computer science. This is because of applications in CRM, information retrieval and filtering, e-commerce, web analytics and Web information systems [7]. Web mining can be classified into three broad categories including:

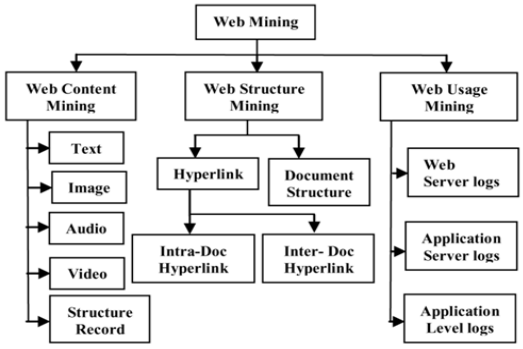

Figure 2: Categories of Web Mining

Web structure mining: It is used to discover valid and useful information using the hyperlinks. These hyperlinks demonstrate the structure of the web. The hyperlink is a link to the web page that refers to the same or another web page. The web structure mining helps to calculate the significance of these web pages which could be said as its most popular application. Google search engine which is used for search results is an example of the web structure mining [11].

Web Content Mining: It extracts the useful information from the content of web pages. Structured data extraction and the text extraction are the two main categories of the web content mining. In structured data extraction, some fixed templates are used by websites to show that important information is extracted from their database. By finding the repeated patterns in web pages, these fixed templates can easily recognize. Along with the structured data, there is some text which is written in the natural language refers to the unstructured data.

Web usage mining: It is used to capture and model profiles and behavior pattern of users who use the website [9]. The organization and structure of the website can be improvedto a great extent by using such patterns to identify the behavior of diverse user segments. Unlike the web content mining and web structure mining, web usage mining use access logs of users to discover the patterns [12].

In the next section, we will review the complete process of web usage mining and how it is used along with data mining.

1.3. WEB USAGE IN DATA MINING

Web usage mining in data mining is used to discover some important patterns from the web pages that can be used to improve the organization of the website. The discovered data provide different paths to access the web pages. The web server automatically collects the data into the access logs. There are three steps in web usage mining that helps to navigate the web in an effective way [10]. These steps include the preprocessing, pattern discovery and pattern analysis. All these steps are describedin detail in the following section:

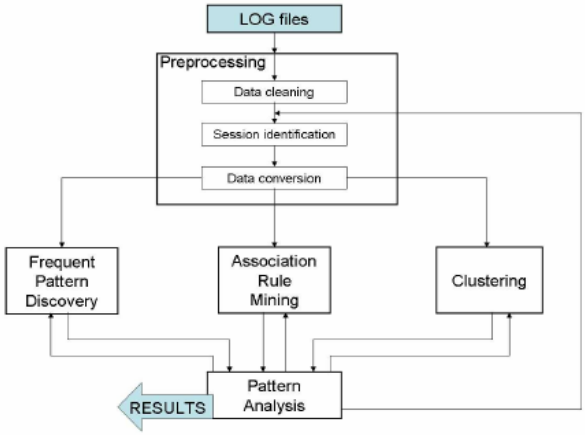

Figure 3: Process of Web Usage Mining

1.3.1. Pre-Processing

Data discovered through the data mining process is abundant. Along with this, the data is noisy and bulky. It includes the data and information regarding each source such as video, image, web page, etc. It is difficult task to identify the useful and relevant data from such huge amount of data without the pre-processing stage [9]. Therefore, pre-processing allows identifying the relevant data and organizing itin terms of sessions and users [3]. The main steps in the pre-processing includes:

- Cleaning the data
- Session identification
- Identification of userand
- Path completion.

1.3.2. PATTERN DISCOVERY

The frequent patterns from the discovered data are identified through the pattern discovery step of web usage mining. The

user clicks numerous hyperlinks for accessing its required data. With pattern discovery, the sequence of those clicks is identified to gather information about the user interest [3]. Even if the pattern are constrained by asession of therequest, threshold and time, pattern discovery can easily recognize them. The pattern discovers some algorithms and patterns which are discoveredfrom numerous fields including data mining, pattern discovery, statistics and machine filtering. Numerous techniques involved in pattern discovery are clustering, classification, statistical analysis, frequent itemset mining, and sequential analysis [7].

1.3.3. PATTERN ANALYSIS

The main requirement of pattern analysis is to filter out the unwanted patterns, models, and rules from the discovered dataset [13]. Visualization technique, usability analysis, Data & Knowledge technique and OLAP techniques are some common techniques used in the pattern analysis for analyzing the discovered data [3].

