A Web Usage Mining Approach Based On New Technique In Web Path Recommendation Systems

R. Khanchana[1] and Dr. M. Punithavalli[2]

[1]Research Scholor, Karpagam University, Coimbatore, Tamilnadu, India.
[2]Director, Department of Computer Application, Sri Ramakrishna Engineering College, Coimbatore, India.

ABSTRACT

The Internet is one of the fastest growing areas of intelligence gathering. The ranking of web page for the Web search-engine is one of the significant problems at present. This leads to the important attention to the research community. Web Prefetching is used to reduce the access latency of the Internet. However, if most prefetched Web pages are not visited by the users in their subsequent accesses, the limited network bandwidth and server resources will not be used efficiently and may worsen the access delay problem. Therefore, it is critical that we have an accurate prediction method during prefetching. To provide prediction efficiently, we advance an architecture for predicting in Web Usage Mining system and propose a novel approach for classifying user navigation patterns for predicting users' requests based on clustering users browsing behavior knowledge. The Expremental results show that the approach can improve accuracy, precision, recall and Fmeasure of classification in the architecture.

Keywords

Data cleaning, Browsing behaviour, Path Recommendation.

1. INTRODUCTION

With the explosive growth of knowledge available on the World Wide Web, which lacks an integrated structure or schema, it becomes much more difficult for users to access relevant information efficiently. Meanwhile, the substantial increase in the number of websites presents a challenging task for webmasters to organize the contents of the websites to cater to the needs of users. Modeling and analyzing web navigation behavior is helpful in understanding what information of online users demand. For decision management, the result of web usage mining can be used for target advertisement, improving web design, improving satisfaction of customer, guiding the strategy decision of the enterprise, and marketing analysis etc [12][13][7][10].

A new technique was proposed by Page and Brin called PageRank to compute the importance of Web pages. PageRank [8] determines the significance of Web pages and helps a search engine to choose high quality pages more efficiently. However, PageRank technique still experiences a drawback: biased ranking to newly pages. Since the newly web pages regularly do not receive enough in-links to illustrate its real importance in the initial time stage. The bias is unfired to the new web pages and thus the search results will be unreliable. In order to produce a better PageRank, two biasing features are considered. They are

- The length of time spent on visiting a page
- The frequency of the visited page.

Predicting the users' browsing pattern is one of web usage mining technique. For this purpose, it is required to recognize the customers' browsing behaviors by means of analyzing the web data or web log files. Predicting the exact user's next needs is according to the earlier related activities. There are several merits to employ the prediction, for example, personalization, building proper web site, enhancing marketing strategy, promotion, product supply, getting marketing data, forecasting market trends, and increasing the competitive strength of enterprises etc. The clustering will perform classification in the browsing features using Fuzzy Possibilistic algorithm for the purpose of clustering

This paper focuses on Web Usage Mining with the help of Classification technique. The classification will perform path recommendation based on users browsing pattern as knowledge. This paper uses Longest Common Subsequence (LCS) algorithm for the purpose of classification.

2. RELATED WORK

Recently, several Web Usage Mining systems have been proposed to predicting user's preference and their navigation behavior. In the following we review some of the most significant Web Usage Mining systems and architecture that can be compared with our system.

Tasawar *et al.*, [1] proposed a hierarchical cluster based preprocessing methodology for Web Usage Mining. In Web Usage Mining (WUM), web session clustering plays a important function to categorize web users according to the user click history and similarity measure. Web session clustering according to Swarm assists in several manner for the purpose of managing the web resources efficiently like web personalization, schema modification, website alteration and web server performance.

Yaxiu *et al.*, [2] put forth web usage mining based on fuzzy clustering. The World Wide Web has turn out to be the default knowledge resource for several fields of endeavor, organizations require to recognize their customers' behavior, preferences, and future requirements, but when users browsing the Web site, several factors influence their interesting, and various factor has several degree of influence, the more factors consider, the more precisely can mirror the user's interest.

Shiguang *et al.*, [3] given the improvement of page ranking algorithm based on timestamp and link. The conventional ranking technique favors the old pages, which makes old pages always emerge on the top of the ranking results when pages are ranked according to the dynamic web by the static ranking algorithm. Therefore, this paper proposes a temporal-link-analysis technique to overcome the problem.

