ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 5, Issue 2, Februray 2016

Efficient Knowledge Representation Integrated with Web Usage Mining for Webpage Recommendation

Namita Ganjewar[1], Prof. Anupama Phakatkar[2]

P.G. Student, Department of Computer Engineering, Pune Institute of Computer Technology, Maharashtra, India[1]

Professor, Department of Computer Engineering, Pune Institute of Computer Technology, Maharashtra, India[2]

ABSTRACT:The explosive growth of information on WWW (World Wide Web) has overwhelmed web users by offering many choices for their searches. Consequently, web users tend to make poor decisions when surfing the web due to an inability to cope up with enormous amounts of information. There is need of recommender systems to help web users by providing useful and effective recommendations or suggestions. In particular, a recommender system can suggest interesting items from a large set of items based on the knowledge gained from input web usage log.The webpage recommendation is popularly used now a day in E-commerce websites to provide the user suggestions or recommendations based on what the user would like to see in the future for his next navigation on the web. In the existing systems where usage mining is used, the recommended pages are limited within discovered web access sequences learnt from web usage data. So domain knowledge generation has become important to increase efficiency of webpage recommendation results. The proposed system discovers frequent web access patterns from the input web log using CM-SPADE (Co-occurrence Map based Sequential PAttern Discovery) algorithm. To generate the domain knowledge, documents containing webpage title and snippet of webpage are clustered using TFIDF (Term Frequency – Inverse Document Frequency) values calculated for documents. The most relevant searches can be obtained by preferring the frequently accessed webpage contained within same cluster of the current web access of user. The proposed approach will enhance the recommendation accuracy and time to discover frequent web access sequences.

KEYWORDS: Web Usage Mining, Webpage Recommendation, Domain Knowledge, CM-SPADE (Co-occurrence Map based Sequential PAttern Discovery), FCM (Fuzzy CMeans Clustering).

I. INTRODUCTION

The explosive growth of information on the World Wide Web with the development of advanced electronic devices has made web information increasingly important in almost everybody's life. The rapid introduction of current websites has overwhelmed Web users by offering many choices. Consequently, web users tend to make poor decisions when surfing the web due to an inability to cope with enormous amounts of information. Recommender systems have proved in recent years to be a valuable means of helping web users by providing useful and effective recommendations or suggestions. The core techniques in recommender systems are the learning and prediction models which learn users' behavior and evaluate what users would like to view in the future. In particular, a recommender system can suggest interesting items from a large set of items based on the knowledge gained about an active user.

Webpage recommender systems are one kind of recommender systems, which can automatically recommend webpages that are most interesting to a particular user based on the user's current web navigation behaviour. Since a website is usually designed to show the index pages on the home page, the index pages take the role of guiding users to the content pages on the website through webpage links, whereas with the index pages, a user usually has to navigate a number of webpages to reach the content page they are interested in. If the index pages of a website are not well designed, which is often a case; web users will struggle to find useful pages and are very likely to leave the site. For a commercial website, this means losing potential customers. For an e-government website, this will mean that the

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 5, Issue 2, Februray 2016

citizen's needs are not satisfied. Therefore, webpage recommender systems have become increasingly valuable for helping web users to find the most interesting and useful webpages on specific websites. Good webpage recommendations can improve website usage and web user satisfaction.

Idea of this dissertation is to propose the enhancement over the approach of recommendation by integrating the domain and web usage knowledge of a website. In the existing systems, frequent sequential web access patterns are extracted using sequence mining techniques to the web usage data. With this approach, webpage recommendations are limited within frequent web access patterns. To increase the efficiency of recommendation results, we are proposing a way to extract frequent web access sequence using CM-SPADE algorithm and to construct the domain knowledge by finding out terms and clustering the similar terms. The combination of web usage knowledge extraction and domain knowledge representation gives the more refined and relevant webpage recommendations.

The goal of the system is to give the relevant webpage recommendations depending on the current clickstreams of the user. The system proposes clustering in domain knowledge representation model to improve the recommendation results.

