INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND CREATIVE ENGINEERING (ISSN:2045-8711)
VOL.9 NO.5 MAY 2019

AN PERFORMANCE COMPARISON ON SPACE COMPLEXITY OF WEB USER TRACKING FOR CLUSTERING AND CLASSIFIERS

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- Web usage mining is the process of examining the web access logs navigation patterns which comprise browsing behaviors of all users over web. The activities of each user on web are stored in the form of weblog files or weblog database. The access activity on web signifies that the number of web pages and the number of times are sequentially visited by the user at different sessions. Through the examination of behavioral navigation patterns, the traffic patterns are mined from weblog database and the future access of the web user is predicted as well as the location of web user is tracked in a significant manner. During the web user behavior analysis, the mining of web traffic pattern is a challenging task. Also, the lack of web traffic pattern mining leads to reduce the performance in the identification of web user location. For the extraction of traffic patterns, the machine learning of clustering and classification techniques are utilized in the mining process to provide accurate results. With this intention, the proposed research work is implemented with web user by effectively mining the patterns from weblog

database. The clustering was developed with the aim of predicting the frequent web pages on weblog database browsed by a user but the prediction time remained unaddressed. In order to address the existing issues, three proposed techniques Clustering and Classifier technique based Web Pattern Clustering technique are implemented. The goal of attaining effective web data usage analysis by achieving higher clustering efficiency with less latency. At the beginning process of proposed method to collect the information of all users from weblog database by using server log files. Further, clustering approach is employed to perform similar user interest web pages from the obtained relevant the space complexity.

Keywords: Web Data, Clustering, Classifier, Space complexity.

I.INTRODUCTION

An existing web usage mining approach predicts the online navigational behavior and failed to provide efficient results on prediction performance of web user behavior. The main aim of web service ranking approach is developed with collaborative filtering to identify the potential user behavior after determining user behavior with their past access. The performance of clustering is not carried out in an effective manner. An existing Hybrid Sequence Alignment Measure (HSAM) is implemented in order to estimate the distance among session pair by considering the user navigated paths but it failed to minimize the latency for analyzing the web user data.

The aim of addressing the existing issues related to the prediction rate and time. In the initial step, the web patterns in a weblog database are grouped based on the different session by performing preprocessing. By grouping the web patterns into a number of sessions based on access time, the time consumption for performing effective web traffic pattern mining is reduced. Further, the classifier is performed in order to classify the web patterns as frequent or non-frequent patterns by measuring the hit ratio. Finally, the analysis determines the correlation of web patterns on different

user sessions to predict the web traffic patterns with higher accuracy.

II. RELATED WORKS

Amit Dipchandji Kasliwal et al. [3] discussed web usage mining method. The process of deletion is carried out on insignificant data and mining log file is provided with mining tool for obtaining adapted access and executed this by making different user to visit the website through processing. Web usage mining is utilized for frequent model in which user visits the website for managing website structure and recommends for users but Web usage mining method failed to improve their true positive rate.

Pablo Loyola et al. [5] designed an ant colony optimization-based algorithm for identifying web usage patterns. A number of data sources namely web content and structure and web usage is incorporated. This followed continuous learning approach is where artificial ants attempt

to fit their sessions with actual sessions by alteration of a text preference vector. The trained ants are set to free onto a web graph and artificial sessions are compared with actual sessions. An exact identification of aggregated patterns of actual usage is achieved and quantitative representation of keywords is attained with navigational sessions. The response time are not reduced to desired level.

Preeti Sharma et al. [6] considered web mining techniques for attaining a viable edge in business. Web mining is employed for electronic means of functions to perform business. Web mining is relevant to the data mining approaches employed for identifying patterns from web by means of content mining, structure mining and usage mining. A competitive merit was achieved in business by web mining but Web traffic prediction did not facilitate in improving the efficiency.

Rekha Jain et al. [7] considered Page Rank, Weighted Page Rank and Hyper-link Induced Topic Search (HITS) algorithms for ranking web pages. Page Rank and Weighted Page Rank are Web Structure Mining. HIT was utilized in both structure Mining and Web Content Mining. The score at indexing time was estimated by Page Rank and Weighted Page Rank for arranging with respect to page significance. HITS could evaluate the hub and ability score of appropriate pages but the computational complexity is not reduced for attaining efficient results.

Lu Dai et al. [8] considered efficient particle swarm chaos optimization mining approach. The chaos optimization mining approach employed feedback model of user for offering superior-matching web pages for user. An initial population of particles moving in D-dimensional search space is considered. Every particle vector is represented to a probable resolution through a subset of web pages. Though the performance of chaos optimization approach is evaluated in terms of response time, execution time, precision and recall for achieving better performance, chaos optimization approach failed in optimization issues.

