Clustering Web Log Files – A Review

R. Suguna
Assistant Professor
Department of Computer Science and Engineering
Arunai College of Engineering,
Thiruvannamalai – 606 603

D. Sharmila
Professor and Head
Department of Electronics and Instrumentation Engineering
Bannari Amman Institute of Technology
Sathyamangalam- 638 401

ABSTRACT

Web mining is the area of data mining. It consists of majorly three subareas: (i) Web content mining, (ii) Web structure mining and (iii) Web usage mining. Web usage mining uses the web log files which are resided at web servers, proxy servers and browser machines as a source to identify user's website access behaviors. The users website visiting details are recorded in various sources in common log format. The web logs are massive in size and not lying in appropriate format. So, careful preprocessing is applied to make the web logs suitable for extracting knowledge. Pattern analysis techniques are applied to the preprocessed web logs to obtain the information from them. This paper gives a review on the well-known pattern discovery algorithm named clustering algorithms.

Keywords: Web logs, Clustering algorithm

1. INTRODUCTION

Today is an information wealth world, billions of users daily accessing the website and navigating the web pages for getting and accessing the information from the web. Web Usage Mining deals with the usage details and behavior of the website visitors. Now a day it becomes an interesting and necessary research field for satisfying customer expectations. The navigation details are maintained in the web servers, proxy servers and client machines in the form of Common Log File format [2].

Preprocessing [1-2] techniques are applied to the weblogs to remove noise and inconsistency. Clustering [1] is a significant and essential activity in data analysis. Clustering is described as the process of grouping the similar kind of objects together with in a same group. The data objects within a group have high similarity and common access behavior than the other groups. Clustering is one of the data mining techniques, used to group the similar data items. In web usage mining, varieties of clustering algorithms [3-32] were proposed by the researchers to group the web log files either in Session based, Link based or User based. The clusters are used to identify the user's website access behavior, this will be helpful for the website analyst for better web recommendation and making smart business decisions, website restructuring for attracting the users.

The rest of the paper is organized as follows: Chapter 2 dealt with literature survey and chapter 3 describe the conclusion on the paper.

2. LITERATURE SURVEY

The authors Kate A. Smith and Alan Ng (2003) used the Self Organizing Map (SOM) to group the web pages based on user's navigation behavior. For dimensionality reduction, most familiar clustering algorithm namely k-means clustering algorithm is used. Their approach proven better result for identifying user's navigation behavior.

Jianhan Zhu et.al. (2004) proposed a clustering algorithm called PageCluster, which clusters conceptually related pages in each level of the link hierarchy based on their in-link and out-link similarities. Clusters are visualized in a archetype called Online Navigation Explorer (ONE) for adaptive Web site steering.

Athena Vakali (2004) described a well-liked clustering methodology for grouping web users and web sessions. They have proposed two algorithms for user's session based clustering such as similarity based and model based approaches. In similarity based clustering, three methods were proposed which are (i) Sequence Alignment Method which measure the similarity between the sessions, (ii) Web pages are clustered using BIRCH algorithm, (iii) Click stream method measures the similarity between the two click streams. In model based approaches, the model structure can be determined by model selection techniques and parameters estimated using maximum likelihood algorithms, e.g., the EM (Expectation-Maximization) algorithm. Markov models are the most indicative models that are used for users' sessions. Once the model is learned, we can use it to assign each user to a cluster or fractionally to the set of clusters. In Link based approach, the Web is treated as a directed graph, the goal is to cluster in the same group

the Web pages with similar content and this can be achieved by eliminating arcs between dissimilar pages.

George Pallis et.al., (2005) focused on session based clustering. Model based clustering algorithm is proposed by the authors to group the web logs based on the sessions. Sophia G. Petridou et.al. (2006) enhanced the k-means algorithm with the KL(Kullback-Leibler) - divergence. The data are normalized before applying the clustering algorithms. Each cell in the normalized table expresses the probability with which a user will visit a page and each row is the probability distribution of each user (p, q).

Castellano, G., et.al., (2006) used the fuzzy C-Means algorithm to categorize user sessions in order to derive groups of users which exhibit similar access patterns. In reformed fuzzy C-means a neighborhood influence parameter □at each pixel is calculated. The probabilistic constraint is removed by equating sum of membership function in a cluster to n.

Raju, G., T., and Satyanarayana, P., S., (2008) proposed a novel clustering algorithm for grouping the users. Prefetching is applied to each cluster to identify the feature needs. Adaptive Resonance Theory (ART) neural network clustering algorithm is newly proposed in this paper. The authors George Pallis et. al. (2008) proposed a new algorithm clustweb, which group the clients based on their domains. This algorithm taken additional effort to cluster the users based on their domain in different websites. (Inter web pages). Graph based approach is proposed to cluster the user group.

The author Peilin Shi, (2009) derived a new algorithm which overcome the difficulties of existing clustering (i.e) soft computing techniques, such as fuzzy theory and rough set theory which are suitable for handling the issues related to understandability of patterns, incomplete/noisy data, and can provide approximate solutions faster. The authors have proposed a novel approach based on rough k-means in fuzzy environments to cluster the website visitors. They have group the users based on user's navigation and time spent on each website. This approach is useful to discover remarkable user access patterns in web log. Jyoti et al., (2009) presented web user clustering approach based on Rough set theory.

