International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 16 (2017) pp. 5648-5654
© Research India Publications. http://www.ripublication.com

An Empirical Analysis of Algorithms to Predict Next Web Page Using Web Log Data

Jothish Chembath[1] and Dr. E.J. Thomson Fredrik[2]

[1]Research Scholar, Computer Science Department, Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India.

Orcid Id: 0000-0002-1070-9027

[2]Associate Professor, Computer Applications Department, Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India.

Abstract

Prediction of next web page that a user may visit is an area of research that is very active as the demand for increased accuracy is high in the ever changing World Wide Web. Several models are proposed for this purpose, which has great use in many applications including social networking, e-commerce and knowledge management systems. Various web usage mining algorithms focus on identifying the user's next web page visit. Most of these techniques are based on data mining techniques like association rule mining, Markov Modeling and clustering. Another technique that gains equal popularity is the usage of pattern search algorithms, like Longest Common Subsequence, for finding navigation patterns by analyzing the current user browsing activities for predicting future requests. This paper presents three hybrid models and compares their performance on prediction. The first model combines clustering, Markov model and Apriori All rule mining algorithm. The second model combines clustering, Markov model and FP-Growth rule mining algorithm, while the third model combines clustering and longest common subsequence method. Experimental evaluation reveals that Markov-based models produce more accurate results but longest common subsequence-based model is much faster.

Keywords: Association Mining, Apriori All, FP-Growth, Longest Common Subsequence, Markov Model, Next Page Prediction, Web Log Data, Web Usage Mining

INTRODUCTION

Modern information technology along with the use of sophisticated devices make it possible to create and transfer huge amounts of data in a geographically independent and cost effective manner. However, this huge amount of information is exploited to gain knowledge is a challenging task. This problem is more severe in World Wide Web (WWW), in which the main goal is to store, publish and deliver information in an easy and fast manner. This information explosion is increasing to make it difficult for people to locate and access online information. To some extent search engines like Google (http://www.google.com) have reduced this difficulty of locating the right information at the right time. Apart from the search engines, many researchers have attempted to provide tools which aid in improving the browsing experience of the user (Nylen and Homstrom,[1] ; Zhou and Jiao[2]). The heart of several of these tools is web mining, a part of data mining, web mining focuses on discovering knowledge from huge databases related to WWW. Two main goals of web mining techniques are to improve user efficiency and effectiveness in searching for information on the web and support businesses in decision-making or business management. Web mining can be classified into three main categories, namely web content mining, web usage mining and web structure mining (Chandel *et al.*[3]). This paper is focused on web usage mining techniques, to predict web pages that can be requested in future using the historical usage data in order to improve the browsing experience of the users. The Next Web Page Prediction (NWPP) systems have high demand in several online applications including e-commerce and e-learning (Narvekar and Banu[4]). The results of these prediction systems can be used for various purposes including personalization of web, reduction of the server response time with proper perfecting and caching strategies, providing guidelines for improving the design of web applications, handling business for specific issues like customer attraction, customer retention, cross sales and customer departure in e-commerce applications. A variety of algorithms are proposed for this purpose, out of which, usage of web access logs to generate navigational profiles for recommending future requests is a popularly adopted method. However, due to the high diversity of on-line navigation and frequently changing web characteristics, building such a prediction system is challenging and an active area of research focuses on identifying techniques that improve these systems in terms of accuracy and speed. The general steps involved in NWPP system are preprocessing, pattern mining and pattern analysis. This paper analyses the different algorithms that can predict future next web requests of the user. The three algorithms selected for empirical analysis are hybrid algorithms that combine (i) clustering, Markov Model and Association mining using Apriori All algorithm, (ii) clustering, Markov Model and Association mining using FP-Growth algorithm and (iii) clustering and Longest Common Sequence (LCS). All these models use the history of web page visits collected from web log data. These are respectively referred to as CMA-NWPP, CMF-NWPP and CL-NWPP models in this paper. The rest of the paper is organized as follows. Section 2 presents detailed description of the three models selected. Section 3 discusses the experimental results obtained during the performance evaluation and Section 4 concludes the work with future research directions.

