RESEARCH ARTICLE OPEN ACCESS

An Weblog based Mining Approach to Improve the Clustering for Efficient Subsequent Page Prediction

Sonam Mittal [1], Devendra Singh Rathore [2], Mukesh Kumar [3]
[1], [2], [3] Computer Science & Engineering, RabindraNath Tagore University - Raisen MP

ABSTRACT
Web mining is one of the data mining methods to discover and mine information from Web documents and services without human intervention. The objective of Web structure mining is to categorize the Web pages and produce information such as the resemblance and link between them, taking advantage of their hyperlink topology. Sequential mining is an extension of basic association rule mining that accommodates ordered set of items or attributes, where the same item may be repeated in a sequence. In this research study, we have applied mining approach over weblog data and through clustered record the analysis is done for next page prediction. The proposed method presented in this paper shows better results as compared to earlier proposed work.

Keywords: - **Clustering**, Page Prediction, Sequential Pattern Mining.

I. INTRODUCTION

Web mining is used to predict user behavior. Web mining [1,2] is a very broad research area emerging to solve the issues that arise due to the WWW phenomenon. The Web mining research is a converging research area from several research communities, such as Databases, Information Retrieval and Artificial Intelligence. This work overview the most important issue of Web mining, namely sequential traversal patterns mining.

Web mining helps to improve the power of web search engine by classifying the web documents and identifying the web pages. The network of worldwide has grown in the past few years from a small research community to the biggest and most popular way of communication and information dissemination. Every day, the WWW grows by roughly a million electronic pages, adding to the hundreds of millions already on-line. At present, 4.5 billions web pages in the world which this number increases with the rate of 8.8 million pages per day. World Wide Web serves as a platform for exchanging various kinds of information, ranging from research papers, and educational content, to multimedia content and softwaredocument is a template. An electronic copy can be downloaded from the conference website. For questions on paper guidelines, please contact the conference publications committee as indicated on the conference website. Information about final paper submission is available from the conference website.

II. TYPE OF WEB DATA MINING

Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World¬ Wide Web. Web mining categorize into three areas of interest -

1) Web Content Mining
2) Web Structure Mining
3) Web Usage Mining

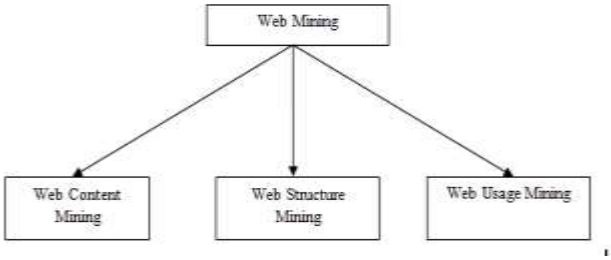


Figure 1 : Web Mining Categories

A. Web Content Mining

Web Content Mining describes the finding of useful information from the Web contents or data or documents. Though, what consist of the Web contents could encompass a very broad range of data. Previously the Internet consists of different types of services and data sources such as Gopher, FTP and Usenet. Now most of those data are either ported to or accessible from the Web. It is mentioned in that in the last several years the growth in the amount of government information has been tremendous. The existence of Digital Libraries that is also accessible from the Web. So many companies are transforming their businesses and services electronically.

B. Web Structure Mining

Web Structure mining discover the model that suits the link structures of the Web. The model is based on the topology of the hyperlinks with or without the description of the links. This model can be used to categorize Web pages and is useful to generate information such as the similarity and relationship

between dissimilar Web sites. Web structure mining could be used to discover authority sites for the subjects (authorities) and overview sites for the subject's that point to many authorities (hubs)[8].

C. Web Usage Mining

Web usage mining make an efforts to make sense of the information generated by the Web surfer's sessions or behaviors. Although the Web content and structure mining make use of the real or main data on the Web, Web usage mining mines the minor data derived from the interactions of the users as interacting with the Web. The Web usage data which includes the data from Web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls, and any extra data as the outcome of interactions[8,9]..