The objectives of a system:

- To mine the frequent web access patterns from web usage log.
- To generate domain knowledge by using clustering of webpages.
- To integrate the domain knowledge with web usage knowledge to improve recommendations beyond the discovered web access sequences.
- To provide effective recommendations relevant to user's current webpage access.

Scope of our system can be elaborated as the webpage recommendation system which extracts the frequent web access sequences from web usage logs and clusters documents containing snippet information as per number of clusters given by user after checking similarity between documents. The use of clustering forms the domain knowledge of a website. By combining the web usage knowledge that is frequent web access patterns with the domain knowledge of a website, the final webpage recommendations are given to the end user.

II. RELATED WORK

Thi Thanh Sang Nguyen et al. proposed a novel method to efficiently provide better webpage recommendation through semantic enhancement by integrating the domain and web usage knowledge of a website. The paper represents domain knowledge using two new models. The first model uses ontology to represent the domain knowledge. The second model uses semantic network to represent domain terms, webpages and the relations between them.This paper focuses on the conceptual prediction model that automatically generates semantic network of web usage knowledge, which is integration of domain knowledge and web usage knowledge.The experimental results given in this paper demonstrate that the proposed method produces significantly higher performance than the web usage mining method. For this demonstration, the performance of webpage recommendation strategies is measured in terms of two performance metrics that is precision and satisfaction [1].

Liu J.N.K et al. proposed a new ontology learning model called DOG (Domain Ontology Graph) in this paper. There are two key components in the DOG that is the definition of the ontology graph and the ontology learning process.The Ontology graph is constructed by using the steps as follows. This is called as ontology and knowledge conceptualization model.

1. Term Extraction
2. Term to class relationship mapping
3. Term to term relationship mapping
4. Concept clustering
5. Ontology graph generation

Liu J.N.K. et al. focus on generating DOGs to represent the domain knowledge and conducting the text classifications based on the generated ontology graph. The experimental results presented in this paper show that the new method can produce significantly better classification accuracy [2].

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 5, Issue 2, Februray 2016

Xindong Wu et al. have presented the survey paper on top ten algorithms of data mining in classification domain. The paper elaborates on the focused study of different supervised and unsupervised classification techniques like k-Means, SVM (Support Vector Machine), Apriori, PageRank, AdaBoost, kNN (k Nearest Neighbor) and Naive Bayes. These top ten algorithms are among the most influential data mining algorithms in the research community. With each algorithm, the paper provides a description of the algorithm, impact of the algorithm, and review of current and further research on the algorithm [3].

Mohammed J. Zaki et al. presented a new algorithm SPADE (Sequential PAttern Discovery) for fast discovery of Sequential Patterns. The existing solutions of patterns mining make repeated database scans and use complex hash structures. This paper follows the approach of decomposing the original problem into smaller sub-problems, that can be independently solved in main memory using efficient lattice search techniques and using simple join operations. In this algorithm, all sequences are discovered in only three database scans [4].

K. Sathiyakumari et al. have presented the approach of ranking documents using TF-IDF (Term Frequency – Inverse Document Frequency) technique. Inclusion of fuzzy logic in clustering results in improved performance in the cluster and its contents. Experimental results in this paper show that the results of FCM algorithm are modified using ranking based approach before clustering is performed and this new improved clustering algorithm is called as MFCM (Modified Fuzzy C Means) which can be efficiently used for document clustering [5].

C.I. Ezeife et al. have proposed mining web log sequential patterns with PLWAP (Position Coded Pre-order Linked Web Access Pattern) tree. In this paper, frequent m-sequences are computed and discovered using frequent (m-1) sequences and appropriate suffix sub-trees. The position codes are assigned at each level in the hierarchy from top to bottom. One level down in the prefix coded tree gives the subsequent access to the parent node in the tree. From the search space, complete set of frequent patterns are efficiently discovered by using traversal of PLWAP tree that is formed from input web usage log and positions codes [6].

III. WEBPAGE RECOMMENDATION

There are mainly three modules in the proposed system.