RELATED WORK

Prediction of web pages and the user's behavior are related to each other. It is needed to asess the browsing history of the user also should know the browsers intention. The Internet is immensely large, makes prediction so complex as the web pages are inter twined with each other and the types of data which are included in web pages are heterogeneous. These facts make it even more difficult to predict, what move will the user make next. Several people tried different techniques for predicting the user's next move, thereby guiding the user with the next pages that the user might possibly browse upon. The researchers has incorporated few review of literature related to this present study. They are as follows.

1. Meera Narvekar, Shaikh sakina banu **(year**) , studied about the Predicting User's Web Navigation Behavior Using Hybrid Approach proposes a Hybrid model which combines Markov model as well as Hidden Markov Model which gives user the list of web pages of their interest. They used various kinds of datasets to analyze, compare and show the effectiveness of Hybrid model using various parameters such as Accuracy, Precision and Miss-Prediction.
2. Mukund Deshpande, George karypis **(year)** Selective Markov Models for Predicting Web Page Accesses University of Minnesota has proposed techniques for intelligently selecting parts of different order Markov models so that the resulting model has a reduced state complexity and improved prediction accuracy. They had tested their models on various datasets and have found that their performance is consistently superior to that obtained by higher-order Markov models
3. Tyagi, N,K. and Solanki, A.K. (2011) Prediction of users behavior through correlation rules, International Journal of Advanced Computer Science and Application recommends interesting Web pages to the users on the basis of their behavior discovered from web log data. Association rules are generated using FP growth approach and has used two criteria for selecting interesting rules: Confidence and Cosine measure. They have also proposed an algorithm for the recommendation process.
4. Khalil, F., Li, J. and Wang, H. (2008) Integrating recommendation models for improved web page prediction accuracy,Thirty-First Australasian Computer Science Conference (ACSC2008), Conferences in Research and Practice in Information Technology Proposes to provide an improved Web page prediction accuracy by using a novel approach that involves integrating clustering, association rules and Markov models according to some constraints. They have proved that the Experimental results prove that this integration provides better prediction accuracy than using each technique individually.

NEXT WEB PAGE PREDICTION MODELS

The selected models perform next page prediction in five steps. They are, preprocessing, feature vector formation, clustering, building Markov model and prediction. In the previous work (Jothish Chembath and S.K.Mahendran[5]), proposed a preprocessing algorithm that improved transaction identification process through the use of pruning algorithm (to remove irrelevant users), automatic procedure (to identify the cutoff time for both session and transaction identification) and improved path completion algorithm to obtain a complete web log data(Jothish Chembath and E,J.Thomson Fredrik[6]) reported algorithms related to feature vector formation and clustering. This algorithm created a feature vector from the preprocessed web log data. The feature vector was created as a representation of number of times a user visited a web page. Using this feature vector, the web pages in a session was grouped (or clustered) according to the service requested. During clustering, an improved version of ensemble K-Means clustering algorithm was used. This algorithm was named as Ensemble Parameterless Fast K-Means (EPFK-Means) clustering algorithm. Experimental results proved that the inclusion of these algorithms in NWPP reduced complexity, improved the property of scalability and reliability of the prediction model. As mentioned previously, this paper analyze three algorithms for the final step, that is, prediction of user's next access.

CMA-NWPP AND CMF-NWPP MODELS

The CMA and CMF versions of NWPP are designed as hybrid models that combine three algorithms, namely, clustering (performed using EPFK-Means algorithm), Markov model

(Deshpande and Karypis[7]) and association rule mining (Tyagi and Solanki[8]). The main motivation of using this hybrid model is to reduce the number of transactions used, thus reducing the time complexity of the prediction model. The number of transactions is reduced through the use of the EPFK-Means clustering algorithm, sessions of groups into different categories based on the services. Then Markov model prediction is performed one with each cluster. Association rules are used when markov models cannot make decision and long historical information is required. The steps involved in CMA-NWPP and CMF-NWPP models are the same except for the algorithm used for association rule mining. The CMA-NWPP model uses Apriori All (Agrawal and Srikant,[11]) algorithm, while the CMF-NWPP model uses FP-Growth (Han, [9]) algorithm. During the prediction, both the prediciton models begin by performing Markov model analysis on each cluster obtained by applying EPFK-Means clustering algorithm on user sessions. Let P = {p_1, ..., p_m) be the set of pages in a website and S be a user session has the series of web pages visited by the user. If the user has visited 'r' pages, then the probability that a user visits a page p_i is denoted as prob(p_i|S). The probability of page p_{r+1} can thus be estimated using Equation (1).

