ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

A Study on Online Web Log Prediction System Using web Mining Technologies: a Review

Megha P. Jarkad, Prof. Mansi, Bhonsle

Assistant Professor, Department of Computer, GHRCEM—Wagholi, Pune, India

College-G.H.Raisoni College of Engineering and Management, Wagholi, Pune, India

Department of Computer Science, University of Pune, MH, India.

ABSTRACT: Web Procedure Mining is the characteristic unearthing of user admittance pattern from web.Web usage mining can be defined in simple words as the discovery and analysis of user access patterns through the mining of web log files and associated data from that particular website. The lots of work has done in the field but basically this paper focuses on user future next request prediction using web log data, click streams record and user information. The objective of this paper is to provide past, current evaluation and update in web usage mining- future request prediction. This paper also presents the comparisons and summary of various methods of future request prediction with application, which gives the overview of development in research.

I. INTRODUCTION

Web is the world's largest knowledge warehouse.With the increasing number of users, extracting the knowledge from the web proficiently and effectively is becoming a monotonous process.While surfing user leaves valuable information which is collected manually in web log file.This collected data in web log file is play an important role in finding out user web navigation behavior. With the help of this data we can predict the user's interest and user's future request.This paper presents different methods of user future request prediction.The pros and cons of these methods have also been discussed. The rest of the paper is organised as below. Section 2 presents

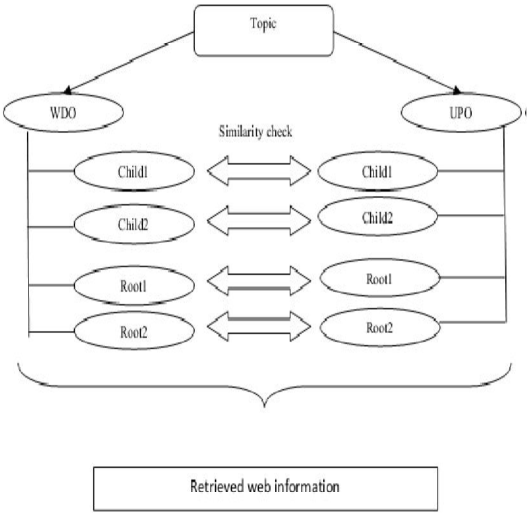
the motivation of paper, Section 3 presents Literature review on users next request prediction, and Section 4 gives the conclusion.

II. MOTIVATION

The world wide web is becoming popular day by day which results in large number of users access the web all over the world. Whenever any user access a website a large volume of information related to that user such as its IP address, requested URL, are collected automatically by servers and saved in access log files as the user may access a same web pages repeatedly.Web access pattern which is nothing but series of all accessed pages play an important role in finding out user behavior.With the help of this behavior we can predict that what will be user future access pattern which will help in reducing browsing time of web and thus reduce load on server as well as save user time. The main objective of this study is to know what research has been done on web usage mining in future request prediction.

