© 2016 IJSRSET | Volume 2 | Issue 6 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

A Review on Finding Users Navigation Behavior Using Web Mining Algorithm

Sana M. Deshmukh*, Krishnakant P. Adhiya

Department of Computer Engineering, SSBT COET, Jalgaon, Maharashtra, India

ABSTRACT

Web mining is the application field of data mining which is useful to extract the knowledge from huge amount of data. So we can use web mining algorithm in understanding users' navigation behavior. In the proposed model first step is the preprocessing of web log data. In a pre-processing step, proposed algorithm will remove nearly all irrelevant entries from web log file. After the data cleaning process, user identification step will be applied and followed by the sessionization with different time limits. After pre-processing, web mining algorithm will be applied to study the user navigation behavior.

Keywords: Web Usage Mining, Pre-Processing, Web Log File.

I. INTRODUCTION

Web mining is the integration of information collected by traditional data mining methodologies and techniques with the information gathered from the World Wide Web [1]. It is used to understand the user's navigation behavior and also to evaluate the efficiency and effectiveness of the website. World Wide Web there is a lot of information available so information should be mined according to the interest of user. Otherwise it will make the trouble for user to get the related information that he/she wants.

World Wide Web is a huge, interconnected, semi-structured, widely distributed, highly heterogeneous and hypertext information repository. The web continues to grow at an incredible rate as an information gateway. Web mining technologies are the proper solutions for knowledge discovery on the web. Web mining is the application field of data mining techniques to discover patterns from the web [3]. Web mining is classified into three categories (Figure 1).

Web Content Mining: Mines the data Sources on the web and extract the patterns from web data sources.

Web Structure Mining Mines the structures on the Web and use linkage information to improve search engines.

Web Usage Mining Mines the usage Patterns on the Web and improve web usability and user experience.

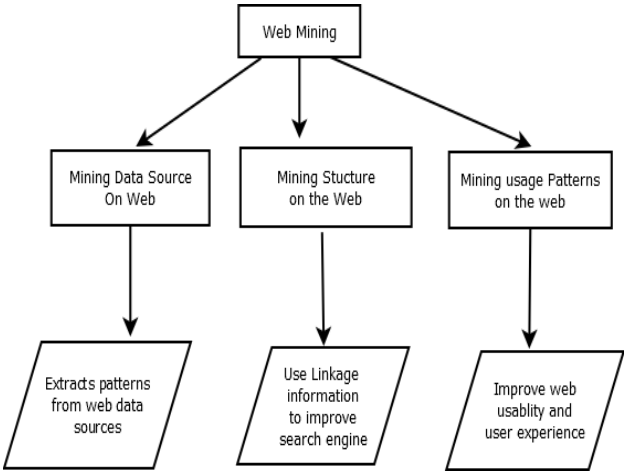

Figure 1. Structure of Web Mining

To understand user's navigation behavior and area of interest recent researchers are focusing on the web usage mining in the proposed method also the web usage mining is used to extract the usage patterns.

Web Usage Mining

Web usage mining is used to extract the usage patterns from web data. It is used to improve the user's web search experience and also improve web usability. Web usage mining has three different tasks to perform. Those three processes of web usage mining are as follows:

1. Pre-processing
2. Pattern discovery
3. Pattern analysis

Web usage mining is the process which extracts the useful usage pattern. In recent search web, usage mining is used to understand the user's interest in web search. That's why it is applied on web log file. Web log file contains the user history and previous access records. To extract user's behavior web usage mining is used.

Pre-processing is the first step of web usage mining process that performs cleaning of web log file. In this process, unnecessary weblog data or noisy data from web log is cleaned and also log file size is reduced.

Pattern discovery is the Second step of web usage mining process of pattern discovery that performs on cleaned log file generated in the prepossessing step to discovering the web patterns.

Pattern analysis is a final step of web usage mining. It processes on pattern discovered in the second step for further analyzed to generate more useful information related to the user behavioral pattern.