$$P_{r+1} = \text{argmax}_{p\in P}\{P(P_{r+1} = p|S)\} = \text{argmax}_{p\in P}\{P(p_{r+1} = p|p_r, p_{r-1}, ..., p_1)\} \quad (1)$$

From Equation (1), it can be seen that the probability is estimated by using sequences of all users in history (or training data) denoted as S. The NWPP is more accurate with larger 'r', which will produce large S. However, large 'r' also results in high complexity. This complexity can be reduced, if it is assumed that all visited pages follows a Markov process, it may impose a limit on the number of pages visited to a constant value k (where K<<r). Thus, while using Markov model, the probability of visiting a page pr depends only on a small set of k preceding pages and not on all the pages in the session. Using this assumption, Equation (1) can be rewritten as Equation (2).

$$P_{r+1} = \text{argmax}_{p\in P}\{P(p_{r+1} = p|p_r, p_{r-1}, ..., p_{r-(k-1)})\} \quad (2)$$

Here, k is the number of previously visited pages and it is used to identify the order of Markov model. The model using Equation (2) is referred as kth Order Markov Model (kOMM). Thus, in order to use the kOMM model, the learning of pr+1 is needed for each sequence of k web pages. Let S_j^k be a state with k number of preceding pages denote the Markov model order and j be the number of unique pages on the website. The probability of P(p_i|S_j^k) is estimated using Equation (3) which is obtained from the historical or training dataset.

$$P(p_i \mid S_j^k) = \frac{\text{frequency}(< S_j^k, p_i >)}{\text{frequency}(S_j^k)} \quad (3)$$

The above equation calculates the conditional probability as the ratio of number of times a sequence S_j^k occurs in the training set to the number of times the page p_i occurs immediately after S_j^k. All pages that satisfy the condition probability are selected as pages which may be visited by the user. In order to improve the coverage and reduce the complexity of the model, three modified Markov models are used. They are, All kth Markov Model, frequency pruned Markov model and accuracy pruned Markov model. The main aim of NWPP is to determine a Markov model that leads to high accuracy with low state space complexity. According to Khalil et al. [10], state space complexity can be reduced through the use of frequency pruned Markov model. Careful analysis revealed that a 2-FP order Markov model can improve prediction accuracy while maintaining the state space complexity involved with higher order Markov models. The complexity is reduced due to the fact that the Markov model prediction is performed only on particular clusters as opposed to the whole web log dataset. When the state prediction probability is not marginal, then the CMA-NWPP and CMF-NWPP models use association mining to predict next web page. The usage of association rules helps to improve the accuracy of prediction as association rules look at more history and examine more states than Markov models. When the association rules are examined only in special cases, the issue related to high number of association rules generated does not increase the complexity of the model. The association rules are built based on window size 4, 90% confidence threshold and 4% minimum support. As mentioned earlier, two association mining algorithms, namely, Apriori All (Agrawal and Srikant[11]) and FP-Growth (Han[9]), are analyzed in this paper. The steps involved are summarized in Figure 1.

- **Training Data**
  - Perform preprocessing and cluster pages in user sessions into K-Clusters
  - For each cluster build l-Markov model
    - Prune results in each cluster using frequency pruned model requirements
  - For each Markov model state majority is not clear, repeat the following steps
    - Collect all sessions satisfying the state
    - Construct the association rules using Apriori All (or) FP-Growth algorithm
    - Store association rules with stages
- **Test Data (Prediction)**
  - For each coming session
    - Find its closest cluster and select its corresponding markov model to perform prediction
    - If predictions fail then use the association rules to predict next page
    - Output next page is predicted

Figure 1: Steps in CMA-NWPP and CMF-NWPP Models

CL-NWPP MODEL

The CL-NWPP model employs a pattern search algorithm for finding navigation patterns by analyzing the current user browsing activities for predicting future requests. For this purpose, the LCS method (Jalali et al.[10]) is used, which analyzes navigational patterns that contain the largest number of similar web pages in each session. The longest common subsequence (LCS) aims to find the longest subsequence common to all sequences in a set of sequences. A fixed-size sliding window over current active session is used to capture the current user activities. In order to classify user session windows, a cluster that includes the larger number of pages in that session is determined.

