International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 10 (2019) pp. 2339-2346

An Analysis of Web User Behavior using Hybrid Algorithm based on Sequential Pattern Mining

Sheetal Sahu, Rajendra Gupta, Amit Dutta

Rabindranath Tagore University, Raisen, Madhya Pradesh (M.P.), India.

All India Council for Technical Education (AICTE), New Delhi, India.

Abstract

An Organization need to understand their customers' behavior, preferences and future needs which depend upon past behavior. Web Usage Mining is an active research topic in which customers session clustering is done to understand the customers activities. It investigates the problem of mining frequent pattern and especially focuses on reducing the number of rules using closed pattern technique. It also reduces scans the size of the database using the SOM clustering technique. It solves the problem through Profile based Closed Sequential Pattern Mining using SOM Clustering (PCSPSC). The user access web pages which consist of distributed environmental patterns and these web pages are accessible by the user in some patterns. These patterns are combined and finding a closed frequent set of web pages. If the user needs next request page in advance, then it searches only partial web data not in whole web data. There is no need to take input as number of clusters. So that this research utilized a personalized weighted recommendation system based on the user's interest with less execution time.

Keywords: Web Usage Mining, Prefix Span, Gap, Recency, Compactness, Data Stream, Closed Pattern, Data Mining, Personalization, Sequential Pattern Mining, Web Services, Self Organizing Map.

I. INTRODUCTION

The popular medium of publishing is the World Wide Web is a very rich source of information gathering. It makes sense, of data is difficult because publication on the web is largely unorganized. Web mining is also knowledge extraction techniques which discover access patterns from the web. It is divided into three parts, a) web usage mining, b) web structure mining and c) web content mining. The commonly used data mining algorithms are Association Rule Mining (ARM), Sequential Pattern Mining, Clustering, and Classification. An ARM technique is used to find out the rules between items found in a transaction database. In the context of web usage mining a transaction is a cluster of web page accesses with an item being a single page access. The problem of discovering sequential patterns is that of finding inter-transaction patterns such that the presence of a set of items is followed by another item in the timestamp ordered transaction set. The data mining algorithms are used to generate the association rule between the items, sequential pattern of access of items, and clustering of items.

Web Usage Mining (WUM) is the application of data mining techniques to large Web data repositories in order to produce results that can be used in the design tasks and improve response time.

ARM techniques, discover the items which having the unordered correlations in a transaction database. In the context of WUM a transaction is a group of Web page accesses, with an item being a single page access. The user finds problems in inter-transaction patterns for discovering sequential patterns, due to the presence of a set of items which is followed by another item in the timestamp ordered transaction set.

Clustering analysis is used to find out those items that have similar characteristics and group into it. It manages the group of user information or data from Web server logs. It also can facilitate the development and execution of future marketing strategies. It dynamically supports or changing a particular site based on a visitor on a return visit.

The rules of association or sequential patterns were not only discovered for adapting existing algorithm, but it is a process of data mining. The WUM is a process which consists of input from web user, which shows the behavior of user in the context of user session files that gives the information regarding the accessed website.

It is also having the information just like what pages were requested and in what order, and how long each page was viewed. A user session is a time interval where a web user accesses the pages that occur during a single visit to a website. The web users access related all the information contained in a raw web server log. It does not reliably represent a user session file for a number of reasons. So that selectively information converts into tabular form and after that apply data mining technique. After getting the result it produces some meaningful and useful information.

II. RELATED WORK

Bing Zhang and Guoyan Huang [33] proposed an efficiently mine influential function for software execution. First, the authors design a novel modelling strategy in which traces of software execution are modeled as sequential patterns. Because of loops, patterns can occur many times in a trace, which leads to high cost of time and extreme complex difficulty of the research.