1. Web usage mining to extract frequent web access sequences using CM-SPADE algorithm

In the dataset we are taking usage logs of multiple users showing navigation from one webpage to another webpage. The web logs taken from web server of the website for specific period of time are used to analyze frequent web access sequences. One log file is considered as one sequence and frequently visited webpages are extracted using CM SPADE algorithm.

2. Domain Knowledge Construction

By using TFIDF i.e. Term Frequency Inverse Document Frequency, we can calculate frequency of each term and occurrences within no. of documents. Based on similarity checking, we perform clustering into different domains so that recommendation results can be improved more.

3. Recommendation Engine

In this module, it takes frequent web access patterns from web usage mining process and term, their required information as well as clustering results from domain knowledge representation. Combining both of these and analyzing current web access, the recommendation engine gives the set of recommended web pages to the user as per his interest.

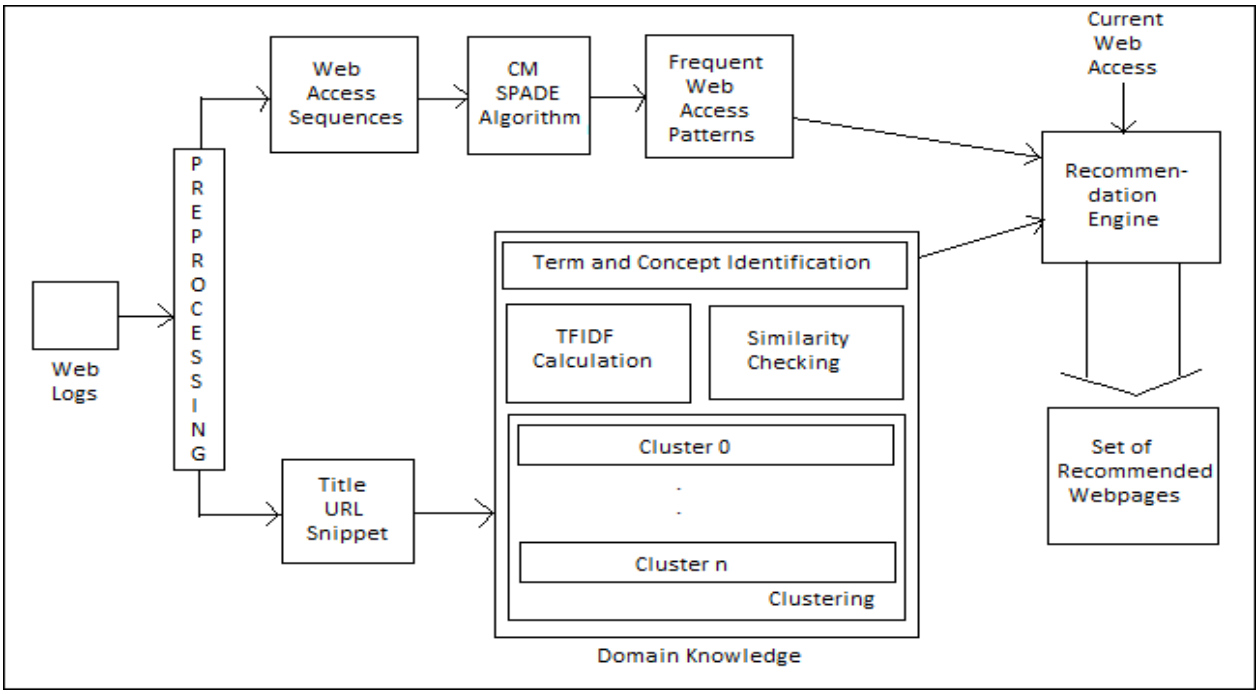
Following is the architectural diagram of the webpage recommendation system.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 5, Issue 2, Februray 2016



IV. ALGORITHMIC SOLUTION

Algorithm: To recommend the webpages relevant to user's current access.

Input: Web usage log files and documents containing URL, title, snippet of webpage.
Output: A set of recommended webpages.