Maryam Jafari et al. [9] analyzed Web Usage Mining (WUM) and pattern extraction approaches. The patterns obtained after discovery are employed in pattern investigation phase. Followed by investigation of Web user navigational patterns, user behaviors and Web structure are recognized for modeling improved Web machinery and Web relevance. Classification accuracy is minimized by using Web Usage Mining.

Satpal Singh et al. [10] explained web usage mining for user identification. Web mining with distinct approaches are proposed for various applications. The classification of web-user was discovered with various dimensions of temporal web mining. Clustering efficiency got minimized by using web usage mining.

Sergio Hernandez et al. [11] described linear-temporal logic model checking technique for structured e- commerce web logs. E-commerce configuration based general mapping log records and web logs are transformed into event logs in which the user behavior is extracted. Various predefined queries are carried out for discovering distinct behavioral patterns with user performance at the time of a session. Specific enhancements are made in website modeling to improve their efficiency. However, Computational complexity is

high when compared to other conventional checking methods.

Shilpa Mahajan et al. [12] discussed user behavior pattern investigation through the estimation of web users in websites. The essential data source in web browsing is web logs which accumulates the user behavior on web pages. The produced logs are examined in phases and classification methods are implemented for identifying the upcoming user behavior. The identified data are employed by E-commerce organizations for recognizing customer necessities to enhance the website data and associations. In addition, E-commerce companies also employed the user behaviors for recognizing customers and employee actions. However, the quality of extracted data decreased.

Tania Cerquitelli et al. [13] discussed Mining Neubot Data (MiND) for examining the features of periodic internet measurements gathered at end user location. MiND is enabled for identifying group of users with an identical internet access behavior and anomalous service. The user measurements are designed by histograms and two-level clustering strategy. A maximum set of users are collected into homogeneous and cohesive clusters with respect to internet access service and users with anomalous services are represented as outliers. Internet Service Provider (ISP) is observed by the users in which ISP effectively discovered anomalous behaviors and acts respectively but MiND failed in reducing the prediction time to desired level.

Vagner Figueredode Santana et al. [14] discussed an identification tool for discovering usage patterns which depended on client-side event logs and event stream composition distinctiveness. A system is employed in recording the usage information at actual utilization and usage patterns are predicted for addressing probable user interface design issues. In addition, discovery of usage patterns and categorization of event streams is performed. The time consumed for identifying the users behavior is increased.

Zheng Xu et al. [15] suggested personalized web search using semantic context. The technique collected user context to present accurate preferences of users in personalized search. The short-term query context was generated to identify related concepts of query. The user context was produced depended on click through the data of users. A forgetting factor was developed for combining the self-governing user context in user session to preserve the evolution of user preferences. Clustering and classification methods of web pages were not included to get accurate outcomes.

III. METHODOLOGY

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

3.1 Web Usage Mining Approach

Abdelghani Guerbas et al. [1] designed Web Usage Mining Approach for web log mining and online navigational behavior prediction. The design of Web Usage Mining Approach includes different phases such as data cleaning and preparation, density based clustering and online pattern prediction. In the initial phase, the approach allows the raw log data and cleans it to provide page views. In this phase, the designed approach uses the time-based heuristic for session's identification to obtain better quality results.

Then, density based clustering algorithm is developed to mine for detecting navigational patterns. Instead of association rules detection and sequential patterns mining techniques, clustering algorithm is employed for designed approach as these techniques are unable to get low frequent and meaningful patterns. In addition, the clustering algorithm was highly sensitive to the input metric (minimum support) and finds the outliers. Finally, efficient online prediction approach called k- Nearest-Neighbors (kNN) approach is designed for obtaining relevant

sessions. The online pattern prediction was very useful. The k-Nearest-Neighbors helped to minimize the server processing time through caching pages where the pages are demanded by a user. It also recommended products, links, online services and so on. Based on the concept of detecting sessions as documents, page references and Web Usage Mining Approach have been designed. This helped to perform pattern prediction of an online session for most relevant documents to a query with higher accuracy. However, the process of clustering could not be carried out in an efficient manner.

3.2 Web Service Ranking Method

Guosheng Kang et al. [4] developed an effective Web service ranking approach for analyzing the user log. Depending on the collaborative filtering (CF), the ranking approach is developed through examining the user behavior. This was done by considering invocation and query history to gather the potential user behavior. With the aid of CF, the similarity between related invocations and related queries are determined. During collaborative filtering, if two users invoke the identical Web services, it is considered to be related to some degree. In case, if the invocations are from similar queries including functional and QoS queries, the value of user similarity tends to be higher, since it denoted similar usage behavior pattern or similar intention from the two users. In Web service selection system, if the user requests a web service, then the services are ranked based on the similar behavior.

