International Journal of Engineering & Technology, 7 (3.20) (2018) 690-694

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research pape

Review of Current Trends in Web Usage Mining

Mantri Gayatri[#1], Dr.P.Satheesh[*2], Dr.R.Rajeshwara Rao[#3]

[1-3#]CSE Department,
[1]MRCET, [2]MVGR College of Engineering, [3]JNTUK UCE
Hyderabad Telangana State. Address Including Country Name
[1] gayatricse312@gmail.com
[2] MVGR College of Engineering,Vizianagaram,India
[2]patchikolla@yahoo.com
[3] JNTUK UCE,Vizianagaram,Andhra Pradesh,India

Abstract

Due to incessantly growing of amount of data that is published over web pages, the world wide web (WWW) has impacted every facet of our lives. Due to the increase in size and also the complexity of web, it is necessary to search and retrieve information in an effective and efficient manner. The data in the web pages are semi-structured and heterogeneous making it difficult to access and a challenging issue. In web usage mining, knowledge is discovered effectively from usage patterns to serve the needs of various users. This paper offers a brief study and study of numerous techniques used for web usage mining

***Keywords*:** Data Mining basics, Web Mining, Web usage mining.

1. Introduction

In Today's world, there is a rapid increase in the usage of Internet applications in day to day life and it grows significantly and steadily day by day, thereby distressing the lives of people in almost all the sectors like health, education, business etc. The web applications are gaining more popularity in the present scenario due to the contributing factors like expediency and flexibility of services provided by web applications. Web applications could able to work with a huge data which consistently consists of various user operations, transactions and user activity logs.
In order to enhance the decision making process , the framework of Knowledge Discovery from Databases (KDD) [1] have been used and various people have conducted many experiments to discover the various ways of retrieving possibly useful information which is embedded in large databases. The main process of KDD, is called as data mining, and its main work is to retrieve the frequent patterns which includes association rules and sequential patterns mining. Web mining is one of the application of mining the data particularly on web data [2].
Many of the researchers are involved in mining the data due to the tremendous increase in the growth of the information sources available on the web and also ecommerce. According to the authors Madria, et al. [3] and Borges and Levene [3] , Web mining have been categorized into three broad areas of interest namely : Web content mining, Web structure mining, and Web usage mining. The above three mining tasks can be used in isolation or it can combined with other tasks since they might contain the links of the web document.
Web mining is categorized under data mining technique and the main purpose of which is to retrieve and extract the information from numerous documents and web services mechanically. The main aim of data mining is to the retrieve the necessary and exciting patterns from a collection of enormous data sets in the current trend as well as used in the typical data mining. Web mining uses big data as the data set from which it tries to retrieve the data. Web data typically consists of various profile, structure, documents, information etc.
Web mining is broadly based on two major concepts namely process-based and data-driven. In Web mining we mainly try to extract knowledge [4] from the web. The steps involved in web mining involves: collection of data, data selection before processing, knowledge discovery and analysis of data [5].

The remaining section of the paper is organized as follows. Section 2 discusses the related work carried out and section 3 discusses the types of web mining. Web usage mining and its processing steps are given in Section 4. Section 5 describes the various steps involved in web usage mining. Section 6 provides the details about the algorithms used in web usage mining. Conclusions are given in Section 7.

2. Related Research

Koutri, Avouris, and Daskalaki [6] have presented the comprehensive survey on web usage mining. Pierrakos et al. [7] used web usage mining in the personalization process. Web mining for personalization has been discussed by Eirinaki, and Vazirgiannis [8]. Kosala and Blockeel [9] he explained the web mining in detail and the realted research work has been discussed. Cooley et al. [10] implemented a prototype called as webminer based on web mining using various web mining tasks.He applied cluster analysis on association rules and sequential pattern discovery. Zaiane [11] implemented the OLAP technique on web mining. He also worked on multimedia data which helped a lot in content mining.

Spiliopoulou [12] discussed about the various conceivable applications in the area of web usage mining .Lee and Liu [13] created an environment called as intelligent Java Development Environment (iJADE) which is used in the field of e-commerce. This application is extended to other applications rather than e-commerce. Mobasher et al. [14] used association rule discovery from usage data. For effective web personalization and it is scalable.

Reddy et. al. [15] he worked on data pre-processing model and this model is good for cleaning the data, session record and unique users. Though the model seems to work good but still it has drawbacks like data quality is not good, user identification, session identification is not accurate and the results are applied to discover the patterns.