Steps:

1. Run the webpage recommender system.
2. Select one input log file as sample file.
3. Show the sample input log.
4. Get the number sequence from the usage log in the form of itemset and sequence where (-1) indicates lexicographical order break and (-2) indicates end of sequence.
5. Discover the frequent web access patterns using CM-SPADE algorithm.
   {
   A. Calculate support count of all items and the possible combinations of items represented as itemsets.
   B. Represent i-extension (Itemset Extension) and s-extension (Sequence Extension) of all items using co-occurrence map structure.
   C. Prune the candidates below minimum support.
   }
6. Recommend a set of webpages that are frequently accessed webpages after some particular webpage based on results of CM-SPADE algorithm.
7. Calculate TFIDF of the documents.
8. Cluster the documents when TFIDF output is provided to clustering module based on user selected similarity measure for clustering of documents.
9. Generate final set of recommendation webpages by combining the pattern mining results and domain knowledge of the webpages.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

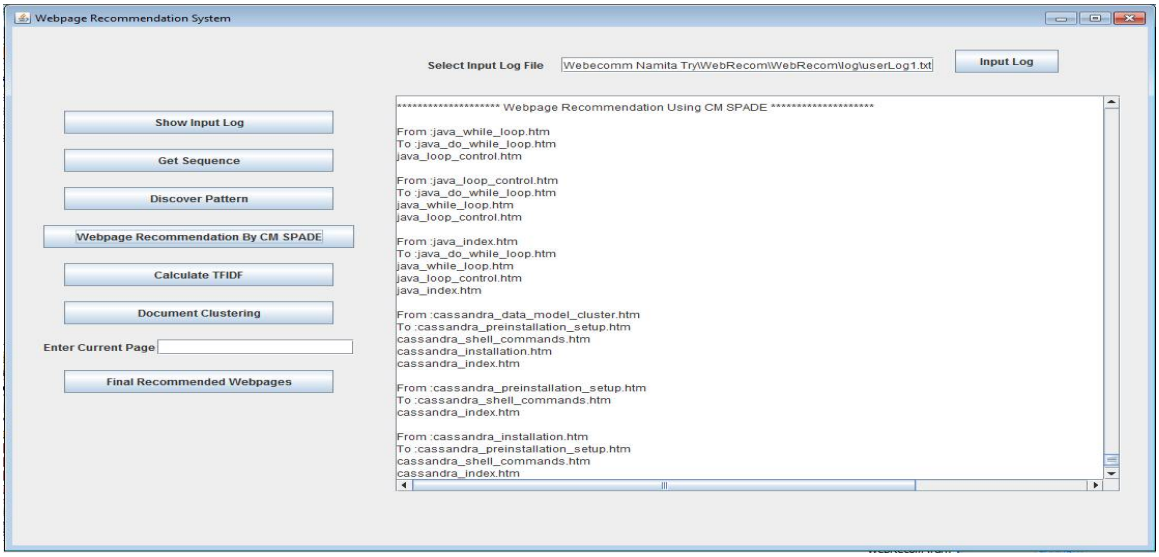
(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 5, Issue 2, Februray 2016

