INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND CREATIVE ENGINEERING (ISSN:2045-8711)
VOL.9 NO.6 JUNE 2019

A Comparative analysis of Various Techniques for Web Traffic and Web User Pattern with Space Complexity

Mr. N . Ulaganathan
Ph.D. (Part-Time) Research Scholar Department of Computer Science Nandha Arts and Science College Erode, Tamil Nadu, India
E-mail ID: ulaganathanjdk@gmail.com
Dr. S. Prasath
Research Supervisor & Ass.Professor, Department of Computer Science Nandha Arts and Science College Erode, Tamil Nadu, India
E-mail ID: softprasaths@gmail.com

Abstract: Classification is used to construct a representation for categorizing a set of pages. It is the process of mapping a page into various predetermined classes. In the Web domain, classification approaches are enabled to develop a profile of users related to a specific group and use particular server files. This in turn necessitated mining and collecting features depending on demographic data accessible on users or access patterns. The existing techniques such as Dominance Fuzzy Clustering and Distributed Probability Graph (, Map Reduce Pearson Correlation Fisher's Linear Discriminant Classifier (MPC-FLDC) technique and Poisson Fragment Frequency based Web Pattern Clustering technique are implemented in Java language by using Apache log samples dataset. Through the use of Apache log samples dataset in the experimental evaluation, web traffic patterns are effectively mined with the goal of tracking the location of web user. The proposed techniques are compared with existing methods. Keywords: Web Data, Clustering, Classifier, Space complexity.

I. INTRODUCTION

Web traffic is generally started through the use of web browsers. Traffic flow begins with a mouse click for delivering browser information to a server utilizing programmed rules and techniques to acquire user browser requests. Depending on these rules, the server chooses the type of action required. Web Traffic analysis is performed for maintaining and classifying the traffic. It also enhances the workload managing ability of the web server.

User communities are created by data collected from Web proxies while users browse the Web. Many hybrid representations are designed over time as search engines integrated directory features to address the problems like categorization and site quality. The key objective is to recognize the behavioral patterns in collected usage data and implement community Web directories depending on patterns. The method of collecting the patterns from data to web directories are called Usage Data Preparation.

Web traffic investigation applications are linked with large amount of data. Web traffic analysis is employed for extracting information and evaluating the performance for efficient input preprocessing. The data in the server logs are irregular, unrelated, noisy and unnecessary for an application of interest. Preprocessing of input data and feature collection approaches are utilized for choosing suitable attributes. The data networking world helps organizations to perform business by assisting companies to converse better with employees, customers and distributors.

II. RELATED WORKS

Marios Belk et al. [1] described user modeling mechanism for designing users cognitive styles depended on navigation patterns and click stream data. The gathering of users through measures acquired from psychometric tests and content navigation behavior with the aid of clustering methods are examined. Also, navigation metrics are employed in identifying specialized user groups with same navigation patterns associated to cognitive style. A psychometric- depended assessment is performed in mining the users cognitive styles. True positive rate is not calculated it decreases user modeling mechanism.

Mohammad Amin Omidvar et al. [2] considered the effectiveness of distinct variables on diverse dependent variables. These variables are times series and a time series regression is discussed. The time series regression is a significant and primary index on Google analytic. In addition, the most appropriate data provided to acquire outcomes of true positive rate has not been addressed.

Neha Goel et al. [3] analyzed Web Log Expert for discovering the user behavior accessing an astrology website. A comparison of accessible log analyzer tools is performed. Web Log Analyzer tools are sector of Web Analytics Software which accepts log file as input and examined input for producing outcomes. Web Log Expert considered web logs of the website and results are analyzed for inclusion in user website. It, in turn sustained in recognizing the customer behavior. The process of user behavior discovery is not accurate in terms of results.

II. METHODOLOGY

In order to overcome the limitations in the existing methodology proposed a method for performing effective web mining.