LCS has a well-studied optimal sub-structure property as given by the following (Jalali et al.[13]). The fundament task in pattern search and matching algorithm is the problem of comparing two sequences SEQ_a and SEQ_b, in order to determine the similarity between them. The LCS algorithm consists of comparison metrics that can measure the subsequence of maximal length that is common to both SEQ_a and SEQ_b (Apostolico, 1997). Let SEQ = $\{seq_1, ..., seq_n\}$ be a sequence of page visits of a user and SS = $\{ss_1, ... ss_n\}$ be a subsequence of SEQ, if there exists a strictly increasing sequence $\{j_1, ..., j_n\}$ of indices of SEQ such that for all i = 1, 2, .., m, $SEQ_{ji} = SS_i$. Given two sequences SEQ_a and SEQ_b, SS is said to be a common subsequence if SS is a subsequence of both SEQ_a and SEQ_b. LCS is concerned with finding the maximum-length or longest common subsequences given two sequences of pages. The LCS can be formulated using Equation (4).

$$\mathrm{LCS}(\mathrm{SEQ}_a, \mathrm{SEQ}_b) = \begin{cases} 0 & \mathit{if}\ \mathrm{i} = 0 \text{ or } \mathrm{j} = 0 \\ (\mathrm{LCS}(\mathrm{SEQ}_{\mathrm{ai\text{-}1}}, SEQ_{bj-1}), SEQ_{ai}) & \mathit{if}\ SEQ_{ai} = SEQ_{bi} \\ \mathrm{longest}(\mathrm{LCS}(\mathrm{SEQ}_{\mathrm{ai}}, SEQ_{bj-1}), LCS(SEQ_{ai-1}, SEQ_{bj})) & \mathit{if}\ SEQ_{ai} \neq SEQ_{bi} \end{cases} \tag{4}$$

To find the longest subsequences common to SEQ_{ai} and SEQ_{bj}, the elements of both the sequences are compared. If equal, then the sequence LCS($SEQa_{i-1}$, SEQ_{bj-1}) is extended by that element, SEQ_{ai}. If they are not equal, then the longer of the two sequences, LCS(SEQ_{ai}, SEQ_{bj-1}), and LCS(SEQ_{ai-1}, SEQ_{bj}), are retained. If they are at the same length, but are not identical, then both are retained.

A worked out example of LCS is given in http://en.wikipedia.org/wiki/Longest_common_subsequence_problem. The following paragraphs explain the usage of LCS in next web page prediction. The application of clustering results in a set of clusters C = <c_1, c_2, …c_n> where c_i = <p_1, p_2, …, p_k> where k is the set of web pages identified as user navigation patterns and 1 ≤ i ≤ n. Let sequence SEQ' = <p_1, …, p_m> be the current active session with m active session windows. Before classifying an active session to construct the prediction list, the pages in active session windows is sorted based on values stored in the co-occurrence matrix M. The CL-NWPP model, finds a cluster c_i with highest degree of similarity using LCS algorithm. This cluster, c_i, thus identified, is treated as the navigation pattern has set of pages, with respect to SEQ', that can be recommended to the user. The NWPP model then ranks the pages in c_i according to the degree of connectivity between pages in the form of adjacency matrix. The page with highest degree of connectivity is reported as the predicted page that the user may next wish to visit. When the prediction engine finds more than one cluster based on LCS algorithm, then the prediction engine selects a cluster in such a way that, if the difference between positions of last elements of longest common subsequence founded in the cluster and the position of first element of this sequence is minimized, then the system chooses this cluster. If the first page in the next user activity is different with prediction list, it needs again to classify with new user activities. The steps involved are presented in Figure 2.

```
Input : User's active session widows, SEQ';

Set of navigation pattern sequences (clusters), C

Procedure
        sort(SEQ');
        for each $c_i$ in C do
                sort($c_i$); //sort navigation pattern
                sequence
        end
        for each $c_i$ in C do
                //find LCS between navigation
                pattern sequence and active session
                Recommend Page Set = |LCS($c_i$,SEQ'|;
        end for
//(Find navigation pattern sequence with maximum LCS)
        Match_Str = NP_str(Max(RecommendPageSet);
        //recommendation page set is found based on the
                difference between active session and
                identified navigation pattern sequence
        RecommendedPageSet = Rank(SEQ'-
Match_Str);
Return (RecommendedPageSet(1));
```