II. METHODS AND MATERIAL

Literature survey

A. Background

On the web, there is a huge amount of information available. In each and every field there is a lot of data is available. Peoples are using this information or data according to their need. But from that huge data extraction of relevant information is not an easy task. For users help and attraction purpose some mechanism were applied to web search. So the user can get the relevant information as soon as possible. Using those techniques user can fire the query on a search engine and it provides the information according to user's area of interest. Then again some improvement has done on from text search to image search. The search result will be image and query will be text. Then research attracted towards the user's search behavior. If the user's search behavior is known then it will be easy to understand his/her area of interest. So research has grown towards users' navigation behavior. Depending on user's navigation behavior prediction of next search by the user is easy to understand. That's why the search work is done on user's navigation behavior. But still, it was tricky because there are so many users and so many websites and many search engines. Understanding Users' navigation behavior is so complicated due to many websites availability. That's why researchers have thought about the only single website and single user navigation behavior on that specific website. But user navigation patterns have different forms like finding users area of interest, domain, browsing path, page interest estimation of the user, Users frequency of search interest in a web page. Afterward, search also attracted towards a group of people having the same area of interest. Then research on their similarity of visiting path. Now a day's research is concentrating users browsing preferred path. Because the existing techniques are only focusing frequency of user's navigation but they are neglecting user's area of interest on visiting path of browsing preference. In the past research, there are many algorithms used based on pre-processing, pre-processing with fuzzy, pre-processing with ontology, graph-based algorithm, using a decision tree, frequent item set mining, using naive bays, using CART, K-means etc. But in recent work improved clustering algorithms are used for users' future web page prediction.

B. Related Work

In this area, many works has done based on different navigation pattern of the user. Some work has done to understand the domain of user some on next page prediction some on next more than one-page prediction also for the group of people having the same area of interest using the different algorithm. Some authors name with their process an approach is described.

K. R. Suneetha et al. in [1], proposed an approach to a website designer to improve their system by determining occurred systems errors, corrupted and broken links by using the web usage mining. This approach is proposed to find top errors, potential

visitors of the website. NASA web log file is used here to finds the user interest. On that web log file pre-processing has done. Pattern discovery and pattern analysis are the further steps of the method. This method is useful for the every user of the website who was visiting the website. But all the website visitors are not interested in the website. That's why this approach is not efficient.

K. R. Suneetha et al. in [2], proposed an approach for an interested group of the people for the website. This method is introduced for improvement of customer relationship management. In this method NASA, web log file is used also pre-processing is done on it. An enhanced version of C4.5 decision tree algorithm is used to classify the interested and uninterested users of the website. This approach is efficient for a frequent user of the website not for all the users.

V. Sujatha et al. in [5], presents the work on prediction of user navigation patterns using clustering and classification algorithm. Clustering is used to extract the information from weblog in the form of a number of visits made to a single web page, web page traffic and most frequently viewed the page with navigation behavior of the users. But the disadvantage of this approach is that it's required the main document and additional files always requires HTTP requests.

Arbelaitz et al. in [6], shows the frequent pattern mining and clustering to adapting the user's preference and requirements. In this method, collaborative filtering is used to the improvement of users browsing behavior towards the website. The user profile is extracted from web log to provide automatically next links in the future. In this process, two phases are used online and offline phase. In offline phase user profile has created using weblog frequent path mining and predicted links also provided to the user when he comes on that website again. But in this method, they are not focusing on any domain specific information.

Pandey et al. in [7], shows the work on large data files. Because the World Wide Web there is a lot of information is available on the web. So due to large data available on web information overloading and information disorientation problems are occurring. To give the better solution to these problems this approach provides the solution. In this method personalized collaborative filtering is combined with association rule mining and FP-growth algorithm. But association rule mining and up-growth algorithm are not efficient in concern of cost of implementation.

R. Thiyagarajan et al. in [8], presents his work for prediction of the user browsing activity and then recommends the web pages to the user according to the area of interest. Weighted k-means clustering algorithm is used here to predict the users' navigational behavior. Weighted k-means is better than k-means that's why in this method weighted k-means is used. In this approach, hamming distance and cosine function are also used to understand the similarity inactive user session to provide the similarity of users browsing behaviors. But in this method, they are not considering overlapping clusters to generate the user profile.