3.1 Dominance Fuzzy Clustering and Distributed Probability Graph Framework

Dominance Fuzzy Clustering and Distributed Probability Graph (DFC-DPG) framework is developed with the aim of extracting the similar web pages which is visited by user with improved clustering efficiency, less latency and space complexity. The implementation of proposed DFC-DPG framework contains four phases such as web user data collection, dominance rank model, fuzzy clustering approach and Distributed Probability Graph Arc model. The DFC-DPG framework is used in order to discover the information about the activities of web user from weblog data base. Initially, the web user data are collected by using server log files. The development of Dominance Rank model in proposed DFC- DPG framework separates the relevant and irrelevant web user data. Following this, fuzzy clustering is carried out on relevant data to form the cluster which contains the users with similar access sequence. Finally, Distributed Probability Graph Arc (DPG) model examines the access history of web user for predicting the future access of web user. Due to this model, the cache utilization and latency are minimized in a significant manner. The description about the process of proposed DFC-DPG framework is discussed below.

3.2 Web User Information Collection

During the implementation of proposed DFC-DPG framework, web user information collection is carried out as first process. The proposed DFC-DPG framework collects the information of web user through the sever log files from the web server data base. The web server log file is considered as text file which contains one line for each web user queries. Every line in log file includes information such as host making the request, timestamp, requested URL, HTTP reply code and bytes in reply which is visited by the user. The other log file is considered as Parse Log which is obtained from web server log file. The obtained files contained IP address, hostname, date, time and request. These data is stored on a web database for successfully handling the data in an effective manner.

3.3 Linear-Temporal Logic Model Checking Approach

Sergio Hernandez [16] developed linear-temporal logic (LTL) model checking approach for investigating structured e-commerce web logs. Based on the mapping log records with e-commerce structure and web logs were changed into event logs to extract the user behavior. Various predefined queries were developed to recognize behavioral patterns of a user during sessions. Certain enhancements in the website design were made to improve its performance efficiency. The product classification and the potential of users assisted in navigating website with respect to such association.

A number of query patterns were changed into LTL formula to enable the extraction of significant correlations among sequences of events acquired from user behaviour. This helped in identifying how various website sections are visited and navigational patterns are associated to buying actions. Several problems, issues and enhancements with respect to product categorization and organization of website sections were resolved. LTL model was also capable of

executing in parallel with the aid of parallel servers. But, it was not a sufficient model to perform effective traffic pattern mining for web user tracking.

3.4 Proposed Technique

The web usage mining approach was implemented to predict the online navigational behavior of web users but it failed to perform the effective prediction of web traffic patterns at the required level. However, a novel method was implemented with the objective of providing better results in the web usage pattern detection by the implementation of client-side logging. It failed to minimize the time consumption for detecting the web usage patterns. Hence, the proposed Map Reduce Pearson Correlation Fisher's Linear Discriminant Classifier technique is introduced with the objective of effectively predicting the web traffic patterns from weblog database with improved accuracy and less time. In the proposed technique, the frequent or the non frequent web patterns on weblog database are effectively classified with higher accuracy by using Fisher's Linear Discriminant Classifier. Thus, the performance of Pearson Correlation Analysis effectively predicted the web traffic patterns with minimized time consumption.

Then, the proposed MPC-FLDC technique is carried out to analyze the web traffic pattern analysis within three phases such as preprocessing, Fisher's Linear Discriminant Classifier and Pearson Correlation Analysis. After performing the web pattern classification, the proposed MPC-FLDC technique carried out Pearson Correlation Analysis in order to achieve effective web traffic pattern prediction (daily/hourly traffic) in weblog database. The web traffic patterns are represented as the web pages which are browsed more number of times by a web user. With the classified frequent web patterns, the web traffic patterns are predicted by using Pearson Correlation Analysis. Through the Pearson Correlation Analysis, the degree of web pages correlation is estimated among different sessions in order to perform web traffic predictions in a significant way.