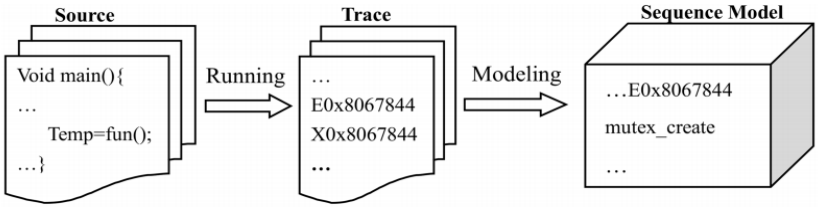

Fig. 1: Sequence Data Modeling Process

Then, an algorithm is designed to remove repetitive patterns in software and software influential nodes mining algorithm is put forward to mine influential functions in software and to rank them by the rank-index. Finally, by comparatively analyzing the top ten functions got from PageRank and those from Degree-Based algorithm, the approach has proven to be an effective and accurate one which combines advantages of the these two classic algorithms.

Minubhai Chaudhari and Chirag Mehta propose a Prefixspan algorithm with GRC constraints which generates sequential patterns by using the prefix projected pattern growth approach is implemented. Other than frequency this algorithm also uses gap, compactness and recency constraints during sequential pattern mining process. The gap constraint applies limit on the separation of two consecutive transactions of discovering patterns, recency constraint makes patterns to quickly adapt the latest behaviors and compactness constraint make sure reasonable time spans for the discovered patterns.

Fan Muhan, Shao Sujie, and Rui Lan propose a method for miing the frequent closed patterns in a sliding window to capture information timely and accurately when new data stream arrives. Data stream is divided into several basic windows. All possible frequent closed patterns are mined in each basic window and be stored in a Closed Pattern - tree in the form of node compression to save space; As the data in sliding window updates, Closed Pattern-tree can be incrementally updated and the infrequent or unclosed patterns will be deleted from the tree.

Doddegowda B J, G T Raju, Sunil Kumar S Manvi proposed a Web Personalization process that adjusts information/ services delivered by a Web to the needs of each user or group of users, taking their behavioral patterns.

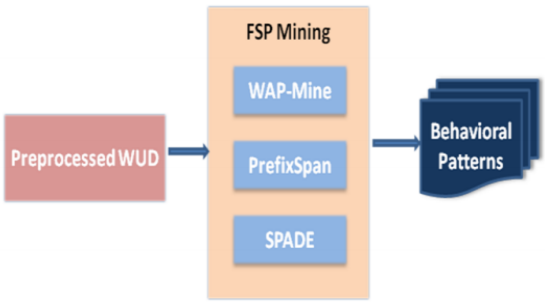

Fig. 2: System Architecture

Frequent Sequential Patterns (FSPs) that are extracted from Web Usage Data (WUD) are very important for analyzing and understanding users' behavior to improve the quality of services offered by the World Wide Web (WWW).

In 2013, Rahul Moriwal [25] - It presented a method for Finding Frequent Sequential Traversal Patterns from Web Logs which is based on Dynamic Weight Constraint, where various frequent sequential pattern mining algorithms have been proposed that mines the set of frequent subsequences pattern which satisfying a min. support constraint in a particular session database. Though, previously sequential pattern mining algorithms give equal weightage to sequential traversal patterns, whereas the pages in sequential patterns have different importance and also have different weightage. Another problem in most of the frequent sequential pattern mining algorithms is that a large number of sequential patterns are generated, when min. support is lowered and here they do not have any alternative ways of adjusting the number of sequential patterns other than increment in the minimum support. The pages are given dissimilar weights and traversal sequences assign a min. and max. weight. For scanning a session database, max. and min. weight in the session database is utilized to cut infrequent sequential subsequence and by this downward closure property is maintained.

Ketki Muzumdar, Ravi Mante, Prashant Chatur (IJRTE-2013) proposed "Neural Network Approach for Web Usage Mining" in which Web usage mining tries to discover useful knowledge from the secondary data obtained from the connections of the users with the Web. Web usage mining has become very dangerous for effective Web site management, business and support services, personalization, and network traffic flow analysis, etc. Earlier study on the Web usage mining using a concurrent Clustering, Neural based method has shown that the practice trend analysis very much depends on the performance of the clustering of the number of requests. In this paper, a novel method Self Organizing Map is introduced, which is a kind of neural network, in the process of Web Usage Mining to detect user's patterns. And analyze the traditional K-Means algorithm result with comparison to SOM. The process details the transformations necessaries to modify the data stored in the Web Servers Log files to an input of SOM.