Ahmad Hawalah et al. in [9], proposed the approach in concern of better providence of users behavior towards the website. Every time it does not happen that user visiting the website for the same area of interest. May be possible that user wants the other information available at the website. That's why there is need of user's behavior changing frequency for better prediction of next page. In the approach, they are concerning about dynamic user profile for personalization. Also, they used the ontology concept for prediction of user's short term and long term search goal. But this method is not fitting to all other approaches.

Meera Narvekar et al. in [10], shows that World wide web so many websites are increasing due to that reason it is very difficult to predict or understand the user behavior towards the website. Because due to large websiteavailability traffic makes the disturbance in understanding user behavior and prediction process. This approach works in the prediction process. Because in the previous method's prediction process requires more training, more time consuming and less accuracy of prediction. To provide the better solution to these problems' this method is introduced in the research area. The proposed method combines the Markov model and hidden Markov model with dempster's rule and also avoids the miss-prediction of web pages. This method is useful but it is working on the 2-tier architecture it can provide better results using 3-tier architecture.

Ravi Khatri et al. in [11], proposed approach to provide the relevant information to the user in a particular time interval and also provides improved personalized web

recommender model, which considers user specific activities and also considers some other additional factors related to websites. This method considers the factors, total number of visitors, number of unique visitors, numbers of users downloading the data, how much amount of data is downloaded, how much amount of data is uploaded and the number of advertisements for a particular URL to provide a better result. For this purpose, they are using the web recommender model with knowledge discovery process to extracting highest fitness value of URL. This approach is useful but they should consider more factors and parameters so that result can improve and in the concern of cost factor it is not useful.

Prajyoti Lopez et al. in [12], proposed a method for the attraction of new user and to retain the existing users. To achieve this aim approach works on the access pattern of the user. This method provides relevant data to the website user including registered non-registered. This approach uses lexical patterns to generate item recommendation in an e-commerce website. But this approach is just for an e-commerce website.

That's why there is need of an approach which suits to other websites also and to provide the better next web page prediction there is need of users browsing preference path to extract the similarity in web page access of users. According to this proposed solution is attracts towards the prediction based on the similarity of users browsing preference path.

III. RESULTS AND DISCUSSION

Proposed work

A. Approach of Proposed Work

Web user browsing preference path mining algorithm is used to analyze of web log records and finds the user access rules [13]. This algorithm is efficiently used in web personalization. But still, some issues are occurring in the reality.

- In the web browsing preference path, they are only considering the frequency of user access.
- In the reality web, log data shows the massive and distributed properties.

To resolve these issues proposed work combines the clustering algorithm and web user browsing pattern mining algorithm. Pattern mining is used to understand the users browsing preference path and clustering is used to make clusters based on similarity of users access path. This work is based on the three tuple model. In this method Improved Preferred Path Mining Algorithm and Extended Improved Clustering Algorithm is used to find user's similarity on visiting path of website. The proposed work is about understanding the user's navigation behavior based on the similarity access the web page for a website. The proposed solution uses the NASA web log file. First preprocessing will be applied on the web log file. In the pre-processing phase nearly all the irrelevant information is removed. Then less ionization of the user identity is performed. In the less ionization process, it extracts the identity of the website visitor. The proposed approach considers the degree of the page interest, access time of web page, the size of the page and the number of visits. According to these factors, user profiles are considered. Then the clustering algorithm is applied to make cluster based on the users' similar visiting path. After clustering process improved preferred path mining algorithm will be applied to predict the user's next web page. This proposed solution is more effective because in this approach it is not concerning about log file but also using the other users' similar access path.

B. Architecture of a Proposed Work

Proposed method works on users visiting path similarity to provide the prediction of next web page for future. In this method extended improved clustering algorithm is combined with improved preferred path mining algorithm. The architecture of the proposed method is described in the "figure.1", Architecture contains the overall process of proposed method. In this method, NASA weblog is used for the prediction process. But web log file also contains the irrelevant information and media files. Those irrelevant files will be removed in the pre-processing step. After pre-processing web log file will be reduced for the further process. Pre-processing is a combination of data cleaning and user identification processes.