Chintan R. Varnagar et. al. [16] proposed the architecture which used the log data of the client and the server side data.Future work is that efficient and better results which can match with the empirical observations could be built. Brijesh Bakaria et. al. [17], proposed a survey paper which discusses the various techniques in user identification and he added that there is no proper solution foe user identification.

Liu Kewen [18], in his paper discussed about the algorithm clearly for data cleaning but the problem of user identification is not discussed. One of the drawbacks is that TB level data need to be handled. Sheetal A. Raiyani et. al. [19], discussed about the DUI (Distinct User Identification) algorithm. The author discussed about the operating system, referer page, website structure, edition of browser, Ip address etc. The algorithm could identify the users as well as the sessions.

V. Sujatha et. al.[22],discussed about the algorithm which is based on Pattern using Clustering & Classification (PUCC), This algorithm separates the probable users from other users. Suneetha and Krishnamoorthy (2010) [23] identified the users who are interested by c4.5 algorithm and decision rules have been used. The drawback is that the details of network robots are completely ignored.

3. Types of Web Mining

Web mining can be generally divided into three categories, as seen in Figure 1:

Web content mining
Web content mining is the process of repossessing the most vital contents from the web pages and it encompasses exploring the web documents. This content can include image, text, video, sound etc. Research on text mining has been carried out by many researchers.

Web structure mining
It is the process of exploring the structural information from the different available documents. It is usually represented as graph which consists of nodes and the links between the documents are the edges.

Web usage mining
In order to understand and meet the needs of the user, Web usage mining is used to discover the patterns from the web and it involves applying the data mining techniques on the web to extract the patterns.it mainly deals with the data related to the web users and their usage. There is no restricted boundary between the mining groups. We can combine the above mentioned techniques to achieve good results.

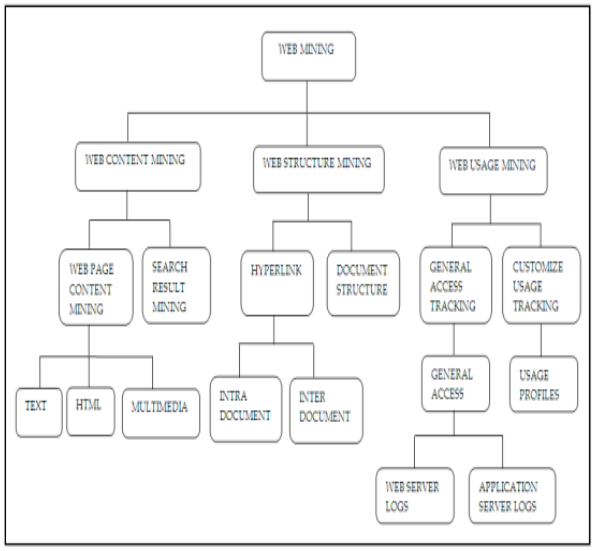

Figure 1: Classification of Web mining

4. Web Usage Mining

Web usage mining mainly puts emphasis on forecasting the behaviour of the customer every time the customer connects with the web. The data obtained from this mining is only the secondary kind of data obtained from the web as a result of user interactions with the web. Generally the data in this mining vary widely but it classifies the data based on the usage of the data in various residents like servers, proxy servers and web clients. The web usage mining is carried out in three steps: data preparation, pattern discovery and pattern analysis phases (See figure 1, Mobasher et al. [14]).

4.1 Data Collection:

The first and foremost work is to collect the data from the web. In this step the relevant and important data are collected from the various data sources. The web log data are pre-processed here to classify the various users, sessions, page views etc. The usage data of the web is here mapped into the relational tables. The main sources of data is classified into three: client data, server side data and intermediate data.

a) Server data: Web servers are the main source of data and data is being collected from web servers.it includes cookies, log files etc.

i) **Server log files:** Logs play an important role. It is the main source for the web usage mining . The server consists of various logs like
File
Protocol
Common Log Format
Remote host
Date
Base Url

ii) Cookies: These are strings which are transferred from web server to the browser of the client. Every time the user visits the site, the cookies are saved in the form of a text file in the browser. Data such as information about all the users i.e., visitors, the pages visited by them, any products purchased, etc. in the cookie log saved within the client's machine.

iii) Explicit user input: These are the data which are provided by the user to the server during the time of user filling the registration pages.

Client data: These data are being collected from the host where the client uses and accesses the website. It is the secondary source of data used for web usage mining.

Intermediate Data:
These are the data composed from proxy servers or packet sniffers.