Li Chaofeng, (2009) presented an new algorithm called WSCBIS (Web Sessions Clustering Based on Increase of Similarities) in this paper. This algorithm defines the number of clusters according to the Web site's structures, Web site's contents and the kinds of user's interesting behaviors which the analyzer desired. It takes advantage of ROCK to decide the initial points of each cluster and determines the criterion function according to the contributions of overall increase in similarities made by dividing Web sessions into different clusters.

Suneetha, K. R., and Krishnamoorthi, R., (2009) proposed a novel approach called Cluster and PreFetch (CPF) for prefetching of web pages based on the Adaptive Resonance Theory (ART) neural network clustering algorithm. First, hey applied algorithm to cluster the users and then prefetch the web pages for each cluster before the users request them. Their CPF approach effectively reduces the user perceived latency without wasting the network resources.

The authors Niranjana Kannan and Elizabeth Shanthi, (2010) used Expectation Maximun algorithm for clustering the users with similar browsing behavior and Maximum likelihood algorithm is applied in each cluster to find the users visited page navigation. Their method yield good memory efficiency, easy implementation with profound probabilistic environment. The authors Santhisree, K., and Damodaram, A., (2010) proposed Similarity Upper Approximation clustering method for grouping the web data with respect to user's visited pages. This method clearly shows the user's web pages visiting behavior, order of occurrences of web pages visited by the user.

An undirected graph methodology is proposed by the authors Dipa Dixit and Jayant Gadge, (2010). In this paper, an undirected graph is used to group the user's visited pages in sessions. A weight is assigned for each page based on (i) no of times the page is visited by the user and (ii) time spent on each web page.

The authors Sujatha, N., and Iyakutty, K., (2010) attempted to enhance the quality of the K-means clustering algorithm further they found that, most of the researchers only accelerate the k-means clustering algorithm but not attempting to enhancing the efficiency of the algorithm. They have used the K-Means clustering algorithm for grouping the user sessions initially. Then mode value is taken into consideration for subsequent clustering. This avoids local minimum. The genetic Algorithms are applied to refine the formed clusters.

Babak Anari et. al., (2011) devised a new clustering algorithm with learning automata. In their method, web access patterns are transferred into weight vector using learning automata. Web pages are grouped based on their weight.

Ravindra Mangal and Akash Saxena, (2011) have proposed a new approach for clustering. It overcomes the difficulties of K-Means clustering algorithm. The automatically determine the no of clusters. Proposed efficient clustering algorithm is based on two specific factors, fuzzy parameter which initially selects the random value from the feature vector and decides the number of cluster. Second is, specific factor which merge the clusters according to the similarity. The same algorithm was enhanced by Rajhans Mishra and Pradeep

Kumar, (2012) with different similarity measures. They have used the rough set theory for identifying the outlier data. The authors used Jaccard, Dice, Levenshtein and S3M similarity measures on msnbc data set to form the clusters. A hybrid Leaders complete linkage algorithm is devised by the author Ilampiray P., (2012) for clustering the users into groups with respect to IP address. Apriori algorithm is used for finding the frequently visited web pages by the users within each cluster between the users.

ART1 NN based Clustering Approach with a Complete Preprocessing Methodology was developed by the authors Ramya, C., and Shreedhara, K., S., (2012). Their approach involves two stage processes. During the first stage the features are extracted from the preprocessed log data are and a binary pattern vector P is generated. In the second stage, ART1 NN clustering algorithm is applied to form the clusters.

Chitraa, V., and Antony Selvadoss Thanaman, (2012) developed an algorithm to enhance the performance of k-means clustering algorithm to group the web logs. In step one, dataset is separated into subsets and initial cluster points are planned. K-means algorithm is applied to discover clusters in the second step.

Web log cluster similarity is calculated by using City Block Measures. Miao Wan et.al., (2012) have proposed Web user clustering approach to prefetch Web pages for grouped users based on Random Indexing (RI).

3. CONCLUSION

Many researchers done their research in the field of web usage mining to extract some kind of knowledge from the web log files. Since the nature of the web log files, it find difficult to get it directly. Some preprocessing techniques and pattern discovery algorithms are needed for the web logs to get meaningful information. Always it is better to split the data into some groups with respect to parameters. So, Variety of clustering algorithms was applied to web log files to group them either user based, link based or session based. The above literature survey reveals some of clustering algorithm developed by the researchers. The algorithms are preferred by the authors based on their application and depth of research. Soft clustering algorithms (fuzzy clustering) are chosen by most of the authors because of its efficiency and accuracy. K-means clustering algorithm is chosen because of its simplicity, thus it is modified in many ways to improve its efficiency.

4. REFERENCES

[1] Srivastava, J., Cooley, R., Deshpande, M., and Tan, P. M. 2000. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKKD Explorations, vol.1.

[2] Cooley, R., Mobasher, B., and Srivastava, J. 1999. Data preparation for mining world Wide Web browsing patterns. Journal of Knowledge and Information Systems,(1) 1.