Different web services like functional significance, CF based score and Quality of Service (QoS) applicability web services are able to perform efficient web service ranking. The CF based score of a web service is obtained according to the historical functional query, and the recent functional query. Functional relevance was obtained depending on the current functional query and web Services of users. QoS applicability of a web service was achieved based on recent QoS query and its information. Finally, the rank aggregation procedure is employed to integrate the web services to ignore the impact of range and distribution of variables for web service ranking. The rank aggregation generates the ranking score which appeared in the top-k web service ranking record to the user. However, the developed approach failed to efficiently extract the relevant information from web log.

3.3 Hybrid Sequence Alignment Measure Method

Poornalatha et al. [POO2013] developed Hybrid Sequence Alignment Measure (HSAM) method to discover the distance among user sessions. The main aim of HSAM method was to find out the distance between a pair of sessions on the user navigation paths for web session clustering and assess the quality of clustering. The sequence alignment is denoted as the association of two sessions. The hybrid distance measure considers navigation path information in order to calculate the distance between two sessions without changing the order. The statistic measure is employed in HSAM method to make a decision regarding the number of clusters to be constructed. Jaccard Index and Davies–Bouldin validity index are applied to evaluate the clustering process. The results obtained through these standard statistic indices prove the goodness of the HSAM method.

3.4 Fuzzy Clustering Technique

Anandhi [2] developed Fuzzy Clustering techniques for detecting patterns like path detection, page aggregation, fuzzy clustering, ant-based clustering and graph separation, etc. Fuzzy Clustering is employed to help the web administrators and web users in order to detect and extract required information in an effective way. The developed Fuzzy Clustering includes preprocessing, user identification, classification and clustering phase. In preprocessing phase, raw data cleaning is carried out to eliminate the unnecessary data and for detecting user and session. The detection of feasible users

minimizes the processing time for robot entries since it does not contain any irrelevant data. In the next phase, classification techniques are employed to categorize the three types of users from log files. The developed technique classifies whether the user is a frequent user, synthetic user or potential user from web log data. Through the classification process, the attributes or class in a web log data like time stamp and users are taken for detecting the class. Moreover, the potential user was taken for identifying navigation pattern. Finally, based on the sequence of primarily accessed pages, the clustering techniques are applied to group the next page which is accessed through the web site user. Thus, clustering provides maximum prediction accuracy while predicting navigation pattern. However, the time taken for predicting web patterns failed to minimize.

3.5 Proposed Methodology

Improved K-means clustering algorithm is developed for finding the browsing activities of user on web. The efficiency for performing the clustering did not increase and the space complexity remained unaddressed. Besides, Hybrid Sequence Alignment Measure (HSAM) was implemented with objective of detecting the distance between any two sessions by utilizing the access path information on website and the reduction of latency remained unaddressed. 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. The proposed Clustering 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. Hence, the proposed 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 Classifier. The performance is effectively predicted the web traffic patterns with minimized time consumption.

IV. 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 performance evaluation of proposed method is compared with the existing Web usage mining approach, Web service ranking approach and HSAM method. 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 a

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

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.

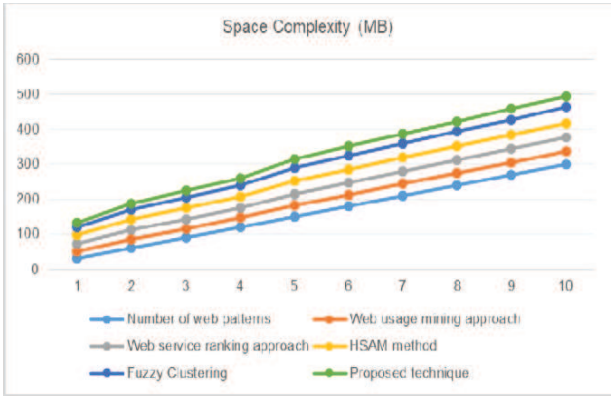
Table 4.1 Performance of Space Complexity

Space Complexity (MB)					
Number of web patterns	Web usage mining approach	Web service ranking approach	HSAM method	Fuzzy Clustering	Proposed technique
30	20	22	25	23	14
60	25	27	30	28	19
90	26	28	31	29	20
120	27	29	32	31	21
150	32	34	37	36	26
180	33	35	38	38	27
210	34	36	39	40	28
240	35	37	40	42	29
270	36	38	41	43	30
300	37	39	42	45	31

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 method is compared with the existing methods for analyzing the results of the space complexity. From table 4.1 shows that the four methods could successfully reduce the space complexity for storing the similar web pages. Comparatively, the proposed method needs less memory space to store the web pages than the other existing methods.