C. Umapathi1, M. Aramuthan, proposed "Enhancing Web Services Using Predictive Caching" where the exponential growth of World Wide Web (WWW) users and network traffic, the development of new services require high bandwidth and also provides high perceived latency. Analyzing the behavior of a web site users also known as Web usage mining, is a research field which consists of adapting

the data mining methods to the records of access log files. Web usage mining method provides information about the activities of the client in order to extract relationships in the registered data. Each record in the log file contains the clients IP address, the requested object and additional information such as the protocol of the request, size of object, etc. The log file contains all the information of different users and incomplete information irrelevant data, noise and errors that need to be filtered out. Hence, preprocessing is required. After preprocessing, the pattern is discovered and analyzed. Based on this, pre-fetching techniques are used to improve the performance of web sites in turn reduce latency. Although, the most essential element in web prefetching is the algorithm based on prediction. The effectiveness of pre-fetching is based on the algorithm which in turn improves the performance of the web. A long-standing use of caching tries to improve the quality of service perceived in web browsing. This paper explains about the technique to perform prefetching. It also implements Prefetch enhanced caching algorithm and provides experimental results.

In 2013, Omar Zaarour, Mohamad Nagi [24] proposed an improvement to the web log mining procedure and to the prediction of online navigational pattern. Their contribution contains three different components. First, they proposed for session identification, a refined time-out based heuristic,. Secondly, suggested the practice for navigational pattern detection by using a specific density based algorithm. Finally, a new method for efficient online prediction is also recommended. The conducted experiment shows the applicability and effectiveness of the proposed method.

Qingqing Gan, Torsten Suel proposed the techniques used to optimize query processing performance. Author initial contribution is the study of outcome caching as a weighted caching problem. Mainly earlier work focused on optimizing cache hit ratios, however, given that processing costs of queries can differ very significantly and argue that overall cost savings also need to be considered. They described and evaluated some algorithms for weighted result caching, and study the influence of Zipf-based query distributions on outcome caching. The next main work is a latest set of feature-based cache eviction strategy to get significant improvements over all previous techniques, significantly narrowing the presented performance gap to the theoretically optimal method. Finally, using the same approach, they also acquire performance gains for the linked problem of inverted list caching.

Jatin D Parmar, Sanjay Garg proposed modified web access pattern (mWAP) method for sequential pattern mining. Web access pattern (WAP), is the sequence of frequently accesses practice by users, practice is of interesting and useful information. Sequential Pattern mining is the process of using data mining method for a sequential database for discovering the correlation relationships which presents with a structured list of events. Web access pattern tree mining is a sequential pattern mining process for web log access sequences, which first saves the original web access sequence database on a prefix tree, alike to the frequent pattern tree (FP-tree) which stores non-sequential data. Web access pattern tree (WAP-tree) method, then, mines the frequent sequences from the Web access pattern tree (WAP-tree) by recursively recreating intermediate trees, started with suffix sequences and ended with prefix sequences. An effort has been made to modify WAP tree method for improving efficiency. mWAP completely remove the want to engage in numerous reconstruction of intermediate Web access pattern trees (WAP-trees) during mining and considerably reduces time of execution.

In 2014, Jerry Chun [26] proposed the prelarge concept is adopted to handle the discovered sequential patterns with sequence deletion. An FUSP tree is first built to keep only the frequent 1-sequences from the original database. The prelarge 1-sequences are also kept in a set for later maintenance approach. When some sequences are deleted from the original database, the proposed algorithm is then performed to divide the kept frequent 1-sequences and prelarge 1-sequences from the original database and the mined 1-sequences from the deleted customer's sequences into three parts with nine cases. Each case is then processed by the designed algorithm to maintain and update the built FUSP tree. When the number of deleted customer sequences is smaller than the safety bound of the prelarge concept, the original customer's sequences are unnecessary to be rescanned, but the sequential patterns can still be actually maintained and updated.