II. LITERATURE REVIEW

S. Sharma and S. S. Lodhi (2016) proposed a web mining techniqueto extract information from the web data stream. The study identified the issue of discovering the relevant information from the abundant information present on the web. The main problem is in identified how to process the raw data to gather information regarding the website use and filtering the search results in order to present only rules and patterns.To mitigate the above-mentioned issues, the data mining method based on the decision tree algorithms is proposed in this paper. The algorithms were developed for mining the weblogs temporarily. The proposed method was able to provide useful information for generating the log files and extracting information, rules, and patterns from the web data stream [1].

S. S. Patil and H.P. Khandagale **(2016)** reviewed the issue of huge data content on the WWW. A large amount of data present on the web makes it difficult to retrieve some useful information in an easy way. Therefore, the paper presented the standard way for aweb developer to identify the user behavioron the server side by accessing the log files.The web log files containing the information about the website navigation are obtained and for proposing the web mining technique. The predictedbehavior of usage was helpful to users to provide efficiency and automated updated links for reducing the time of developer [2]

C. Bull et.al **(2015)**provided an open sourceSoftware Architecture for Mental Health Self-Management (SAMS) framework. The main purpose of the SAMS is to deliver the non-invasive method for secure storage, analysis of an individual's computer usage to help detection of cognitive decline and large-scale collection. The framework allowed the evaluation and study by medical professionals in which data and textual features can be linked to deficits in cognitive domains that are characteristic of dementia. This research was entirely based on the evaluation of the medical professionals that linked both the data and text to shortages in cognitive domains of the dementia features. Along with this, the paper discussed the concern in the previous research and focussed on the implementation of text and collection components [3].

P. Ristoski and H. Paulheim(2016)defined the 'Data Mining, and Knowledge Discovery in Database (KDD)' methods with a large amount of data. This method performed by applying many approaches that integrated Semantic Web data and Knowledge discovery process. This research analyzed an application domain in biomedicine and life science. The linked open data (LOD) was used for build content-based recommender systems, but still,the KDD methods were unblocked to be used. An example was also included on how to use the Linked Open data. The paper providesan overviewof the approaches used in Knowledge discovery process through different stages [4].

A. Raiyani and Prof. S.S. Pandya (2013) provided an overview of the complete pre-processing techniques including the data cleaning, session identification, user identification activities. The paper introduces a new technique called as Distinct User Identification (DUI) which is based on Agent and session time, IP address and the preferred page on specific time. The DUI (Distinct User Identification) algorithm has been processed which get the efficient result of log files while web usage mining procedure. The technique helps to mitigate the issues of user security by encountering fraud detection, detection of regular user access behavior, terrorism, detection of unusual access to the confidential data, etc. thus the proposed method enhance the performance and designing of upcoming access of pre-processing step outcome [5].

P.Sukumaret, al (2016) investigated various data mining techniques and classified them into theeffectiveway and with limitations. The effectiveness and limitations of all the reviewed algorithms are explained in depth that helps to identify an effective technique for the data mining. The web contained a large amount of data, which enhanced the volume and dimension of data day by day. In fact, it was a 'Web Structure Mining' method. Several heuristic and pre-processing algorithms were applied by using programming language. The pre-processing of data was used to parse the 'Raw Log Files' that involved splitting of the log files and then cleansed to obtained the higher quality of data. The unique users can be then identified using the obtained data. Even session identification is possible when the user identification is successful. However, there was a lack of accuracy for getting the better result in the collected data [6].

B. K. Malviya and J. Agrawal (2015) provided a glance of numerous application of web usage mining. The paper defines that how web usage mining become a dynamic region of study in data mining. defined the Web Usage Mining technique related to the log data files that extracted user's performance, that they used with different applications such as Pre-fetching, Modified services, E-Commerce, etc. Web log data usually has confusing and noisy behavior, but yet pre-processing and pattern analysis technique was needed to be determined. While doing the web usage mining with the weblog data files, there were various problems occurred like desirable information not found, personalized of data was not analyzed and didn't found

the associated information of the gathered data. All the stages in the web usage mining are described in detail along with the problems encountered by various researchers in the same field

[7]. ***Sunena and K. Kaur (2016)*** summarized the concept of data mining. For this, data type generalization and various mining algorithms were compared based on the type of data and type of application. Most of the content of the web mining and its categories were based on the internet. It exposed the comparison between the pattern discovery technique by using its parameters and provided web services to the user. A large amount of data was added to the source in every second, but still, there is a needto discover more the web service related to the user. Along with this, the comparison between the pattern discovery techniques has been provided[8]

III. FINDINGS

The findings table include all the relevant information obtained from the research review. We researched many algorithms, techniques, methods and results obtained from all of them. Till now, many types of researchhas been done on the web usage mining techniques. Many positive results have been obtained by implementing these different web mining algorithms which help businesses in structuring and organizing the website. The following table provides insightinto the findings from the previous research in the data and web mining.