The establishment of Pearson Correlation computes the similarity of web pages between the user sessions. From the determined degree of correlation i.e. similarity, the daily and hourly traffic volume are predicted in an efficient manner. As a result, the prediction rate is enhanced while performing the web traffic pattern mining on weblog database.

III. EXPERIMENTATION AND RESULTS

An effective Clustering framework is implemented in Java language using Apache log samples dataset. The Apache log samples datasets identifies the access activities of several web users namely IP address, Date, Time of Access, Port Number and accessed Web page. The tables and the graphs generated depend on the performance values obtained from experiments to assure the effectiveness of the proposed technique.

4.1 Performance Analysis of Space Complexity

The space complexity is defined as the amount of memory space required to store the similar web pages from the web server log files. The space complexity is measured as the difference between the entire memory space and the unused memory space on weblog database. The mathematical expression of space complexity is given as

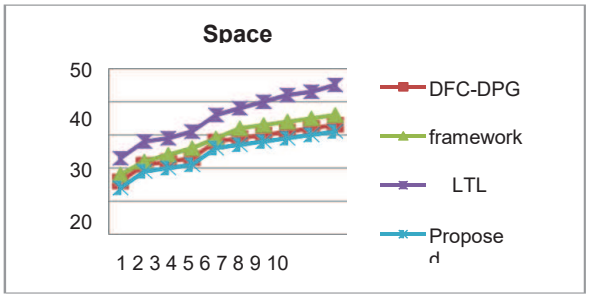
$$SC = \text{Total NeNory space} - \text{unused NeNory space} \quad \ldots\ldots\ldots\ldots\ldots\ldots(4.1)$$

In the above equation (4.1), the space complexity is represented as 'SC' which is measured in terms of Mega Bytes (MB). The lower value of space complexity enhanced the performance of DFC-DPG framework.

Table 4.1 Space Complexity

Number of web patterns	DFC-DPG framework	LTL based model	Fuzzy Clustering	Proposed technique
30	16	18	23	14
60	21	22	28	19
90	22	24	29	20
120	23	26	31	21
150	28	29	36	26
180	29	32	38	27
210	30	33	40	28
240	31	34	42	29
270	32	35	43	30
300	33	36	45	31

Fig. 4.1 Space Complexity



According to the different number of web patterns, the experimental result of space complexity is determined as shown in table 4.1. While carrying out the experiment, the number of web patterns considered ranges from 30 to 300 which are taken as input. After the experiment, the proposed is compared with the existing methods for analyzing the results of the space complexity. From table 4.1, it is comparatively proposed framework needs less memory space to store the web pages than the other existing methods. In the above Fig.4.1explains the performance analysis of space complexity with the number of web patterns for proposed framework and three existing methods. This proposed framework relatively consumed less memory space for storing web pages when compared with the existing methods.

I. CONCLUSION

The proposed framework performing the web user data analysis in an effective manner. The fuzzy clustering approach is carried out on the obtained relevant data regarding the user to form the clusters with similar user interest web pages. The DPG model is performed for changing the web user session into the graph which results in the reduction of latency. The experiment result shows that the proposed DFC-DPG framework groups the similar user interest web pages with the improvement of clustering efficiency with other existing methods. An effective PFF- WPC technique is implemented to track the web user location through the performance of web traffic pattern mining. By carrying the Poisson fragment process, the web pages are grouped at different sessions which result to attain effective web user tracking.
Through the deployment of frequency-based web patterns clustering, frequent or non-frequent web patterns are clustered from web pages. For the detected frequent web patterns, temporal similarity is determined to find the web traffic patterns. In this proposed technique, the clustering efficiency has improved and computational complexity is reduced for web user when compared to existing methods.

REFERENCES

[1] Marios Belk, Efi Papatheocharous, Panagiotis Germanakos and George Samaras, "Modeling users on the World Wide Web based on cognitive factors, navigation behavior and clustering techniques", Journal of Systems and Software, Elsevier, Vol. 86, Iss. No:12, Pp. No. 2995-3012, 2013.