Fig.4.1 Performance of Space Complexity



V. CONCLUSION

The proposed method during web user data extraction from the web database. Through this method user with the visited pages is examined in a sequence manner and the relevant web pages from the web server log files to the web user are stored in memory with less space consumption. The result is that the space consumed for storing web pages is effectively reduced in the proposed method by 14% and 19% when compared to Web usage mining approach and Web service ranking approach. Similarly, the proposed method reduced the space complexity by 26% when compared to HSAM method and Fuzzy Clustering.

REFERENCES

[1] Abdelghani Guerbas, Omar Addam, Omar Zaarour, Mohamad Nagi, Ahmad Elhajj, Mick Ridley and Reda Alhajj, "Effective Web log mining and online navigational pattern prediction", Knowledge Based Systems,Elsevier,Vol.49, Pp.No.50-62, 2013.

[2] D. Anandhi and M. S. Irfan Ahmed, "Prediction of user's type and navigation pattern using clustering and classification algorithms", Cluster Computing, Springer, Pp.No.1-10, 2017.

[3] Amit Dipchandji Kasliwal and Girish S. Katkar, "Web Usage mining for Predicting User Access Behaviour",International Journal of Computer Science and Information Technologies, Vol. 6 , Iss. No:1, Pp No.201-204,2015.

[4] Guosheng Kang , Jianxun Liu, Mingdong Tang , Buqing Cao and Yu Xu, "An Effective Web Service Ranking Method via Exploring User Behavior", IEEE Transactions on Network and Service Management, Vol.12,Iss.No:4,Pp.No.554-564,2015.

[5] Pablo Loyola, Pablo E.Roman and JuanD.Velasquez, "Predicting web user behavior using learning-based ant colony optimization", Engineering Applications of Artificial Intelligence, Elsevier,Vol.25, Iss.No:5,

Pp. No.889-897, 2012.

[6] Preeti Sharma and Sanjay Kumar,"An Approach for Customer Behavior Analysis Using Web Mining", International Journal of Internet Computing, Vol. 1, Iss. No: 2, Pp. No. 1-6, 2011.

[7] Rekha Jain and G. N. Purohit, "Page Ranking Algorithms for Web Mining", International Journal of Computer Applications, Vol. 13, Iss. No:5, Pp. No. 22-25, 2011.

[8] Lu Dai,WeiWang and Wanneng Shu, "An Efficient Web Usage Mining Approach Using Chaos Optimization and Particle Swarm Optimization Algorithm Based on Optimal Feedback Model", Hindawi Publishing Corporation, Mathematical Problems in Engineering,Vol.2013,

Pp.No.1-8, 2013.

[9] Maryam Jafari, Farzad SoleymaniSabzchi and Shahram Jamali,"Extracting Users'Navigational Behavior from Web Log Data: a Survey", Journal of Computer Sciences and Applications,Vol. 1, Iss. No.:3, Pp. No. 39-45,2013.

[10] Satpal Singh and Vivek Badhe, "An Exclusive Survey on Web Usage Mining for User Identification",International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Iss. No:11,Pp. No. 6852-6859, 2014.

[11] 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.

[12] Shilpa Mahajan and Shilpa Yadav, "Analyzing HTTP Traffic Patterns for Monitoring and Analyzing User Behavior", Indian Journal of Science and Technology, Vol. 9, Iss. No:48, Pp. No. 1-7, 2016.

[13] Tania Cerquitelli , Antonio Servetti and Enrico Masala, "Discovering users with similar internet access performance through cluster analysis", Expert Systems with Applications, Elsevier,Vol.64,Pp.No.536–548, 2016.

[14] Vagner Figueredo de Santana and Maria Cecília Calani Baranauskas, "WELFIT: A remote evaluation tool for identifying Web usage patterns through client-side logging", International Journal of Human-Computer Studies, Elsevier, Vol. 76, Pp. No. 40-49, 2015.

[15] Zahid Ansari, Syed Abdul Sattar, A. Vinaya Babu and M. Fazle Azeem, "Mountain density-based fuzzy approach for discovering web usage clusters from web log data", Fuzzy Sets and Systems, Elsevier, Vol. 279, Pp.No. 40–63, 2015.