III. LITERATURE REVIEW

The focus of literature review is to study and contrast the available technique to predict the users' future movements. AlexandrasNanopoulos et al. [1] focused on "web pre-fetching‟ because of its importance in reducing user perceived latency present in every web based application. As a popularity of web growing tremendously, there is heavy traffic in the internet which led to delay of response. The web servers underheavy load, Low bandwidth, Network congestion, propagation delay and Bandwidth underutilization are the few reasons of delay. One of the way to overcome this is to increase the bandwidth but this is not optimal solution as we also have to consider economic cost.As a result of this a technique was proposed in which the delay of client future requests for web objects is reduced by transferring those objects into the cache in the background before an explicit request is made for them. Order of dependencies between web documents access and interleaving of requests belonging to patterns with random ones within user transactions these are the important factors which affect on web pre-fetching were discussed in their paper.[1] V. Sujatha, Punithavalli [2] proposed the Prediction of User navigation patterns using Clustering and Classification (PUCC) from web log data. The first stage is the cleaning stage where all the web log data is preprocessed, which removes all entries which willhave no use during analysis or mining as well as an unformatted data is converted into the form which can be directly applied to web mining process.This phase also includes session and user identification. In the second stage, 1.all the entries which has accessed robot.txt file are identified and removed.2.All the entries which have visiting time of access as midnight are detected and removed.3.All the entries having access mode HEAD instead of GET or POST are removed.4.All the entries whose browsing speed exceeds threshold T1 are removed.The browsing speed is the ratio of number of viewed pages to session time.The potential users are then identified from others using cleaned data. From the potential users, a graph partitioned clustering algorithm was used to find the navigation pattern. An undirected graph having connectivity between each pair of web pages is used.Weight is assigned to every edge in the graph which is based on frequency and connectivity time.An LCS classification algorithm was then used to predict user future requests.[2]S. Vigneshwari, S. Vigneshwari[3] proposed a model for web information gathering using ontology mining method. Web Data Ontology(WDO) and User Profiling Ontology(UPO) are usd for extracting information from web in the proposed approach. This ontologies stand as the building blocks of the proposed web information gathering model. The web data ontology is considered as the main ontology of the proposed web information gatheringModel.The web ontology is constructed from the web site under consideration.The web sites are gathered and saved for extracting the ontology.Extracting the concepts from web documents is the initial stage of ontology construction. Based on the interrelation ship between the keywords in the web documents the concepts are extracted.Stop words in the web documents are removed in the initial preprocessing step and after that the documents are transferred to stemmer algorithm. The root words of the keywords in each of the document are extracted in stemmer.Duplicate keywords and redundancy is removed from each document.Thus the document contains only unique keywords. The keywords are then subjected to theconcept extraction. The concepts are extracted based on balanced mutual information (BMI) which is also known as semantic similarity measure.Based on BMI value all the concepts are arranged and the result is ontology.The second ontology that is user profiling based ontology is constructed based on user given tags for the web retrieval. The processing of the tags is same as the WDO model.As tags are simple keywords, there is no need for muchpreprocessing to separate the exact contentlike WDO model.In ontology mining method, based on the web log data and the web documents two ontologies are developed. The main feature for the ontology mining method is user given query. The ontology mining mainly focuses onthe interesting measure between the ontologies to collect the information from them.cross ontology similarity measure is used for semantic similarity measure.



Then, the ontology mining technique was devised to extract the information using the interesting measures that considered both the ontologies developed in the previous step. Finally, the experimentation has been carried out using the extracted log data and web documents. The results showed the information extracted from both the ontologies to validate the information provided in the web documents and the user perspective[3]. TeenaSkaria, Prof.T.Kalaikumaran, Dr.S.Karthik[4] proposed an ontology model for gathering web information.Personalized ontologies are constructed using user profiles and concept models.Identifying the user context and to arrange them in such a way that improves the search precision are two major challenges in personalization of information retrieval.Because of this user profiles are developed in hierarchical structure known as ontology for user profiles.Web information is gathered more accurately by clustering. Clustering is done using K-means algorithm.Clustering is performed on the basis of content and locations.A better partition of input data set is done by k-mean algorithm. It is employed to learn the user's preferences and to gather information from web according to preference of users.[4] DilpreetKaur, A.P. SukhpreetKaur[5] proposed KFCM method of fuzzy clustering to predict the user future request.First step is read web log file.Second step is preprocessing where only required attributes are selected from log file like User request,requestmethod,ipaddress,date,time.Irrelevant entries like all the entries having file name .jpeg,.jpg, .gif, robots files, error code,Request method HEAD, POST are removed from the log file and thus cleaned log file is prepared. From cleaned log fileunique users are identified according to unique webpages and IP Address. After user identification, session identification is performed. In session identification step, sessions are identified for all users by taking 30 minute time threshold value.Pagesvisited by user less than or equal to 30 minute time put into one session and another pages which are visited after 30 minute put into another session. unique session id is assigned to all sessions. Third step is clustering in which whole data of web pages visited by each user and user session and user session ids is putted in an array to make a clusters. Thendata is divided into clusters using Fuzzy C-MeansandKernelized Fuzzy C-Means algorithms. Then webpages with highest grade of membership in each cluster are searched. In the fourth step according to grade of membership, weightage is assigned to each webpage, page having low weightage has lowMembership and page having high weightage has high membership. User future webpage is predicted using Fuzzy C-Means and Kernelized Fuzzy C-Means algorithms [5].