In 2016, Doddegowda [30] having approached to personalize the information available on the Web according to user requirements. This is called Web Personalization process that adjusts information/services delivered by a Web to the needs of each user or group of users, taking their behavioral patterns. Frequent Sequential Patterns (FSPs) that are extracted from Web Usage Data (WUD) are very important for analyzing and understanding users' behavior to improve the quality of services offered by the World Wide Web (WWW). User behavioral patterns are required to build profiles for each user, using which Personalization of a website is made.

In 2016, Minubhai [31] proposed a prefixspan algorithm with GRC constraints which generates sequential patterns by using the prefix projected pattern growth approach is implemented. Other than frequency this algorithm also uses gap, compactness and recency constraints during sequential pattern mining process. The gap constraint applies limit on the separation of two consecutive transactions of discovering patterns, recency constraint makes patterns to quickly adapt the latest behaviors and compactness constraint make sure reasonable time spans for the discovered patterns.

In 2016, Fan Muhan [32] proposes a method for mining the frequent closed patterns in a sliding window to capture information timely and accurately when new data stream arrives. The data stream is divided into several basic windows. All possible frequent closed patterns are mined in each basic window and be stored in a Closed Pattern - tree in sliding window updates, Closed Pattern-tree can be incrementally updated and the infrequent or unclosed patterns will be deleted from the tree.

In 2017, Bing Zhang [33] proposed a new approach to efficiently mine influential functions based on software execution sequence is proposed. First, the authors design a

novel modeling strategy by which software execution traces are modeled as sequential patterns. Owing to loops, patterns can occur multiple times in a trace, which leads to high cost of time and the extreme complexity of the research. Then, an algorithm is designed to remove repetitive patterns in software and software influential nodes mining algorithm is put forward to mine influential functions in software and to rank them by the rank-index. Finally, by comparatively analyzing the top ten functions got from Page Rank and those from Degree-Based algorithm.

In 2017, H. Ryang [34] propose a novel algorithm and list structure for finding high utilization patterns over data streams on the basis of a sliding window mode. Unlike existing algorithms, the proposed algorithm does not consume huge computational resources for verifying candidate patterns because it can avoid the generation of candidate patterns. Therefore, the algorithm efficiently works in complex dynamic systems.

III. ANALYSIS OF PREVIOUS WORK

The following table shows the analysis of previous and current work –

S .N.	Authors	Title	Advantage	Disadvantage
1	Bing Zhang and Guoyan Huang	Approach to mine influential functions based on software execution sequence	An effective and accurate one which combines advantages of the Page Rank and Degree based algorithms.	It suffers from privilege protection, loss in a release pair when it was definitely protected on all execution paths.
2	Minubhai Chaudhari and Chirag Mehta	Extension of Prefix Span Approach with GRC Constraints for Sequential Pattern Mining	It provides latest behaviors with reasonable time spans for the discovered patterns.	The observed composition rules into the guessing rule set.
3	Fan Muhan, Shao Sujie, and Rui Lanlal	A Mining Algorithm for Frequent Closed Pattern on Data Stream Based on Sub Structure Compressed in Prefix-Tree	The window partitioning method to balance the time cost of mining closed patterns.	The verification of their opacity without need for the original models.
4	Doddegowda B J, G T Raju, Sunil Kumar S Manvi	Extraction of Behavioral Patterns from Preprocessed Web Usage Data for Web Personalization	It provide better services in the web of each user or group of users for their behavioral patterns.	The local conditional probability distribution of each node, which is calculated accordingly.

IV. PROPOSED APPROACH

The proposed Profile based Closed Sequential Pattern Mining using SOM Clustering (PCSPSC) approach is next applied to discovering frequent sequential patterns item in a cluster by using the Self Organization Map algorithm of Neural Network for producing the cluster of web data set. This cluster is used to access the partial web data set not whose web data set. So at this time this tree having the web pages of website in proportional sessions can access partially. To applied this method call the procedure Clustering_Method.

The following Fig. 3 shows that the process of PCSPSC algorithm which generate useful closed sequential pattern using web data.