III. PROCESS OF MINNING WEBLOG DATA

A. *Data Collection*

In any user session, all the navigation activity on the web site is cached using a log file in the web server. The web log files collected are given as input for the analysis in the web page prediction process. The web log file mainly contains the following fields like IP address, user_ID, base_url, date, method, file, catdesc, protocol, code, bytes, referrer, user_agent etc[10-12].

B. Pre-Processing Log Files

Generally in the web usage mining, the preprocessing [3] is considered as a basic and essential task. Preparing cached log data for analysis by removing irrelevant data items is known as Preprocessing. The quality of the data is an important issue in the mining process. About the 80% of mining efforts are often spend to improve the quality of data [4]. We obtain mostly incomplete, noisy and inconsistent data from the server logs. The attributes that we may have seen for in quality data depends upon the accuracy, completeness, consistency, timeliness, believability, interpretability and accessibility. Preprocessing is needed for obtaining the above mentioned attributes to extract the interesting patterns of knowledge. The following steps explain the preprocessing of web log files [5,6].

Removal of noisy and irrelevant data from the web logs [7] is the basic step for data cleaning. When the user request to HTML web pages, the embedded images are also be downloaded and gets stored in the web server. But these are not explicitly requested by the users which are avoided.

In the proposed system, data cleaning removes the irrelevant data from the given log files. For the given web log file, the log files before data cleaning was 4520 and after data cleaning it is 1288. It nearly removes the 70%-80% of unwanted data from the log file which in-turn also reduces the processing time of the recommender system.

C. Clustering Log Files

After data cleaning process, identification of user is performed. Identification of individual web users who accessed a website is an important step in web usage mining and it is based on the "log_id" field in the collected weblog file..

IV. SEQUENTIAL PATTERN MINNING FOR WEB LOG FILE

The proposed method is based on Sequential Pattern Mining of log files. It is used for trend analysis to identify user pattern in the process of Web Usage Mining. It depends on the performance of the clustering of the amount of requests. Here, SOM is used with Sequential Pattern Mining (SPM) algorithm to find more frequent sequential patterns.

The existing algorithm access whole database multiple times. The proposed method clusters the session data using SOM algorithm called Self Organizing Map (SOM). It clusters the data according to similarity with the help of Sequential Pattern Mining. The developed algorithm is the combination of both of it and named as SPMSOM.

Algorithm 1 (Self Organizing Map)
Select output layer network topology

- Initialize current neighborhood distance, D (0), to an affirmative value.
- Initialize weights from input to output to small random values. Let t = 1
- As calculation limits are not exceeded do

1) Choose an input sample a1
2) Calculate the second power of the Euclidian metric of a1 from weight vectors (wj) linked with each output node $\Sigma nk = 1 (al, k - wj, k (t))2$
3) Choose output node j* which has weight vector with minimum value from step 2)
4) Modify weights to each & every node within a topological distance given by Dt (t) from j*, using the weight modify rule: $Wj(t+1) = wj(t) + \eta (t) (a1 – wj(t))$
5) Increment t

End while

V. RESULT ANALYSIS

A. About DataSet

We presented the performance over various data sets for accuracy purposes. The synthetic data set is used for evaluating the performance of the web recommendation system. The data set is taken from UCI Machine Learning data source that is available on www.kaggle.com.

The report of the experimental results on the performance of SPMSOM in comparison with a recently developed algorithm; VoMM [2], which is the fastest algorithm for mining sequential patterns. The main purpose of this experiment is to demonstrate how effectively the sequential traversal patterns with weight constraint can be generated by incorporating a weight page, weight of sequence with a support. First, it shows how the number of sequential traversal patterns can be adjusted through user assign weights, the efficiency in terms of runtime of the SPMSOM algorithm, and the quality of sequential traversal patterns. Secondly, shows that SPMSOM has good scalability against the number of sequence transactions in the data sets with clustering.