Houqun *et al.*, [4] proposed an approach of multi-path segmentation clustering based on web usage mining. According to the web log of a university, this paper deals with examining and researching methods of web log mining; bringing forward a multi-path segmentation cluster technique, that segments and clusters based on the user access path to enhance efficiency.

3. METHODOLOGY

Web mining can be categorized into three categories (as Fig. 1) which are web content mining, web structure mining and web usage mining.

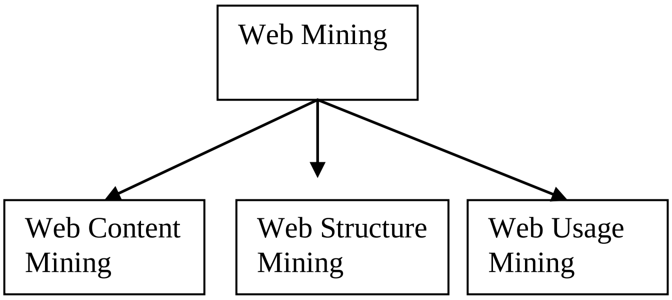

Fig 1. Web Mining Taxonomy

Web content mining focuses on useful knowledge which is extracted from web pages. Web structure mining is used to analyze the links between web pages through the web structure to infer the knowledge. Web usage mining is extracting the information from web log file which is accessed by users.

3. Architecture Overview

In this system we advance architecture for predicting users' next request by applying novel approach in classification module of this architecture. According to different functions the architecture contains three phases: weblog processing, clustering and classification. The architecture is shown in Fig.2. But these phases need strongly work together.

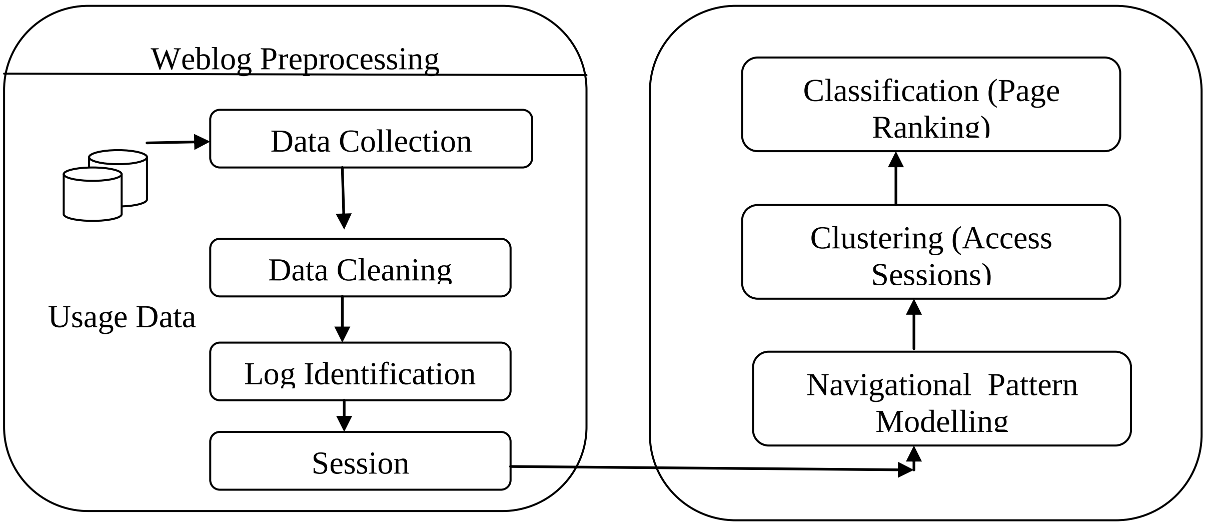
Weblog Preprocessing
Data Collection
Data Cleaning
Usage Data
Log Identification
Session
Classification (Page Ranking)
Clustering (Access Sessions)
Navigational Pattern Modelling

Fig.2 The Architecture for Predicting users' Browsing Behavior

3.1. Weblog Preprocessing

A web usage mining aims to reformat the original web logs to identify all web access sessions. The web server registers all users' activities of the website as web server logs. Generally, several pre-processing tasks need to be done before performing web mining algorithms on the web server logs. For our work, these including

- Data Cleaning
- Log Identification and
- Session Identification

These preprocessing tasks are common for all web usage mining problem.