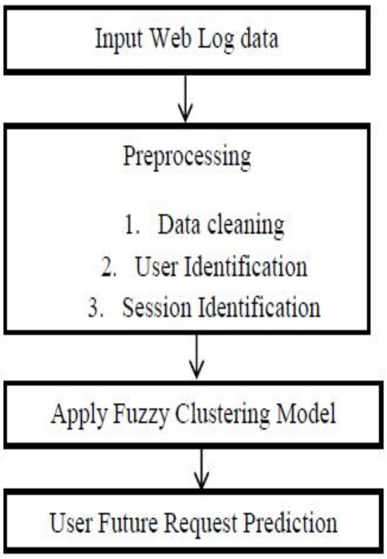

Input Web Log data
Preprocessing
1. Data cleaning
2. User Identification
3. Session Identification
Apply Fuzzy Clustering Model
User Future Request Prediction

Fig. 1 Flow Diagram of Proposed Work

Xiaohui Tao, Yuefeng Li, and NingZhong[6]Proposed an ontology model thatprovides a solution for emphasizing local and global knowledge in a single computational model.In this paper personalized ontologies are used.Ontology model attempts to improve web information gathering performance by using ontological user profiles. User’s local instance repository (LIR) and the world knowledge are used in the proposed model. World knowledge is acquired by people from education and experience; an LIR is a user’s personal collection of information. For analyzing concepts specified in ontologies, a multidimensional ontology mining method is introduced in proposed model. The users’ LIRs are then used to populate the personalized ontologies as well as to discover background knowledge. The model also has extensive contributions to the field of information system,informationretrieval and web intelligence. Userpersonalized ontologies are constructed by extracting world knowledge from the LCSH system and finding out user background knowledge from user local instance repositories. In evaluation stage, for experiments large testedand standard topics were used.[6] Maryam Jafari, ShahramJamali, FarzadSoleymaniSabzchi[7] proposed a novel algorithm named as PD-FARM for FP tree mining process. To find fuzzy association rule the proposed algorithm uses fuzzy FP-tree. For pattern discovery PD-FARM algorithm is used which is based on Fuzzy Association Rules Mining (FARM).The mining algorithm has two phases. In first phase, FP-tree is constructed from database and in second phase frequent patterns are derived from FP-tree. From quantitative data fuzzy association rules are find out by fuzzy FP-tree mining algorithm.The tree structure for frequent fuzzy terms(regions) is generated using fuzzy FP-tree construction algorithm. Quantitative values of attributes in transactions are transformed into linguistic terms. The linguistic term with the maximum cardinality is only used for each term. The frequent fuzzy items are derived from the fuzzy FP-tree and represented by linguistic terms . The paper uses fuzzy partition method using CURE clustering algorithm. Next FP-Growth is applied, in the first scan regions with maximum count for each visited page are find out and in the second scan FFP tree is constructed.Finally recursive method is used Finally, recursive method is used to extract all fuzzy association rules according to min FC .The fuzzy FP-tree structure is used to handle the page visit time efficiently and effectively for a deeper understanding of user navigational behavior.[7] MehrdadJalali, Norwati Mustapha, Ali Mamat, Md. Nasir B Sulaiman[8] proposed novel contribution to clustering user navigation patterns. Graph partitioning algorithm is base for this approaching first step web log data preprocessing is done. Web log data preprocessing includes session identification, user differentiation and data cleaning. Afterprepressing task navigation modeling is done. In this paper the degree of connectivity in each pair of web pages depends on two main factors: the occurrence of two pages in a session and thetime position of two pages in a session. They proposed an algorithm to model thepages accesses information as an undirected graph $M= (V, E)$. For approximating the connectivity degree of each two web pages in sessions they propose a weight measure. User navigation patterns are modeled using graph partitioning algorithm. An undirected graph based on connectivity between each pair of web pages is established in order to mining user navigation pattern. For assigning weights to edges of the graph they propose novel formula. The experimental results shows that the quality of clustering for user navigation pattern in web usage mining systems can be improved by this approach. [8]KobraEtminani,Mohammad-R.Akbarzadeh-T, NooraliRaeejiYanehsari[9] Proposed a new method to extract user navigational patterns from web log data. For this purpose ant-based clustering has been used. A neighborhood function needs to be defined. Once the clustering is completed, alignment processing has been applied to the extracted sequences in individual cluster and frequent patterns are identified.[9]