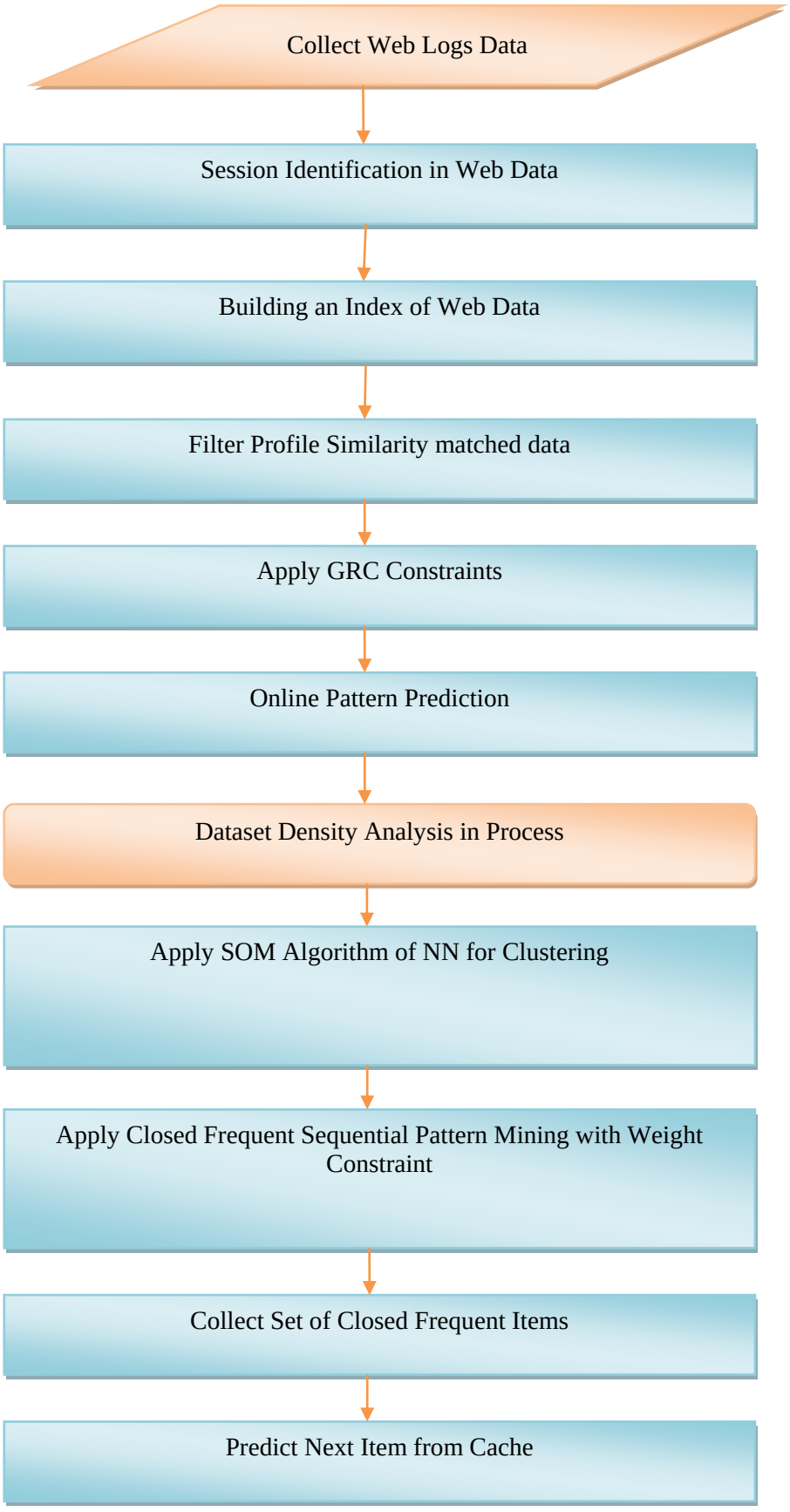

Fig. 3: The Process of PCSPSC Algorithm

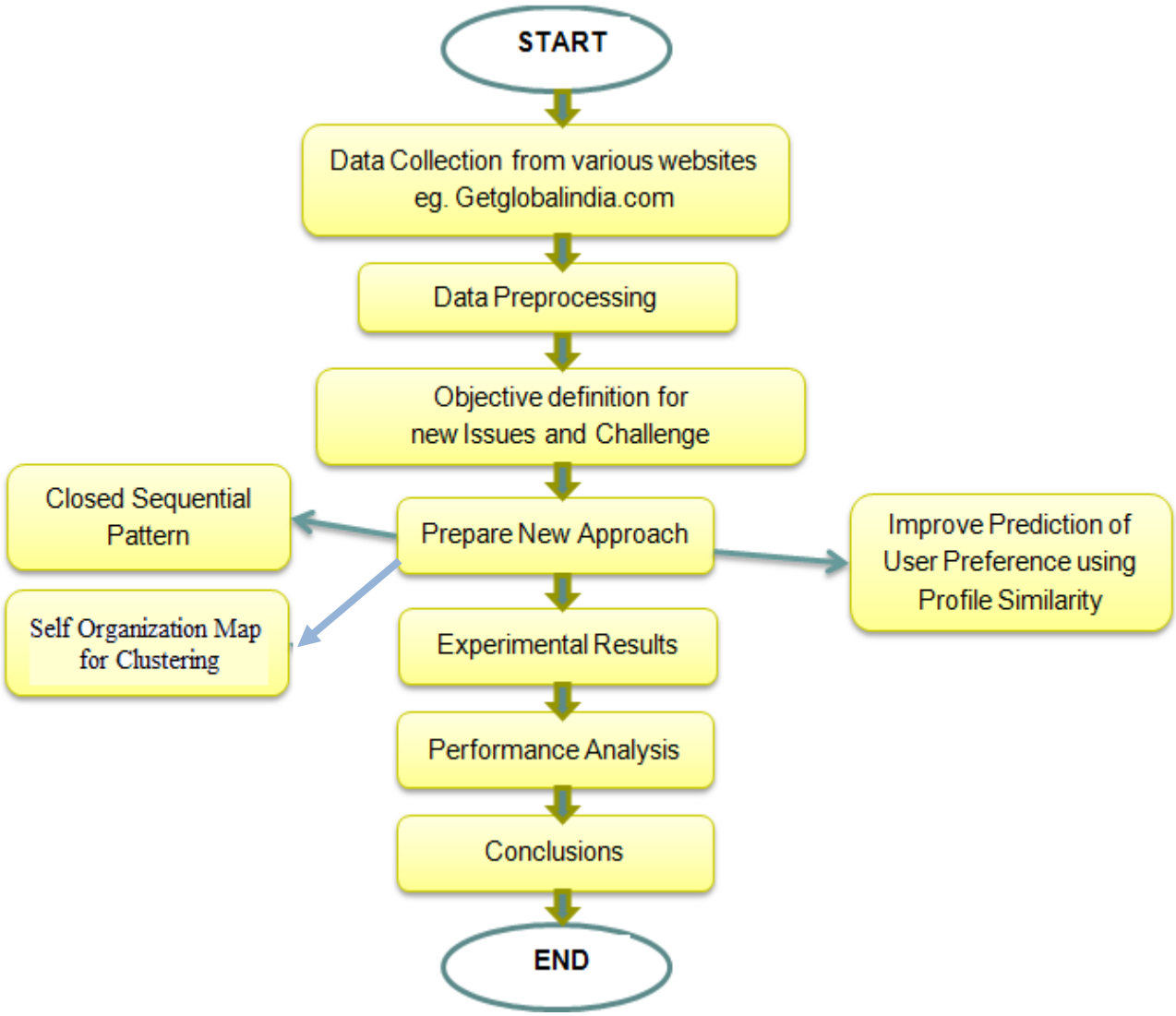

Fig. 4: Steps of Research Methodology in Research Area

V. CONCLUSION

In this research it scans only partial database not the whole database so that multiple scanning of the database will be reduced and response time is increased. It enhanced reflection of the importance of pages by using min-max weight and support of every page by using min-max weight of pages updating automatically by using web services.