5. Steps in Data Mining:

5.1 Data Preprocessing

After collecting huge amount of data, the next step is to pre-process the data. In order to use the data for pattern discovery it should be consistent and integrated together. the main task of pre-processing is to achieve high quality and accuracy of data. To achieve this we eliminate the irrelevant data from the log file so as to easily access the web log file. As part of this process, identification of user and session is done. The reason is to know that in one session how many times a user who will access the same web page. These sessions obtained are formatted properly so that they can be analysed and useful patterns can be generated [21].

There are four steps involved in the data preparation namely:
1. Data cleansing
2. User identification
3. Session identification
4. Path completion.

1) Pattern discovery

In order to extract the necessary patterns, the pre-treated information is analysed properly in this step. To mine the various patterns, various methods such as Statistical and machine learning approaches are used

Some of the known approaches used are: classification, clustering, path analysis, association rules, sequential patterns and order model discovery. The issues like pre-processing, data quality plays an important role here. The users need to be uniquely identified and their sessions also in the presence of proxy server and caching [22].

2) Pattern analysis

In pattern analysis, after the various patterns of the user are being discovered, the next step is to analyse the patterns for which the techniques and tools needed are discussed and the analysed information should be understandable by the analysts and to get the maximum benefits from these analysed patterns. The other techniques used for exploring the various patterns include database querying, graphics and visualization, statistics and usability analysis.

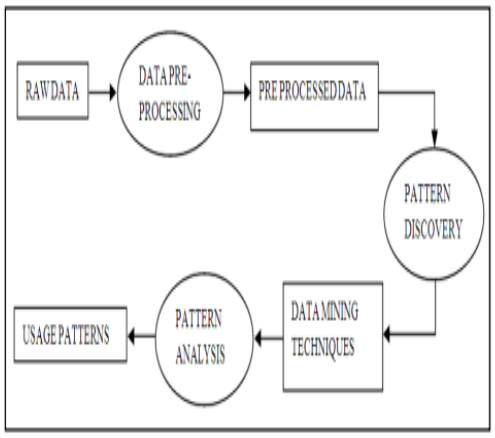

Figure 2: Data mining Steps

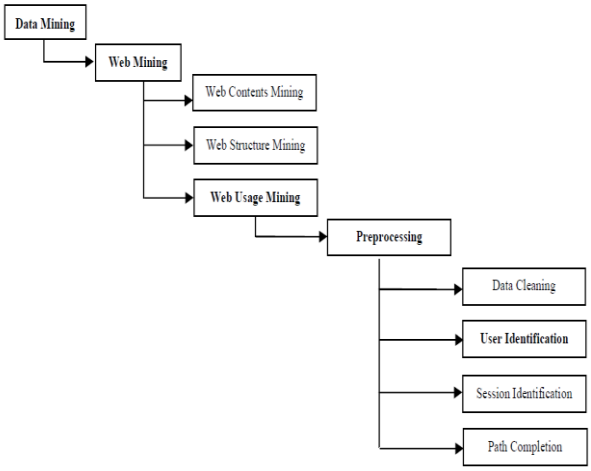

Figure 3: Broad classification of Data Mining

Web usage mining applications could broadly be categorized into two categories: personalized and impersonalized. Personalized involves knowing about a user profile and impersonalized deals with learning the navigation patterns of the user.

6. Web Usage Mining Techniques

In Web usage mining, pattern discovery plays a crucial role. It combines the techniques and algorithms taken from data mining, pattern recognition and machine learning. In data mining, various methods like association rule mining, classification, clustering etc are used for pattern discovery.

Association Rule Mining:
To find which web pages were accessed together on a web, association rules can be generated. In a single server session which pages were accessed together can also be retrieved. In association rule mining, a minimum support is defined which is the count of how many times the web pages were accessed together. If the count is greater than the minimum support then we say that the web pages are frequently accessed together. This helps in pre-fetching the web pages which in turn will reduce the time delay and latency. Accordingly the websites also can be restructured from the access logs. The drawback of using association rules is that if the minimum support is less, than many rules will be generated which may not be relevant.
Many algorithms have used association rule mining in web usage mining, some of them are: In [23] association rules were generated from the scaling Apriori algorithm. This algorithm defines a new protocol suite which also takes confidence as a measure for generating the association rules. In [24] the frequent usage patterns were generated from the log data which were collected in log files. Both Apriori and FP-Growth are combined to find the usage patterns for a particular website. In [25] Apriori algorithm was extended to include set size and frequency for calculating the significant web pages. The proposed method improves the efficiency of the memory.