IV. CONCLUSION

The conclusion from literature review different researches has been done on user’s future behavior prediction. In this paper of review different algorithms like graph partitioning,LCS,K-mean, Fuzzy C-Means,Ant-based clustering and Kernelized Fuzzy C-Means algorithm s are used for finding out user’s future behavior.

REFERENCES

[1] Yan Wang ”Web Mining and Knowledge Discovery of Usage Patterns”, 2000.

[2] R.Cooley, B. Mobasher, and J. Srivastava“Data Preparation For Mining World Wide Web Browsing Patterns”,1999.

[3] AlexandrosNanopoulos, DimitrisKatsaros and YannisManolopoulos “Effective prediction of web-user accesses: A data mining approach,” in Proc. Of the Workshop WEBKDD, 2001.

[4] Yi-Hung Wu and Arbee L. P. Chen, “Prediction of Web Page Accesses by Proxy Server Log” World Wide Web: Internet and Web Information Systems, 5, 67–88, 2002.

[5] Mathias Gery, Hatem Haddad “Evaluation of Web Usage Mining Approaches for User‟s Next Request Prediction” WIDM‟03 Proceedings of the 5th ACM international workshop on web information and datamanagement p.74-81, November 7-8,2003.

[6] GerdStumme, Andreas Hotho, Bettina Berendt, “*Semantic web mining a state of the art andfuture directions*”, Institute of Information Systems, Humboldt University Berlin, Spandauer,Vol. 78, pp. 1-36, 2004.

[7] Bettina Berendt, Andreas Hotho, DunjaMladenic, “*A Roadmap for Web Mining From Web toSemantic Web*”, Institute of Technical and Business Information Systems, Otto–von–Guericke University Magdeburg, Vol. 18 , pp. 1-21, 2008.

[8] Andrew Clearwater, “*The new ontologies: the effect of copyright protection on public scientificdata sharing using semantic web ontologies*”, Vol. 10, pp. 182-205, 2010.

[9] Paul Buitelaar, Philipp Cimiano and Bernardo Magnini, “*Ontology Learning from Text: AnOverview*”, DFKI, Language Technology Lab AIFB, University of Karlsruhe, Vol. 3, pp. 1-10,2003.

[10] Xiaohui Tao, Yuefeng Li, and NingZhong, Senior Member, “A Personalized Ontology Modelfor Web Information Gathering”, *IEEE Transactions on Knowledge and Data Engineering*,Vol. 23, No. 4, pp. 496-511, 2011.
[11] Cat ledge, L. and Pitkow, J., “Characterizing browsing behaviours on the World Wide Web”, *Computer Networks andISDN Systems*,1995, Vol. 27, No. 6, Pp. 1065-1073.
[12] Cooley, R., Srivastava, J. and Mobasher, B. , “Web mining: Information and pattern discovery on the world wide web,Tools with Artificial Intelligence”, *Ninth IEEE International Conference on In Tools with Artificial Intelligence*, 1997.Proceedings., Vol. 10, pp. 0558-567.
[13] Eirinaki, M. and Vazirgiannis, M. , “Web mining for web personalization”**,** *ACM Transactions on Internet Technology*(TOIT), 2003, Vol. 3, Issue 1, Pp. 1-27.
[14]http://en.wikipedia.org/wiki/Longest_common_subsequence_problem, Last Accessed on 27-02-2011.
[15] Huysmans, J., Baesens, B. and Vanthienen, J.), “Web Usage Mining: A Practical Study”, *KatholiekeUniversitiesLeuven*,Dept. of Applied Economic Sciences (2003).
[16] Inktomi, A, “ Web surpasses one billion documents”, www.inktomi.com/new/press/billion.html 2000