Year	Researchers	Algorithm /Method	Input	Results
2016	S. Sharma and S. S. Lodhi	Decision Tree Algorithm	Web log files	The proposed web mining method withstand with all the applied input Parameters.
2016	S. S. Patil and H.P. Khandagale	Various web mining algorithm	Website navigation log files	The proposed method help to identify the user pattern in an effective way and reduces the time expended by the developer.
2015	C. Bull et.al	Software Architecture for Mental Health Self-Management	Hardware log files	The proposed method allow asecurestorage and large-scale data collection.
2016	A. Raiyani and Prof. S.S. Pandya	Reviewed Knowledge Discovery in Database	-	provide information about approaches used in Knowledge discovery process through different stages
2013	A. Raiyani and Prof. S.S. Pandya	Distinct User Identification (DUI)	User Log files	Proposed method is effective for fraud detection and unusual, suspicious activities on the secure data.
2016	P.Sukumar et, al	Investigate and implemented various web mining algorithm	Raw Log Files	Provided benefits and limitations of the reviewed algorithms.
2015	B. K. Malviya and J. Agrawal	Studied various web usage mining techniques	-	Described the whole web mining proves and the problems encountered in it.
2016	Sunena and K. Kaur	Compared various data mining techniques based on the type of data an application	-	The comparison is helpful for providing web services to the users.

IV. CONCLUSION

In this paper, we have reviewed various web mining algorithms and techniques that have been used by the previous researchers. In-depth analysis of the data mining and web mining help to identify the benefits and limitations of these techniques. Along with this, we have provided aproper process of web usage mining with three phases including the preprocessing, pattern discovery and pattern analysis. The reviewed research will help us in the further research on the topic.

V. REFERENCES

[1] S. Sharma and S. S. Lodhi, "Development of Decision Tree Algorithm for Mining Web Data Stream," International Journal of Computer Applications *(*0975 – 8887*)*, vol. 138 *(2)*, March (2016).

[2] S. S. Patil and H.P. Khandagale, "Enhancing Web Navigation Usability Using Web Usage Mining Techniques," International Research Journal of Engineering and Technology *(*IRJET*)*, vol. 4 *(*6*)*, June (2016).

[3] C. Bull, D. Asfiandy, A. Gledson, J. Mellor, S. Couth, "Combiningdataminingandtextminingfordetectionofearlystage dementia: the SAMSframework," LREC Workshop*:* RaPID-Portorož Slovenia, (2016).

[4] P. Ristoski and H. Paulheim, "Semantic Web in Data Mining and Knowledge Discovery: A Comprehensive Survey," Preprint submitted to Journal of Web Semantics, (2015).

[5] A. G. Raiyani and Prof. S. S. Pandya, "Discovering user identification mining technique for preprocessed Web Log Data," Journal of Information, Knowledge, and Research in Computer Engineering, ISSN: 0975 – 6760, Vol. 2 (2), (2013).

[6] P.Sukumar, L.Robert and S.Yuvaraj, "Review on Modern Data Preprocessing Techniques in Web Usage Mining (WUM)," International Conference on Computational Systems and Information Systems for Sustainable Solutions, 978-1-5090-1022-6/16/IEEE (2016).

[7] B. K. Malviya and J. Agrawal, "A Study on Web Usage Mining: Theory and Applications," International Conference on Communication Systems and Network Technologies, 978-1-4799-1797-6/15/IEEE (2015).

[8] Sunena and K. Kaur, "Web Usage Mining-Current Trends and Future Challenges," International Conference on Electrical, Electronics, and Optimisation Techniques (ICEEOT), 978-1-4673-9939-5/16/IEEE (2016).

[9] P. Lopes and B. Roy, "Dynamic Recommendation system using web usage mining for e-commerce users," Elsevier, International Conference on Advanced Computing technologies and applications (ICACTA), (2015).

[10] N. Kaur and H. Aggarwal, "Web Log Analysis for Identifying the Number of Visitors and their Behaviour to Enhance the Accessibility and Usability of Website," International Journal of Computer Applications (0975 – 8887), vol. 110 (4), (2015).

[11] J. Srivastava, R. Cooley, M. Deshpande and P. Tan, "WebUsageMining," ACM SIGKDD Explorations Newsletter, vol. 1, no. 2, p. 12, 2000.

[12] M.Alphy and A. Sharma, "Study on online community user motif using the web usage mining," Journal of Physics: Conference Series, vol. 710, p. 012015, 2016.

[13] H. Wang, C. Yeng,and H.Zeng, "Design and Implementation of a Web Usage Mining Model Based On Upgrowth and Preflxspan," vol. 6:. No.2, 2006.

[14] Adeniyi, Z. Wei and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method," Applied Computing and Informatics, vol. 12, no. 1, pp. 90-108, 2016.

[15] M. Thilagu, & R.Nadarajan (2012). Efficiently Mining of Effective Web Traversal Patterns with Average Utility. *Procedia Technology*, *6*, 444-451. doi:10.1016/j.protcy.2012.10.053