It is identifying the user previous search subjects and topics so that current search will be more up to the point. So information gathered could be used to offer feedback to users on their use of the internet. It enables the effective tracking for the development and also improvement of the user interface in software by analyzing user behavior.

In future work, other data mining algorithms can be implemented in cloud to efficiently handle big data of many Hospital website in a distributed environment for finding any critical diseases.

REFERENCES

[1] R. Agrawal and R. Srikant, "*Fast Algorithms for Mining Association Rules in Large Databases*", in Proc. Int. Conf. Very Large Data Bases, pp. 487–499, 1994.

[2] R. Agrawal and R. Srikant, "*Mining Sequential Patterns*", In Proceedings of the 1995 International Conference on Data Engineering, pp. 3-14, 1995.

[3] R. Agrawal and R. Srikant, "Mining Sequential Patterns: Generalizations and Performance Improvements", In Proceedings of the 5th International Conference on Extending Database Technology, pp. 3-17, Avignon, France, 1996.

[4] M. N. Garofalakis, R. Rastogi, K. Shim, "SPIRIT: Sequential Pattern Mining with Regular Expression Constraints", In Proceedings of 25th VLDB Conference, pp. 223-234, San Francisco, California, 1999.

[5] M. J. Zaki, "SPADE: An Efficient Algorithm for Mining Frequent Sequences", Machine Learning Journal, Vol. 42, Issue (1-2), pp. 31-60, 2001.

[6] T.-P. Hong, C.-Y. Wang, and Y.-H. Tao, "A New Incremental Data Mining Algorithm using Pre-large Itemsets", Intell. Data Anal., vol. 5, no. 2, pp. 111–129, Apr. 2001.

[7] Jian Pei, Jiawei Han and Helen Pinto, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth", In Proceedings of 12th International Conference on Data Engineering, pp. 215-224, Heidelberg, Germany, 2001.

[8] Freire J., Kumar B., and Lieuwen D., "WebViews: Accessing Personalized Web Content and Services", In Proceedings of the Tenth International World Wide Web Conference, 2001.

[9] Antunes, A. L. Oliveira, "Generalization of Pattern-growth Methods for Sequential Pattern Mining with Gap Constraints", Machine Learning and Data Mining in Pattern Recognition, Third International Conference, MLDM 2003, Leipzig, Germany, July 5-7, 2003, Proceedings 2003.

[10] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach", Data Mining Knowledge Discovery, vol. 8, no. 1, pp. 53–87, 2004.

[11] Show-Jane Yen and Yue-Shi Lee, "Mining Sequential Patterns with Item Constraints", DaWaK 2004: data warehousing and knowledge discovery: International conference on data warehousing and knowledge discovery, Zaragoza, ESPAGNE, vol. 3181, pp. 381-390, 2004.

[12] Rigou, M., Sirmakessis, S., and Tsakalidis, A. K., "A Computational Geometry Approach to Web Personalization", In Proceedings of CEC, pp. 377-380, 2004.

[13] J. Pei et al., "Mining Sequential Patterns by Pattern-Growth: The Prefix Span Approach", IEEE Trans. Knowledge Data Eng., vol. 16, no. 11, pp. 1424–1440, Nov. 2004.

[14] P. Berkhin, "A Survey of Clustering Data Mining Techniques", in Grouping Multidimensional Data. Berlin, Germany: Springer-Verlag, 2006, pp. 25–71.

[15] Yen-Liang Chen, Ya-Han Hu, "The Consideration of Recency and Compactness in Sequential Pattern Mining", In Proceedings of the second workshop on Knowledge Economy and Electronic Commerce, Vol. 42, Iss. 2 , pp. 1203-1215, 2006.

[16] Jian Pei, Jiawei Han, Wei Wang, "Constraint-based Sequential Pattern Mining : The Pattern Growth Methods", J Intell Inf Syst, Vol. 28, No.2, pp. 133 - 160 , 2007.

[17] T.-P. Hong, C.-W. Lin, and Y.-L. Wu, "Incrementally Fast Updated Frequent Pattern Trees", Expert System Application, vol. 34, no. 4, pp. 2424–2435, May 2008.