3.1.1 Data Cleaning

Initially, the data cleaning process is carries out. It removes records with graphics and videos format such as gif, JPEG, etc. The obtained record consists of 1150 records in the log file. After the data cleaning process, which removes graphics and videos format such as gif, JPEG, etc., 560 records are obtained.

3.1.2 Log Identification

There are several types of web logs according to server setting parameters, but typically the log files share the same basic information such as client IP address, user name, request time, requested URL, date, time, server IP address, client bytes sent, server bytes sent, server name, service and instance, HTTP status code etc. The Internet Information Service (IIS) log file format records the above data. It is a fixed ASCII text-based format. Because HTTP system handles the IIS log file format, this format record HTTP system kernel-mode cache bits.

3.1.3 Session Identification

After the data cleaning and log identification, we perform navigation pattern mining on the derived user access sessions. As an important operation of a navigation pattern mining, clustering aims t to group sessions into cluster based on their common properties.

3.2 Navigational Pattern Modeling

The whole Web can be considered as a directed graph in which the N nodes represent the N Web pages and the edges indicates the links among the web pages. Here Fv denotes the set of web pages that are linked to the web page v and PR(v) represents the rank value for the web page v, then the connection v→u will contribute PR(v)/|Fv| units of page v's ranking value to page u. The number of web pages that are included in Fv is denoted as |Fv|. The total of ranking values of the backlinks of a web page is considered as a PageRank and the recursive process is necessary for its computation. In order to bound the effect of rank sinks and assure the recursive computation of PageRank is capable to converge to a certain value, a damping factor (1-α) is used. Usually α is a small value which can be set to 0.15. In addition, for a page which does not have outlinks, a link to all other pages is added in the Web graph. In this manner, ranks that are lost because to pages without outlinks can be redistributed uniformly to all other pages.

3.3 Clustering Algorithm

Forecasting the users' browsing behaviors is one of web usage mining issues. Due to the heterogeneity of users' browsing features, the hierarchical agglomerative clustering algorithm is used to class users' browsing behaviors [9] [11][5][6]. Many different user clusters will be acquired and seem as cluster view for replacing of the global view. The fuzzified version of the k-means algorithm is Fuzzy C-Means (FCM). It is a clustering approach which allows one piece of data to correspond to two or more clusters. Dunn in 1973 developed this technique and it was modified by Bezdek in 1981. The features of both fuzzy and possibilistic c-means techniques is combined for better result. Memberships and typicalities are very significant characteristics for the proper characteristics of data substructure in clustering problem. To illustrate the process, consider the navigational patterns set given in table 1.

Table 1. Browsing patterns generated by Clustering Algorithm

BP Number	Browsing Pattern
1	< $P_2,P_4,P_7,P_3,P_1,P_{20}$>
2	<$P_5,P_7,P_2,P_9,P_{11},P_3,P_5,P_{10}$>
3	<P_4,P_6,P_{12}>
4	<$P_3,P_5,P_8,P_4,P_9,P_{18},P_{26}$>
5	<$P_2,P_{37},P_{23},P_{18},P_{27}$>
6	<$P_{50},P_{24},P_{28},P_{34},P_{14},P_9,P_{17}$>

FPCM constructs memberships and possibilities simultaneously, together with the normal point prototypes or cluster centers for every cluster. Hybridization of Possibilistic C-Means (PCM) AND Fuzzy C-Means (FCM) is the PFCM that frequently rejects several drawbacks of PCM, FCM and FPCM. The noise sensitivity fault of FCM is solved by PFCM, which conquers the concurrent clusters drawbacks of PCM.

Here the FPCM clustering done with two distance measures like Euclidean and Magalanobis distance. When compared Euclidean distance the Magalanobis distance not only consider the distance between two data points and also concentrate on covariance value in the matrix form. If the covariance value is nothing it normally consider the Euclidean distance. The two distance measures use error convergence criteria for better clustering. These functions eliminate the number of iteration and also find the centroid accurately.

3.4 Classification

3.4.1 Extreme Learning Machine

Extreme Learning Machine (ELM) meant for Single Hidden Layer Feed-Forward Neural Networks (SLFNs)] will randomly select the input weights and analytically determine the output weights of SLFNs. This algorithm tends to afford the best generalization performance at extremely fast learning speed. ELM contains an input layer, hidden layer and an output layer.