Figure 2: Steps in CL-NWPP

EXPERIMENTAL RESULTS

In order to evaluate the performance of the preprocessing algorithms, four web log datasets were used (Table 1). The CMA-NWPP and CMF-NWPP algorithms generate association rules by using window size 4, 90% confidence threshold and 4% minimum support. The experiments are designed to evaluate the performance of the prediction algorithms with respect of two performance metrics, namely, accuracy, coverage, F1-Measure and speed are used for this purpose. Figure 3 shows the efficiency of the prediction models with respect to accuracy performance measure. From the accuracy results, it can be seen that the performance of the three selected models are higher that is conventional counterparts. Maximum performance is shown by the model that combined clustering, FP-growth algorithm and Markov model. This model improved the precision process by 18.05%, 13.71% and 7.64% respectively over the conventional LCS, Markov and association rule based prediction models. The CMF-NWPP also showed an average efficiency gain of 0.95% and 2.0% when compared with CMA-NWPP and CL-NWPP respectively. A similar trend was envisaged with coverage (Figure 4) and F1-Measure (Figure 5) performance measures also. But this trend changed while analyzing the algorithms with executing speed (Figure 6). While considering the speed of prediction, even though the three selected hybrid models are faster than the conventional models, the CL-NWPP model was faster than the CMA-NWPP and CMF-NWPP models. Thus, from the results it is clear that when the usage of hybrid model based on Markov model produces accurate prediction, it is slow, when the hybrid model combining clustering and LCS algorithm is fast.

Table 1. DATA SETS USED

Data Set	*Code*	*Period*	*Size(MB)*	*No of records*
NASA Kennedy Center Space (http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html)	NASA	01-07-95 to 31-08-95	205.2	34,61,612
University of Saskatchewan's (http://ita.ee.lbl.gov/html/contrib/Sask-HTTP.html)	SASK	01-06-95 to 31-12-95	233.4	24,08,625
ClarkNet Internet Service Provider (http://ita.ee.lbl.gov/html/contrib/ClarkNet-HTTP.html)	CN	24-08-95 to 10-09-95	171	33,28,587
University of Calgary's, Department of Computer Science (http://ita.ee.lbl.gov/html/contrib/ Calgary-HTTP.html	CL	24-10-94 to 11-10-94	52.3	7,26,739

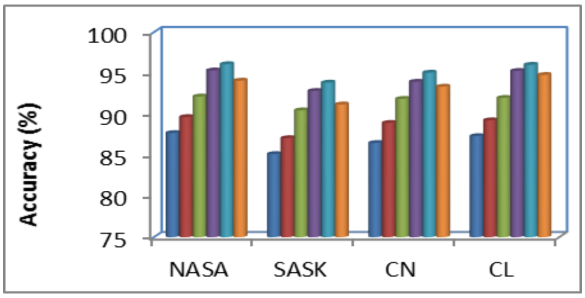


Figure 3: Accuracy (%)

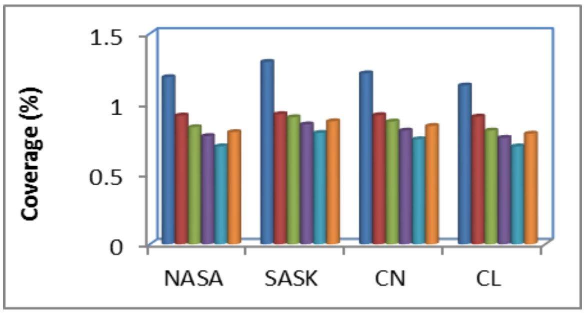


Figure 4: Coverage (%)

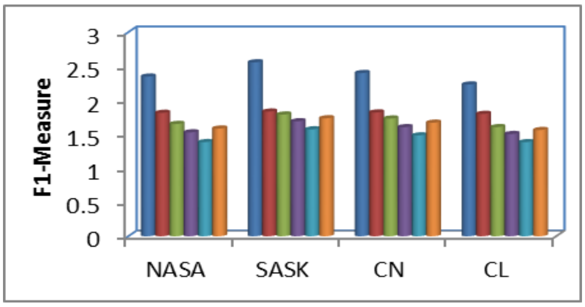


Figure 5 : F1-Measure

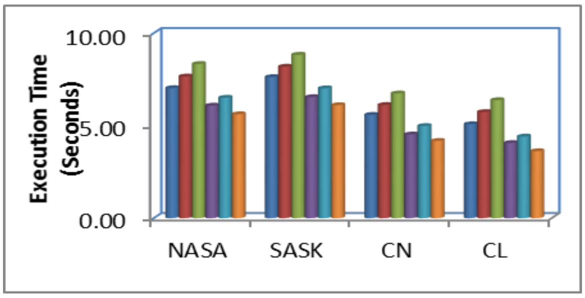
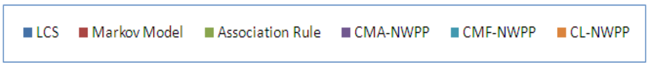