[18] Krzysztof D., Wojciech K., Marcin S., "Effective Prediction of Web User Behaviour with User-Level Models", Fundamental Informatics, IOS Press , Vol. 89, No. 2-3, pp. 189, 2008.

[19] K. R. Suneetha, Dr. K. R. Krishnamoorthy, "Identifying User Behavior by Analyzing Web Server Access Log File", IJCSNS International Journal of Computer Science and Network Security, Vol. 9, No.4, pp. 327, 2009.

[20] T.-P. Hong, C.-W. Lin, and Y.-L. Wu, "Maintenance of Fast updated Frequent Pattern Trees for Record Deletion", Comput. Statist. Data Analysis, vol. 53, no. 7, pp. 2485–2499, May 2009.

[21] Dhirendra Kumar Jha, Anil Rajput, Manmohan Singh. & Archana Tomar, (2010) "An Efficient Model for Information Gain of Sequential Pattern from Web Logs based on Dynamic Weight Constraint", IEEE International Conference on Computer Information Systems and Industrial Management Applications, pp. 518-523.

[22] C.-W. Lin, T.-P. Hong, and W.-H. Lu, "An Effective Tree Structure for Mining High Utility Itemsets", Expert System Application, vol. 38, no. 6, pp. 7419–7424, Jun. 2011.

[23] C.-W. Lin and T.-P. Hong, "A New Mining Approach for Uncertain Databases using CUFP Trees", Expert System Application, vol. 39, no. 4, pp. 4084–4093, Mar. 2012

[24] Omar Zaarour, Mohamad Nagi, "Effective Web Log Mining and Online Navigational Pattern Prediction", ELSEVIER, 2013.

[25] Rahul Moriwal and Vijay Prakash, "An Efficient Algorithm for Finding Frequent Sequential Traversal Patterns from Web Logs based on Dynamic Weight Constraint", Proceedings of the Third International Conference on trends in Information, Telecommunication and Computing, Vol. 150, 2013.

[26] Jerry Chun, Wensheng Gan, Tzung Pei Hong, "Efficiently Maintaining the Fast Updated Sequential Pattern Trees With Sequence Deletion", IEEE Access - The Journal for Rapid open access publishing, Vol. 2, pp. 1374-1383, 2014.

[27] Sahu S., Saurabh P. and Rai S. ,An enhancement in clustering for sequential pattern mining through neural algorithm using Web logs Proceedings of International Conference on Computational Intelligence and Communication Networks 758-764 IEEE Press, 2014.

[28] Dmitriy Fradkin, Fabian Mörchen, "Mining Sequential Patterns for Classification", Vol. 45, Issue 3, pp 731- 749, December 2015.

[29] Wei She, "Role Based Integrated Access Control and Data Provenance for SOA Based Net Centric Systems", IEEE Transactions on Services Computing, Vol. 9, Issue 6, pp. 940-953, 2016.

[30] Doddegowda B. J., G. T. Raju, Sunil Kumar, "Extraction of Behavioral Patterns from Pre processed Web Usage Data for Web Personalization", IEEE International Conference On Recent Trends In Electronics Information Communication Technology, pp. 494-498, 2016.

[31] Minubhai Chaudhari, CHirag Mehta, "Extension of Prefix Span Approach with GRC Constraints for Sequential Pattern Mining", International Conference on Electrical, Electronics, and Optimization Techniques, pp. 2496-2498, 2016.

[32] Fan Muhan, Shao Sujie, Rui Lanlan, "A Mining Algorithm for Frequent Closed Pattern on Data Stream based on Sub Structure Compressed in Prefix Tree", IEEE Proceedings of CCIS, pp. 434-439, 2016.

[33] Bing Zhang, Guoyan Huang, Haitao He, Jiadong Ren, "Approach to Mine Influential Functions Based on Software Execution Sequence", International Journal of Engineering and Technology, Vol. 11, Issue 2, pp. 48-54, 2017.

[34] H. Ryang and U. Yun., "Efficient High Utility Pattern Mining for Establishing Manufacturing Plans with Sliding Window Control", IEEE - Expert Systems with Applications, Vol. 57, pp. 214-231, 2017.