[2] Mohammad Amin Omidvar, Vahid Reza Mirabi and Narjes Shokry, "Analyzing the Impact of Visitors on Page Views With Google Analytics", International Journal of Web & Semantic Technology, Vol. 2, Iss. No:1, Pp. No. 14-32, 2011.

[3] Neha Goel and C.K. Jha, "Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool", International Journal of Computer Applications, Vol. 62, Iss. No:2, Pp. No. 29-33, 2013.

[4] Salah Sleibi Al-Rawi, Rabah N. Farhan and Wesam I. Hajim, "Enhancing Semantic Search Engine by Using Fuzzy Logic in Web Mining", Advances in Computing, Vol. 3, Iss. No:1, Pp. No. 1-10, 2013.

[5] Christian Banse, Dominik Herrmann and Hannes Federrath, "Tracking Users on the Internet with Behavioral Patterns: Evaluation of Its Practical Feasibility", Information Security and Privacy Research, Springer, Pp. No. 235-248, 2012.

[6] Anurag kumar, Vaishali Ahirwar and Ravi Kumar Singh, "A Study on Prediction of User Behavior Based on Web Server Log Files in Web Usage Mining", International Journal Of Engineering And Computer Science, Vol. 6, Iss. No:2, Pp. No.20233-20236, 2017.

[7] Tawfiq A. Al-asadi, Ahmed J. Obaid, Rahmat Hidayat and Azizul Azhar Ramli, "A Survey on Web Mining: Techniques and Applications", International journal on advanced science engineering information technology, Vol.7, Iss. No:4, pp. No. 1178-1184, 2017.

[8] Peng-Yeng Yin and Yi-Ming Guo, "Optimization of multi-criteria website structure based on enhanced tabu search and web usage mining", Applied Mathematics and Computation, Elsevier, Vol. 219, Pp. No. 11082– 11095, 2013.

[9] Rupinder Kaur and Kamaljit Kaur, "An Improved Web Mining Technique to Fetch Web Data Using Apriori and Decision Tree", International Journal of Science and Research (IJSR), Vol. 3, Iss. No:6, Pp. No. 2094- 2098, 2014.

[10] V.V.R.Maheswara Rao and Valli Kumari, "An Efficient Hybrid Successive Markov Model for Predicting Web User Usage Behavior using Web Usage Mining", International Journal of Data Engineering, Vol. 1, Iss. No:5, Pp. No. 43-62, 2011.

[11] Amit Vishwakarma and Kedar Nath Singh, "A Survey on Web Log Mining Pattern Discovery", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5, Iss. No:6, Pp. No. 7022-7031, 2014.

[12] Vedpriya Dongre and Jagdish Raikwal, "An Improved User Browsing Behavior Prediction Using Web Log Analysis", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol. 4 Iss. No:5, Pp. No. 1838- 1842, 2015.

[13] S.Padmaja and Ananthi Sheshasaayee, "Clustering of User Behavior based on Web Log data using Improved K-Means Clustering Algorithm", International Journal of Engineering and Technology (IJET), Vol. 8, Iss No:1, Pp. No. 305-310, 2016.

[14] Mohammed Asad and Girish P. Potdar, "A Survey on Different Clustering Techniques for Web Usage Mining", International Journal of Computer Science and Information Technology & Security, Vol. 6, Iss. No:2, Pp. No. 200-204, 2016.

[15] S.Vijaya Kumar, A.S.Kumaresan and U.Jayalakshmi, "Frequent Pattern Mining in Web Log Data using Apriori Algorithm", International Journal of Emerging Engineering Research and Technology, Vol. 3, Iss No:10, Pp. No. 50-55. 2015.

[16] Sergio Hernández, Pedro Álvarez, Javier Fabra and Joaquín Ezpeleta, "Analysis of Users' Behavior in Structured e-Commerce Websites", IEEE Access, Vol. 5, Pp. No. 11941 – 11958, 2017.