The analysis of SPMSOM algorithm is similar to *FP-growth* [10]. At first, given an FSTP-tree *T*, Weight Range WR, and Average Weight Range of Session AWR, mining the frequent sequential traversal patterns with weight constraint from *T* with traversal strategy. If *T* only contains a single path of FSTP-tree in which each node only has a single child, then gets directly the frequent sequential traversal patterns with weight constraint. When *T* is multi-path FSTP-tree, each generate 1-SPMSOM and construct prefix traversal sequence pre. By adding the suffix, generates the frequent sequential traversal pattern.

B. Experiment

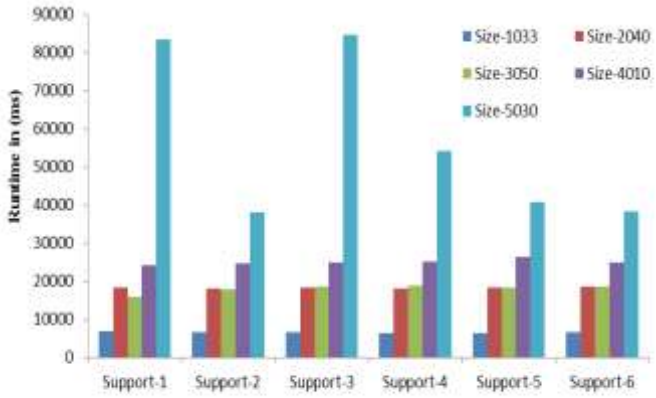
All the experiments are performed on a 3.5 GHz Pentium V processor with 1.5 Gegabytes main memory, running on Microsoft Windows 10. In addition, all the programs are written in .Net platform. The experiments were carried out on real data sets to evaluate the performance of SPMSOM algorithm. The data sets, which contain several months worth of click sequence data from two e-commerce web sites. Collecting the same number sequence data of the two data sets, which are divided into the different length sessions, and the average session contains 5-11 pages[13].

The following Table 1 shows page details with Support and Min-Max weight range.

TABLE 1
PAGE NAME WITH WEIGHT RANGE

S.No.	Page ID	Page Name	Support	Min. Weight	Max. Weight
1	P1	Glass	9	3	40
2	P2	Mobile	7	4	10
3	P3	Garments	7	5	20
4	P4	Hardware	6	6	8
5	P5	Furniture	6	4	12
6	P6	Entertainment	1	1	2
7	P7	Base	2	1	3

The above Table represents the inputs that are given at initial stages of SPMSOM. The table contains Page ID from P1 to P7, with page name: Glass, Mobile, Garments, Hardware, Furniture, Entertainment and Base. The pages from P1 to P7 have different support with different minimum and maximum weight.



. Fig 2 : Running Time (in ms) with different size and different support

The Figure 2 show the Running time (in ms) of SPMSOM under different record size with different support. Running time (in ms) of SPMSOM varies from record size 1033 to record size 5000 with support 1, support 2, support 3, support 4, support 5 and support 6. The graph shows that while taking record size 1000 with support 1 then running time of SPMSOM is 3810 ms, similarly with support 2,3,4,5 and 6 running time of SPMSOM is 5708 ms, 5723 ms, 5723 ms, 5505 ms, 5474 ms and 5708 ms respectively. In the above mentioned graph, support 4 is taking minimum time and support 1 is taking max time with their corresponding size. Similarly, with record size 3000, SPMSOM running time with support 1 is minimum and with support 6 is maximum with their corresponding size. Similarly with record size 4000, SPMSOM running time with support 2 is minimum and with support 4 is maximum with their corresponding size. Additionally, with record size 5000, SPMSOM running time with support 1 is minimum and with support 5 is maximum with their corresponding size. Similarly with record size 5000, SPMSOM running time with support 2 is minimum and with support 3 is maximum with their corresponding size, which is desired output[14,15].

VI. CONCLUSION

The goal of Web Usage Mining is to collect interesting information about user's navigation patterns. To improve the Website from the users' viewpoint, this information can be exploited later. The results produced by the mining of Web logs can be used for various purposes like to personalize the

delivery of Web content; to improve user navigation behaviour through pre-fetching and caching; to improve Web design or e-commerce sites and to improve the customer satisfaction.