Clustering Algorithm
In Clustering similar web pages are extracted and grouped together. This helps in knowing which users or group of users have similar navigational patterns while using the web and also we can extract patterns of data also. It also helps in knowing the user demographics with which we can provide personalized web content to the individual users. Clustering is mainly used in search engines and web service providers. There are various kinds of clustering such as incremental clustering, hierarchical clustering and partitional clustering.

Some of the algorithms which have used clustering in web usage mining are: In [26] global Fuzzy-C means algorithm was used for clustering and for optimizing the objective function. This method uses incremental approach and is not dependent on initial conditions. This improves clustering and is obtained through a deterministic global search process. In [27] also fuzzy-C means algorithm was used for clustering. But the main focus was given in increasing the performance of the algorithm. The idea is to choose a good initial centroid cluster which will influence the performance of the algorithm. This paper discusses a way of selecting a good initial centroid and to increase the quality of clustering. In [28] author discussed the clustering algorithms, namely K-Means algorithm and representative object based algorithms are compared. They are analysed based on the number of dataset points and number of clusters.

Classification:
In web usage mining, classification is widely used. It is a supervised learning technique, in which the models are designed to be classified into data classes. It uses various algorithms which act as classifier. The classification techniques can also be used for studying the user-client behaviour and also interesting patterns can be generated. We can classify the relevant and irrelevant links which are visited by a particular user. This can be identified based on the time spent on a particular web page and also number of hits. Few of the Classification based web usage mining algorithms are discussed here:
[29] deals with personalization of various web services. User sessions are separated based on the users access and then these sessions are accessed from the server web log. Two new approaches were defined for web mining. The first method also known as "process centric view" defines web mining as a sequence of tasks and in the second method which is also known as " data centric view" web mining depends on the type of data used. In [30] both temporal pattern extraction and association rule mining are combined to frame the classification framework. This method uses IF-THEN rules which have temporal patterns on left hand side and prediction is done on right hand side. The prediction is done on temporal patterns and important events. In [2] classification is done in three phases: first phase is the training phase which uses labelled records. Second is the test phase which uses unseen labelled records. Final phase is the deployment phase which classifies the unlabelled records.

7. Applications

The main motto of Web Usage Mining is to collect the relevant and interesting details about the navigation patterns of the users mainly to describe the web users. This gathered information provides a base further for improving the website based on the user's point of view.
Log results obtained from the web can be used in various applications:

- User navigation can be improved by pre-fetching and caching
- Improving the websites
- Improving the user needs
- Useful in e-commerce
- Personalization of web content

Personalization of web content
Based on the profiles of the user as well as their behavior different techniques of web usage mining can be applied to personalize the websites. There are many advantages of personalization which includes automating the products for the customers, establishing a profound relationship, used to plan new strategies in the market. It is also used to provide more information about the users which helps them in improving their website designs [27].

E-learning environment
Web usage mining plays a very important role in e-learning environment mainly to keep track of the activities of the websites and tries to gather the behavior and the patterns which further can be used to make modifications to the website. Mining tool allows us to understand the frequently visited pages, links, etc.[14]

Security
Web usage mining can used in finding out the security breaches and it provides the patterns that are applicable and useful in finding fraud, intrusion etc. [28].

Site design support
There are many issues for the designers while designing the website. Web usage mining provides more vital information about the behavior of the users so that they can change the content or redesign the website .It provides more help to the designers and makes their work much easier after tracking the behavior and pattern of the users [28].

Business Intelligence
Web usage mining is good in improving the market strategies and it helps companies to maintain their market place and also good in providing decisions to the companies to improve their performance. It extracts the information about the customer behavior and from the stored database it could be used in the building the company's marketing strategies.

8. Conclusion

This paper has attempted to provide an up to-date survey of the rapidly growing area of Web usage mining, which is the demand of current technology. In this paper a general overview of Web usage mining is presented in introduction section. Web usage mining is used & more research must be made in many areas such as e-Business, e-CRM, e-Services, eEducation, advertising, marketing, bioinformatics and so on. The main techniques for pattern discovery are sequential patterns, association rules, Classification, Clustering, and path analysis. Web usage mining helps in designing the website based on the user navigation tools and they can create a new user page or modify existing ones based on the user experience. Different tools are already available in the market to assist web usage mining. Still more research needs to be taken in this direction to bring out many changes in this area.