Figure 6 : Execution Speed (Seconds)



CONCLUSION

In WWW, various types of users visit a website, each with different browsing practices. This would require different unique models to simulate their behaviours and construction of separate models for each user is very inefficient and it is almost impossible due to the huge number of users. To solve this issue, the selected web page prediction models for predicting a user's next access used a combination of clustering, Markov model, association rules and LCS algorithms. Performance evaluation of the algorithms showed that all the three models are improved version to the conventional algorithms. The Markov models combined with clustering and association algorithms are more efficient in terms of accuracy when compared the model combining with LCS algorithm. However, the LCS-based model was much faster when compared to the other models. Future research ideas include the methods to identify these three models, with the aim of further improving the accuracy of web page prediction.

REFERENCES

[1] Nylen, D. and Homstrom, J. (2015), Digital innovation strategy: A framework for diagnosing and improving digital product and service innovation, Business Horizons (ScienceDirect), Vol. 58, Issue 1,

Pp. 57–67.

[2] Zhou, F. and Jiao, R.J. (2013), An improved user experience model with cumulative prospect theory, Procedia Computer Science, Conference on Systems Engineering Research, Vol. 16, Pp. 870-877.

[3] Chandel, G.S., Patidar, K. and Mali, M.S. (2016), A result evolution approach for web usage mining using fuzzy C-means clustering algorithm, International Journal of Computer Science and Network Security, Vol. 16, No. 1, Pp. 135-140.

[4] Narvekar, M. and Banu, S.S. , P(2015), Predicting User's Web Navigation Behavior Using Hybrid Approach, International Conference on Advanced Computing Technologies and Applications, Procedia Computer Science, Elsevier, Vol. 45, Pp. 3-12.

[5] Jothish Chembath and S.K.Mahendran(2015), Transaction Identification Algorithm enhanced with user pruning and combined maximal forward reference and reference length approach for improving prediction of next web page from web log entries, International Journal of Multidisciplinary research and development, vol 2, issue 6, Pp 426 - 430

[6] Jothish Chembath and E.J.Thomson Fredrik(2016), Ensemble Clustering Based Approach for Predicting Next Web Page Using Web Log Data – A study, European journal of scientific research, Volume 140, issue 4

[7] Deshpande, M. and Karypis, G. (2004), Selective markov models for predicting web page accesses, ACM Transactions on Internet Technology, Vol. 4, No. 2, Pp.163-184.

[8] Tyagi, N,K. and Solanki, A.K. (2011), Prediction of users behavior through correlation rules, International Journal of Advanced Computer Science and Application, Vol.2, No. 9, Pp.77-81.

[9] Han (2000) ,Mining Frequent Patterns Without Candidate Generation, Proceedings of the 2000 ACM SIGMOD, International Conference on Management of Data, SIGMOD '00, Pp. 1–12

[10] Khalil, F., Li, J. and Wang, H. (2008), Integrating recommendation models for improved web page prediction accuracy,Thirty-First Australasian Computer Science Conference (ACSC2008), Conferences in Research and Practice in Information Technology (CRPIT), Vol. 74, Pp. 91-100.

[11] Agrawal, R. and Srikant, R. (1996), Mining sequential patterns: Generalizations and Performance Improvements, Fifth International Conference on Extending Database Technology: Advances in Database Technology, Springer-Verlag, London, Pp. 3-17.

[12] Jalali, M., Mustapha, M., Mamat, A. and Sulaiman, M.N.B. (2008), A new clustering approach based on graph partitioning for navigation patterns mining, 9th International Conference on Pattern Recognition, Pp. 1-4.

[13] Jalali, M., Mustapha, N., Sulaiman, N. and Mamat, A. (2010), WebPUM: A web-based recommendation system to predict user future movements, Expert systems with applications, Vol. 37, Pp. 6201-6212.