This performance study shows that SPMSOM mines the complete set of patterns and is efficient and runs considerably faster as compared to VoMM algorithms. The SPMSOM algorithm is able to quickly determine the suffix of any frequent pattern prefix under consideration by comparing the assigned binary position codes of nodes of the tree. Dealing with unvisited or recently added pages is one of the challenging problems in web page recommendation systems. So, further improvement was hybridization of the efficiency of earlier algorithms. The target web link is give to the new user by matching the pattern tree with the new user's current web access sequence. A weight range is used to adjust the weight of every page with own minimum and maximum weight range (importance of page) number of sequential patterns. The extensive performance analysis shows that SPMSOM is efficient and scalable in mining sequential patterns.

VII. REFERENCES

[1] Chen, G.Y. "Research on computer information processing technology in the era of big data", Network Security Technology and Application, No.3, pp.44-52, 2020

[2] T. Gopalakrishnan, P. Sengottuvelan, A. Bharathi1, R. Lokeshkumar, "An Approach to Webpage Prediction Method using Variable Order Markov Model in Recommendation Systems", Journal of Internet Technology, Volume 19, No.2, pp. 415-425, 2018

[3] Kesavan, S., Saravana Kumar, E., Kumar, A., & Vengatesan, K. "An investigation on adaptive HTTP media streaming Quality-of-Experience (QoE) and agility using cloud media", International Journal of Computers and Applications, pp.1–14, 2019

[4] Pan, L. "Research on Personalized Recommendation System Based on Web Mining", Jiangsu University of Science and Technology, 2020

[5] Charu C. Aggarwal. "Introduction to Special Issue on the Best Papers from KDD", ACM Transactions on Knowledge Discovery from Data (TKDD), Vol.11, pp.4, 2017

[6] Rahul Moriwal, Vijay Prakash, "An Efficient Algorithm for Finding Frequent Sequential Traversal Patterns from Web Logs based on Dynamic Weight Constraint", 2019.

[7] Chitraa, V., Antony Selvadoss Thanamani, "An Enhanced Clustering Technique for Web Usage Mining", International Journal of Engineering Research & Technology (IJERT), Vol.1, Issue 4, June-2018.

[8] Ketki Muzumdar, Ravi Mante, Prashant Chatur, "Neural Network Approach for Web Usage Mining", International Journal of Recent Technology and Engineering (IJRTE), Vol.-2, Issue-2, May-2017.

[9] Omar Zaarour, Mohamad Nagi, "Effective web log mining and online navigational pattern prediction", Knowledge Based Systems, ELSEVIER, 2017, pp.50-62.

[10] Umapathi, C., Aramuthan, M., Raja, K., "Enhancing Web Services Using Predictive Caching", International Journal of Research and Reviews in Information Sciences, Vol.-1, No.-3, Sept-2016.

[11] Song Sun, Joseph Zambreno, "Design and Analysis of a Reconfigurable Platform for Frequent Pattern Mining", IEEE, Vol.22, No.9, Sept-2016.

[12] Vijayalakshmi, S., Mohan, V., Suresh Raja, S., "Mining of Users Access behavior for Frequent Sequential Pattern from Web Logs", International Journal of Database Management Systems (IJDMS) Vol.2, No.3, August 2016.

[13] Mahdi Esmaeili, Fazekas Gabor, "Finding Sequential Patterns from Large Sequence Data", International Journal of Computer Science (IJCSI), Vol.7, Issue 1, No.1, January 2014.

[14] Priyanka Makkar, Payal Gulati, Dr. A.K. Sharma, "A Novel Approach for Predicting User Behavior for Improving Web Performance", International Journal on Computer Science and Engineering (IJCSE), Vol. 02, No. 04, 1233-1236, 2014.

[15] Utpala Niranjan, Dr. R.B.V. Subramanyam, Dr. V.Khanaa, "An Efficient System Based On Closed Sequential Patterns for Web Recommendations", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 4, pages 26-34, May 2010.