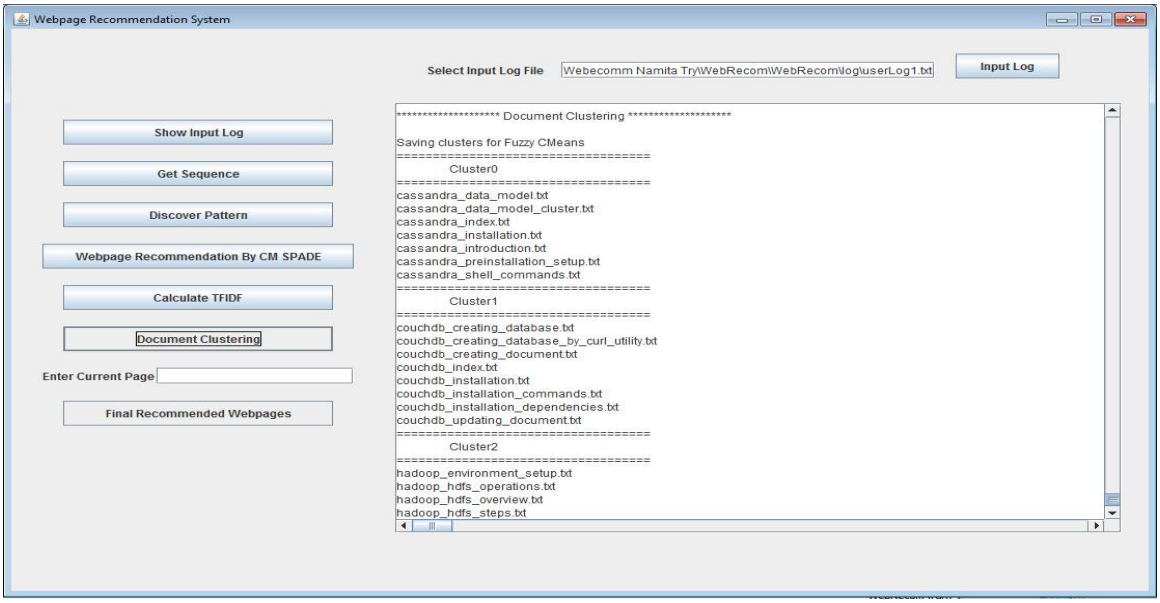
10. Show the results that are the suggested webpages that can be accessed as next navigation step from the current webpage of user.

V. EXPERIMENTAL RESULTS

- Webpage Recommendation Using CM-SPADE



- Clustering Results



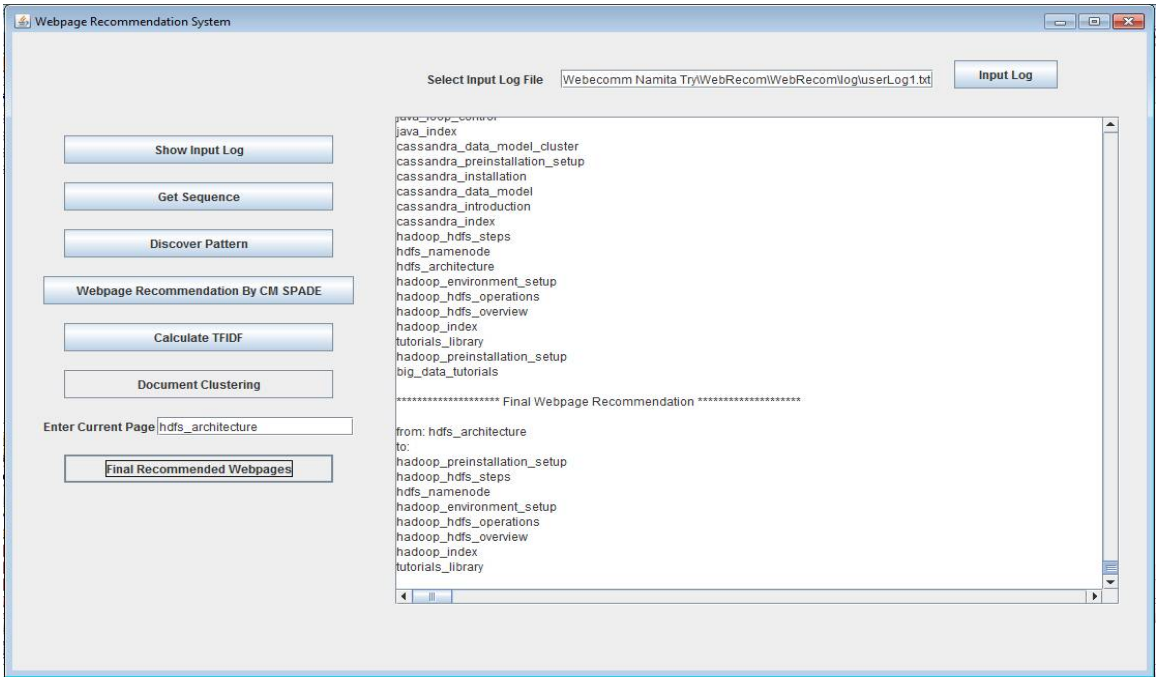
ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 5, Issue 2, Februray 2016