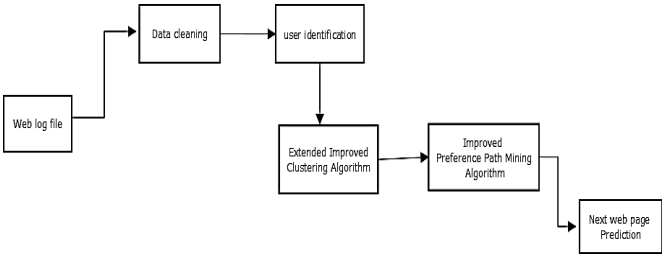

Figure 2. Proposed Architecture

According to the figure, 2 web log file is used to analyze the websitevisitors' area of interest for that website. To remove the irrelevant information and all other media from web log file and to reduce the size of a web log file pre-processing is applied. In a pre-processing first data cleaning is applied to a log file then User identification process is applied to identify the each website visitor so it will help for future web page prediction. After user identification extended improved clustering algorithm and improved preferred path mining algorithm is applied on web log file. At the end next web page prediction will be achieved by using both extended improved clustering algorithm and improved preferred path mining algorithm.

IV. CONCLUSION

This paper focuses on web mining algorithm used in finding the user behavior to provide the future prediction. According to this review, it is concluded that for better web page prediction users visiting path similarity is more attractive than the other strategies. That's why the proposed solution works on users visiting path similarity to provide next web page prediction.

V. REFERENCES

[1]. K. R. Suneetha, Dr. R. Krishnamoorthi, "Identifying User Behavior by Analysing Web Server Access Log File", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, page 327-332 April 2009.

[2]. K. R. Suneetha, R. Krishnamoorthi,"CLASSIFICATION OF WEB LOG DATA TO IDENTIFY INTERESTED USERS USING DECISION TREES", researchgate 2010.

[3]. Naga Lakshmi, Raja Sekhara Rao, Sai Satyanarayana Reddy, "An Overview of Pre-processing on Web Log Data for Web Usage Analysis", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-4, page 274-279, March 2013.

[4]. Xiaojing Li and Yanzhen Cheng, "The Improved Clustering Algorithm for Mining Users Preferred Browsing Paths", D.S. Huang et al. (Eds.): ICIC 2016, Part III, LNAI 9773, pp.329337, 2016.

[5]. V. Sujathaa, punithavalli, "IMPROVED USER NAVIGATION PATTERN PREDICTION TECHNIQUE FROM WEB LOG DATA", International Conference on Communication Technology and System Design 2011, page 92-99, 2011.

[6]. Olatz Arbelaitz, Ibai Gurrutxaga, Aizea Lojo, Javier Muguerza, Jesus M. Perez and Inigo Perona,"Adaptation of the User Navigation Scheme using Clustering and Frequent Pattern Mining Techiques for Profiling", International Conference on Knowledge Discovery and Information Retrieval, KDIR-2012, page187-192, 2012.

[7]. Gopal Pandey, Swati Patel, Vidhu Singhal, Akshay Kansara,"A Process Oriented Perception of Personalization Techniques in Web Mining", International Journal of Science and Modern Engineering (IJISME), ISSN: 2319-6386, Volume-1, Issue-2, January 2013, page 26-30, 2013.

[8]. R. Thiyagarajan, K. Thangavel, R. Rathipriya," Recommendation of Web Pages using Weighted K-Means Clustering", International Journal of Computer Applications (09758887), Volume 86 No 14, January 2014, page 44-48, 2014.

[9]. Ahmad Hawalah, Maria Fasli, "Dynamic user profiles for web personalisation", Expert Systems with Applications 42 (2015) 25472569.

[10]. Meera Narvekar, Shaikh Sakina Banu," Predicting Users Web Navigation Behavior Using Hybrid Approach", International Conference on Advanced Computing Technologies and Applications (ICACTA-2015), pages 3-12, 2015.

[11]. Ravi Khatri, Daya Gupta," An Efficient Periodic Web Content Recommendation Based on Web Usage Mining", 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), page 132-137, 2015.

[12]. Prajyoti Lopez, Bidisha Roy," Dyanamic Recomendation System Using Web Usage Mining For E-Commerce Users"Procedia Computer Science 45-2015, page 60 69, 2015.

[13]. Xiaojing Li, Yanzhen Cheng," The Improved Clustering Algorithm for Mining Users Preferred Browsing Paths", Springer International Publishing Switzerland 2016.