A SURVEY ON NEW TECHNIQUES IN WEB PATH RECOMMENDATION SYSTEMS IN DATA MINING

Ms. Archana Patel[1], Ms. Prachi Joshi[2]
[1]M.E. Student, Department of Computer Engineering
[2]Asst. Prof, Department of Computer Engineering
HGCE, Gujarat, 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

I. 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. 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.[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.. 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 as below :

1. The length of time spent on visiting a page
2. 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 Possibility 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.

II. 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.

[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.

[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,segmentation cluster technique, that segments and clusters based on the user access path to enhance efficiency.

III. METHODOLOGY

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

Several factors influence their interesting, and various factors has several degree of influence, the more factors Consider, the more precisely can mirror the user's Interest.

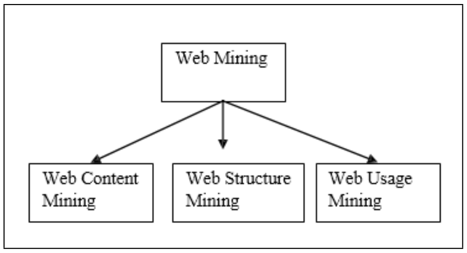

Figure 1: Web mining categories

[3] 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

IV. ARCHITECTURE OVERVIEW

a) **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 :

1) Data Cleaning
2) Log Identification and
3) Session Identification

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

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.

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) **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 to group sessions into cluster based on their common properties.

b) **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][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.

V. EXPERIMENTAL RESULTS

For evaluating the proposed technique, the database is selected from reputed educational institution website Dataset. Some sample browsing patterns are provided to test the web prediction results by using the proposed technique. 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.

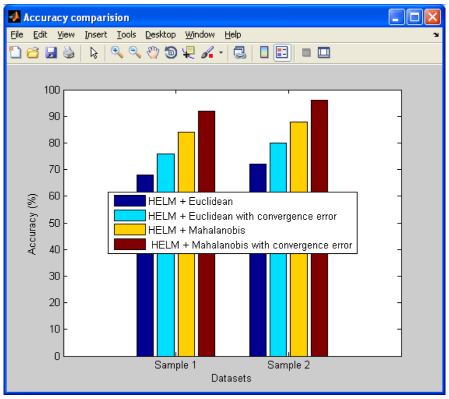

Figure 2: Accuracy of Web Page Prediction

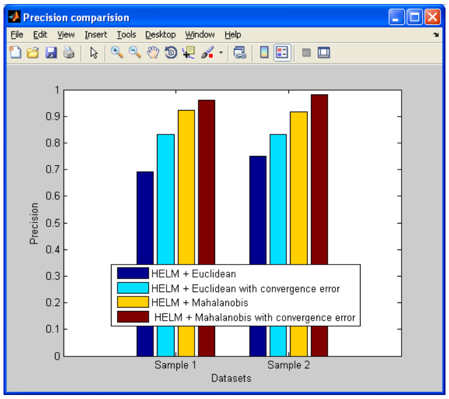

Figure 3: Precision of Web Page Prediction

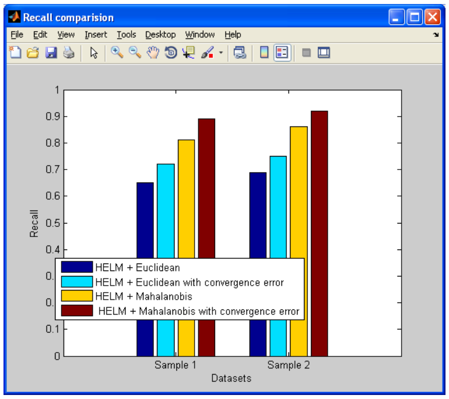

Figure 4: Recall of Web Page Prediction

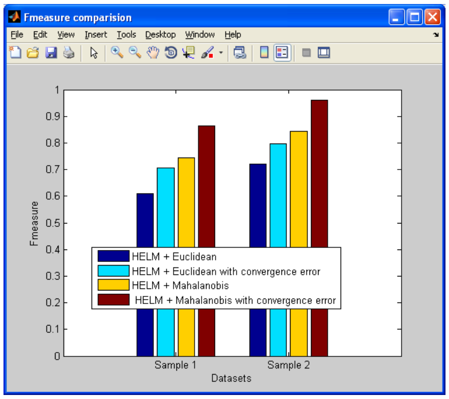

Figure 5: Fmeasure of Web Page Prediction

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 the webpages efficiently

VI. CONCLUSION

This paper uses Fuzzy Possibility 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 Measure of prediction metrics which shows the clustering performance Evaluation.

References

[1] Asghar Sohail Hussain Tasawar and Fong Simon. A hierarchical cluster based preprocessing methodology for web usage mining. *th International Conference on Advanced Information Management and Service (IMS)*, pages 472–477, 2010.

[2] Yaxiu Yu and Xin-Wei Wang. Web usage mining based on fuzzy clustering. *nternational Forum on Information Technology and Applications*, pages 268–271, 2009.

[3] Zheng Wang Shiguang Ju and Xia Lv. Improvement of page rankingalgorithm based on timestamp and link. *International Symposiums on Information Processing (ISIP)*, pages 36–40, 2008.

[4] Jingsheng Lei Houqun Yang and Fa Fu. An approach of multi-path segmentation clustering based on web usage mining. *Fourth International Conference on Fuzzy Systems and Knowledge Discovery*, 4:644–648, 2007.

[5] R.S. Bapi P. Kumar, P.R. Krishna 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*, pages 622–630, 2006.

[7] F. M. Facca and P. Luca Lanzi. Mining interesting knowledge from weblogs: A survey. *Data and Knowledge Engineering 53*, pages 225–241, 2005.

[8] M. A. Nacar M. S. Aktas and F. Menczer. Personalizing pagerank based on domain profiles. *Processing of WEBKDD Workshop*, 2004.

[9] S.K. De and P.R. Krishna. Clustering web transactions using rough approximation. *Fuzzy Sets and Systems*, pages 131–138, 2004.

[10] M. Deshpande J. Srivastava, R. Cooley and P. N. Tan. Web usage mining: Discovery and applications of usage patterns form web data. *SIGKDD Explorations*, 1(2):12–23, 2000.

[11] M.N. Murty A.K. Jain and P.J. Flynin. Data clustering: A review. *ACM Computing Surveys*, 31(3):265–323, 1999.

[12] W. Bin and L. Zhijing. Web mining research. *ICCIMA'03 IEEE*, pages 84–89, 2003.

[13] H. Blockeel R. Kosala. Web mining research: A survey. *SIGKDD Explorations*, 2(2).