[3] Kate A., Smith, Alan Ng. 2003. Web page clustering using a self-organizing map of user navigation patterns. Decision Support Systems 35, 245– 256.

[4] Jianhan Zhu, Jun Hong, and John G., Hughes. 2004. PageCluster: Mining Conceptual Link Hierarchies from Web Log Files for Adaptive Web Site Navigation. ACM Transactions on Internet Technology, Vol. 4, No. 2, Pages 185–208.

[5] Athena Vakali, Jaroslav Pokorn, and Theodore Dalamagas. 2004. An Overview of Web Data Clustering Practices. 3, Springer-Verlag Berlin Heidelberg.

[6] George Pallis, Lefteris Angelis, and Athena Vakali. 2005. Model-Based Cluster Analysis forWeb Users Sessions. M.S. Hacid et al. (Eds.): ISMIS 2005, LNAI 3488, pp. 219–227, Springer-Verlag Berlin Heidelberg.

[7] Sophia G., Petridou, Vassiliki A., Koutsonikola, Athena I., Vakali, and Georgios I., Papadimitriou. 2006. A Divergence-Oriented Approach for Web Users Clustering. ICCSA 2006, LNCS 3981, pp. 1229 – 1238. Springer-Verlag Berlin Heidelberg.

[8] Castellano, G., Fanelli, A., M., and Torsello, A., M. 2006. Mining usage profiles from access data using fuzzy clustering. Proceedings of the 6th WSEAS International Conference on Simulation, Modelling and Optimization, Lisbon, Portugal.

[9] Raju, G., T., and Satyanarayana, P., S. 2008. Knowledge Discovery from Web Usage Data: A Novel Approach for Prefetching of Web Pages Based on Art Neural Network Clustering Algorithm. International Journal of Innovative Computing, Information and Control , ICIC International, ISSN 1349-4198 Volume 4, Number 4, April 2008l.

[10] George Pallis A., Athena Vakali A., and Jaroslav Pokorny. 2008. A clustering-based prefetching scheme on a Web cache environment, Computers and Electrical Engineering 34,309–323.

[11] Peilin Shi. 2009. An Efficient Approach for Clustering Web Access Patterns from Web Logs. International Journal of Advanced Science and Technology, Volume 5.

[12] Jyoti, A., K., Sharma, Amit Goel. 2009. A Novel Approach for clustering web user sessions using RST. International Journal on Computer Science and Engineering Vol.2(1), 56-61.

[13] Li Chaofeng, 2009. Research on Web Session Clustering. Journal of Software, Vol. 4, No. 5.

[14] Suneetha, K., R., and Krishnamoorthi, R. 2009. Identifying User Behavior by Analyzing Web Server Access Log File. IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4.

[15] Niranjana Kannan, and Elizabeth Shanthi. 2012. Classification and Clustering of Web Log Data to Analyze User Navigation Patterns. Volume 1, No. 1. Journal of Global Research in Computer Science.

[16] Santhisree, K., and Damodaram, A.. 2010. Clustering on Web usage data using Approximations and Set Similarities. International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 4 27.

[17] Dipa Dixit, and Jayant Gadge. 2010. A New Approach for Clustering of Navigation Patterns of Online Users. International Journal of Engineering Science and Technology, Vol. 2(6), 1670-1676.

[18] Sujatha, N., and Iyakutty, K. 2010. Refinement of Web usage Data Clustering from K-means with Genetic Algorithm. European Journal of Scientific Research, ISSN 1450-216X Vol.42 No.3,pp.478-490.

[19] Babak Anari , Mohammad Reza Meybodi and Zohreh Anari. 2011. Clustering Web Access Patterns Based on learning Automata, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1.

[20] Ravindra Mangal, and Akash Saxena. 2011. Efficient Clustering Algorithm to Discover User Pattern Applying on Weblog Data. International Journal on Emerging Technologies 2(2): 51-55.

[21] Rajhans Mishra, and Pradeep Kumar. 2012. Clustering Web Logs Using Similarity Upper Approximation with Different Similarity Measures. International Journal of Machine Learning and Computing, Vol. 2, No. 3.

[22] Ramya, C., and Shreedhara, K. S. 2012. Clustering of Web Users using ART1 NN based Clustering Approach with a Complete Preprocessing Methodology. International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 1.

[23] Chitraa, V., and Antony Selvadoss Thanaman. 2012. An Enhanced Clustering Technique for Web Usage Mining. International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 4.

[24] Ilampiray, P. 2012. Efficient Resource Utilization of Web Using Data Clustering and Association Rule Mining. Journal of Theoretical and Applied Information Technology, Vol. 37 No.2.

[25] Miao Wan, Arne Jonsson, Cong Wang, Lixiang Li and Yixian Yang. 2012. Web user clustering and Web prefetching using Random Indexing with weight functions. Knowledge and Information Systems, (33), 1, 89-115.

[26] Khalil, F., and Li, I. 2009. An Integrated Model for Next Page Access Prediction, Copyright 2009 Inderscience Enterprises Ltd.