3.4.2 Hybrid Extreme Learning Machine

A hybrid ELM technique which uses ELM and LM technique can be described as below:Initially, the input weights and hidden biases are created by with the help of AHP technique.Next, the equivalent output weights are analytically determined with the help of ELM algorithm only in first step and randomly produce the output hidden biases. Then, the parameters (all weights and biases) are restructured with the help of LM algorithm.

4.Experimental Results

For evaluating the proposed technique, the database is selected from reputed educational institution website Dataset. Every sequence data is consequent to use The categories are presented in sequential dat rs' page views. Some sample browsing patterns are provided to test the web prediction results by using the proposed technique. Sample browsing behaviors considered are 1-3-4, 4-3-1, 1-7-2, 3-1-2 and 9-11-16. The accuracy of suggested pages are compared with FPCM algorithm with to test the proposed classifier.

Figure 3 represents the web page prediction accuracy by using various patterns suggested by the users.

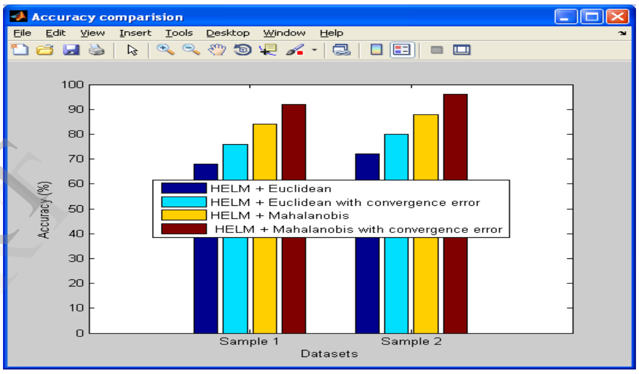

Figure 3: Accuracy of Web Page Prediction

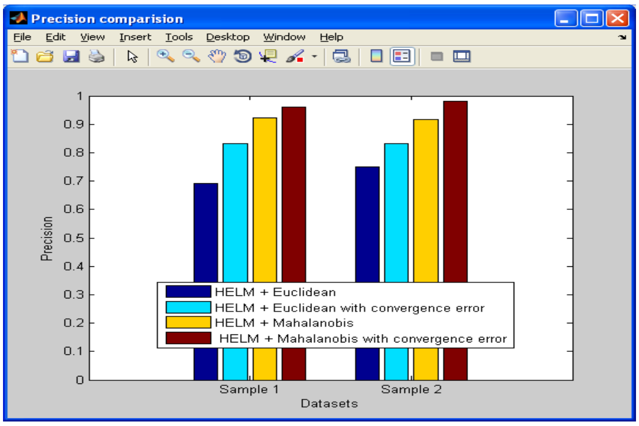

Figure 4: Precision of Web Page Prediction

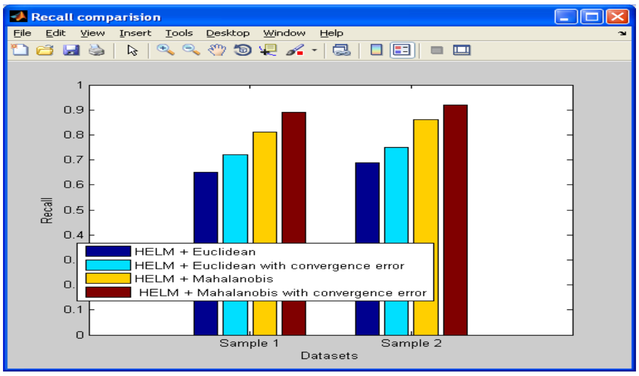

Figure 5: Recall of Web Page Prediction

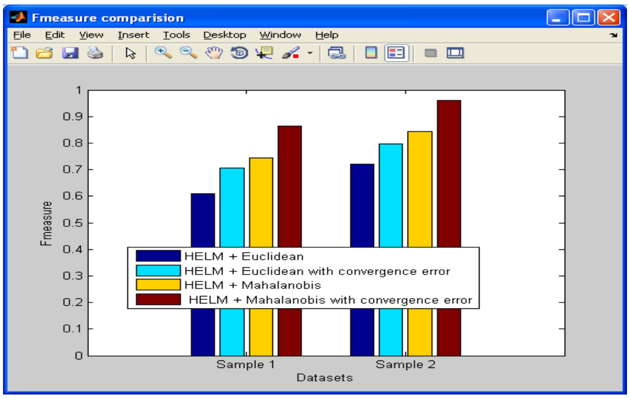

Figure 6: Fmeasure of Web Page Prediction

From the figure, it can be observed that the proposed technique results in better accuracy of prediction when compared with various FPCM class labels. And the following figures 4, 5 and 6 represents the prediction results based on Precision, Recall, Fmeasure respectively. Based on these prediction results rank thewebpages efficiently