References

[1] J.W. Han, M .Kamber, Data Mining-- Concepts and Techniques, Elsevier Science & Technology Books, 2006.
[2] J. Srivastava, R. Cooley, M. Deshpande, P. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, 1(2), 20001, pp.2-23.
[3] S. K. Madria, S. S. Bhowmick, W. K. Ng, and E.-P. Lim. Research issues in web data mining. In Proceedings of Data Warehousing and Knowledge Discovery, First International Conference, DaWaK '99, pages 303–312, 1999.
[4] Hidenao Abe , "Development of a Classification Rule Mining Framework by Using Temporal Pattern Extraction", New fundamental technologies in data mining, January 2011, pp. 493-504.
[5] Bamshad Mobasher .Web Usage Mining .Springer Berlin Heidelberg, 2007, 449-483.
[6] Koutri M, Daskalaki S., Avouris N., Adaptive interaction with web sites: an overview of methods and techniques, Proc. CSIT 2002, Patras, September 2002
[7] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C. D. Spyropoulos, "Web usage mining as a tool for personalization: A survey," User Modeling and User-Adapted Interaction, vol. 13, pp. 311-372, 2003.

[8] Magdalini Eirinaki and Michalis Vazirgiannis, “Web mining for web personalization”, ACM Transactions on Internet Technology, 03(01):1-27, February 2003.
[9] Kosala and Blockeel, “Web Mining Research: A Survey”, SIGKDD Exploration, Newsletter of SIG on Knowledge Discovery and Data Mining, ACM, Vol.2, 2000
[10] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: information and pattern discovery on the World Wide Web," in Proc. Ninth IEEE International Conference on Tools with Artificial Intelligence, 1997, pp. 558-567.
[11] O. R. Zaïane, "Web usage mining for a better web-based learning,"presented at the Conference on Advanced Technology for Education,2001.
[12] Pawel Matuszyk, Georg Krempl, Myra Spiliopoulou, Correcting the Usage of the Hoeffding Inequality in Stream Mining,published in 2013 IDA
[13] R. S. T. Lee, and J. N. K. Liu, iJADE Web-Miner: An Intelligent Agent Framework for Internet Shopping, IEEE Transactions on Knowledge and Data Engineering, 16(4), 2004, 461- 473.
[14] B. Mobasher, R. Cooley, and J. Srivastava, J. Automatic personalization based on Web usage mining, Communications of the ACM, 43(8), 2000, 142-151
[15] K. S. Reddy, G. P. S. Varma, and S. S. S. Reddy, "Understanding the scope of web usage mining & applications of web data usage patterns, “in Proc. International Conference on Computing, Communication and Applications, 2012, pp. 1-5.
[16] Chintan R. Varnagar; Nirali N. Madhak; Trupti M. Kodinariya; Jayesh N. Rathod,” Web usage mining: A review on process, methods and techniques”, IEEE International Conference on Information Communication and Embedded Systems (ICICES),pp,40-46,2013
[17] Brijesh Bakariya, Krishna K. Mohbey and G.S. Thakur, “An Inclusive Survey on Data Preprocessing Methods Used in Web Usage Mining”, Springer-2011.
[18] Liu Kewen, “Analysis of Preprocessing methods for web usage mining”, International Conference on measurement, Information and Control, IEEE, 2012.
[19] Sheetal A. Raiyani, Shailendra Jain and Ashwin G. Raiyani, “Advanced Preprocessing using Distinct User Identification in web log usage data”, ISSN : 2278 – 1021, IJARCCE, Vol. 1, Issue 6, August 2012
[20] MuhammedShafi. P,Selvakumar.S*, Mohamed Shakeel.P, “An Efficient Optimal Fuzzy C Means (OFCM) Algorithm with Particle Swarm Optimization (PSO) To Analyze and Predict Crime Data”, Journal of Advanced Research in Dynamic and Control Systems, Issue: 06,2018, Pages: 699-707
[21] Selvakumar, S & Inbarani, Hannah & Mohamed Shakeel, P. (2016). A hybrid personalized tag recommendations for social E-Learning system. 9. 1187-1199.
[22] V. Sujatha and Punithavalli, “Improved User Navigation Pattern Prediction Technique From Web Log Data”, ELSEVIER-2012
[23] Suneetha, K. R. and D. R. Krishnamoorthi, "Identifying User Behavior by Analyzing Web Server Access Log File”, (IJCSNS) International Journal of Computer Science and Network Security, VOL.9, No.4, April 2009.