- Webpage Recommendation Results



VI. CONCLUSION

With the substantial growth of information on the internet, webpage recommendation has become crucial part of the websites today. In the existing systems that incorporate web usage mining, the recommendation results are limited within the frequent web access patterns generated from input web log for example, e-commerce websites suggest the user with product relevant to web usage history of that user. To enhance the performance of webpage recommendation beyond frequent web access sequence, the proposed system introduces clustering of documents containing webpage information which generates the domain knowledge. The use of domain knowledge along with web usage knowledge generated by discovering frequent web access sequences using CM-SPADE algorithm, provides semantic enhancement in the existing webpage recommendation system. The most relevant searches can be obtained by preferring the frequently accessed webpage contained within same cluster of the current web access of the user. With this approach, the accuracy of recommended webpages can be improved. The experimental results show that the proposed system has increased the performance of webpage recommendation in terms of time required to get recommendation results.

REFERENCES

[1] Thi Thanh Sang Nguyen, Hai Yan Lu, Jie Lu, "Web Page Recommendation Based on Web Usage and Domain Knowledge," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 10, pp. 2574-2587, October 2014.
[2] James N.K. Liu, Yu Lin He, Edward H.Y. Lim, Xi Zhao Wang, "A New Method for Knowledge and Information Management Domain Ontology Graph Model," *IEEE Transactions on Systems, MAN, and Cybernetics Systems*, vol. 43, no. 1, pp. 293-307, January 2013.
[3] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, "Top 10 algorithms in data mining," *Springer Knowledge and Information Systems*, vol. 14, no.1, pp. 1-37, 2008.
[4] Mohammed Zaki, "SPADE - Efficient Algorithm for Mining Frequent Sequences," *ACM Journal of Machine Learning*, vol. 42, no. 2, pp. 31-60, February 2002.
[5] K.Sathiyakumari, V.Preamsudha, "Unsupervised Approach for Document Clustering
Using Modified Fuzzy C Mean Algorithm," *International Journal of Computer and Organization Trends*, vol. 1, no.3, pp. 10-16, January 2013.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 5, Issue 2, Februray 2016

[6] C. Ezeife, Y. Liu, "Fast Incremental Mining of Web Sequential Patterns with PLWAP Tree," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 721–732, April 2012.
[7] Olfa Nasraoui, Maha Soliman, Esin Saka, Antonio Badia, Richard Germain, "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 2, pp. 911-923, February 2012.
[8] Amal Zouaq, Roger Nkambou, "Evaluating the Generation of Domain Ontologies in Knowledge Puzzle Project," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 11, pp. 1247-1259, November 2009.
[9] Lizhi Liu, Jun Liu, "Mining Web Log Sequential Patterns with Layer Coded Breadth-First Linked WAP Tree," *International Conference of Information Science and Management Engineering*, vol. 19, no. 5, pp. 124-139, May 2010.
[10] N.R. Mabroukeh, C.I. Ezeife, "Semantic-Rich Markov Models for Web Prefetching," *ACM International Conference on Web Search and Data Mining*, vol. 34, no. 1, pp. 11-29, March 2008.
[11] C.I. Ezeife, Y. Lu, "Mining Web Log Sequential Patterns with Position Coded Pre-Order Linked WAP-Tree," *Proceedings of 30th Annual International ACM Conference on Research and Development in Information Retrieval*, vol. 16, no. 2, pp. 72-81, October 2011.
[12] T.T.S. Nguyen, H. Lu, T.P. Tran, J. Lu, "Investigation of Sequential Pattern Mining Techniques for Web Recommendation," *International Journal of Scientific and Engineering Research*, vol. 4, no. 4, pp. 293–312, April 2012.
[13] B. Zhou, S.C. Hui, A.C.M. Fong, "CS-Mine: An efficient WAP-Tree Mining for Web Access Patterns," *Proceedings of Advanced Web Technologies and Applications*, vol. 9, no. 7, pp. 523–532, November 2010.