5. Conclusion

This paper uses Fuzzy Possibilistic algorithm for clustering. Finally, the prediction based on the clustering result is performed by means of using the Hybrid Extreme Learning Machine (HELM) which has the better capability of better prediction than other conventional techniques. The experimental result shows that the proposed technique results in better accuracy, Precision, Recall and Fmeasure of prediction metrics which shows the clustering performance Evaluation.

REFERENCES

[1] Hussain Tasawar, Asghar Sohail and Fong Simon, "A hierarchical cluster based preprocessing methodology for Web Usage Mining", 6th International Conference on Advanced Information Management and Service (IMS), Pp. 472-477, 2010.

[2] Yaxiu Yu and Xin-Wei Wang, "Web Usage Mining Based on Fuzzy Clustering", International Forum on Information Technology and Applications, Pp. 268-271, 2009.

[3] Shiguang Ju, Zheng Wang and Xia Lv, "Improvement of Page RankingAlgorithm Based on Timestamp and Link", International Symposiums on Information Processing (ISIP), pp. 36-40, 2008.

[4] Houqun Yang, Jingsheng Lei and Fa Fu, "An Approach of Multi-path Segmentation Clustering Based on Web Usage Mining", Fourth International Conference on Fuzzy Systems and Knowledge Discovery, Vol. 4, Pp. 644-648, 2007.

[5] P. Kumar, P.R. Krishna, R.S. Bapi and S.K. De, “Rough clustering of sequential data,” *Data and Knowledge Engineering*, 2007.

[6] Q. Song and M. Shepperd, “Mining web browsing pattern for E-commerce,” *Computers in Industry 57*, 2006, pp.622-630.

[7] F. M. Facca and P. Luca Lanzi, “Mining Interesting Knowledge from Weblogs: A Survey,” *Data and Knowledge Engineering 53*, 2005, pp. 225-241.

[8] M. S. Aktas, M. A. Nacar and F. Menczer, "Personalizing PageRank Based on Domain Profiles”, Processing of WEBKDD 2004 Workshop, 2004.

[9] S.K. De and P.R. Krishna, “Clustering web transactions using rough approximation,” *Fuzzy Sets and Systems 148*, 2004, pp.131-138.

[10] J. Srivastava, R. Cooley, M. Deshpande and P. N. Tan, ”Web Usage Mining: Discovery and Applications of Usage Patterns form Web Data,” *SIGKDD Explorations*, Vol. 1, Issue 2, Jan 2000, pp. 12-23. 285

[11] A.K. Jain, M.N. Murty, and P.J. Flynin, “Data Clustering: A Review,” *ACM Computing Surveys*, Vol. 31, No 3, September 1999, pp. 265-323.

[12] W. Bin and L. Zhijing, “Web Mining Research,” *ICCIMA’03 IEEE*, 2003, pp. 84-89.

[13] R. Kosala, H. Blockeel, “Web Mining Research: A Survey,” *SIGKDD Explorations*, Volume 2, Issue 2, pp. 115.

First. A. R.Khanchana graduated with M.Sc., Computer Science in 2004 from J.K.K Nataraja college of Arts and Science for women, Erode, India and completed M. Phil., from Bharathiar University, India during 2006-2007. She has presented number of papers in National conferences and International seminarsand conferences. She is guiding research scholars and has published many papers in national and international conference and in many international journals. Her areas of Interest include Software Engineering and Data Mining. She has about 6 years of teaching experience. Currently she is working as a Lecturer in Sri Ramakrishna college of Arts & Science for Women, India.

Second. B. Dr. M. Punithavalli is presently working as Director and Head of the Dept of Computer Science, SNS College of arts and science, India. She has published more than twenty papers in National/International Journals. Her areas of interest include E-Learning, Software Engineering, Data Mining, Networking and etc. She has about 19 years of teaching experience. She is guiding many research scholars and has published many papers in national and international conference and in many international journals.