Web Page Recommendation Using Domain Knowledge and Improved Frequent Sequential Pattern Mining Algorithm

Ms.Harshali H. Bendale
Master of Engineering (Computer) Department ofComputer Engineering
JayawantraoSawant College of Engineering,Hadapsar
SavitribaiPhule Pune University, Pune, Maharashtra, India
e-mail:bendaleharshali@gmail.com

Prof. H. A. Hingoliwala
Head of Department and Asso. Prof (Computer)Department of Computer Engineering
JayawantraoSawant College of Engineering, Hadapsar
SavitribaiPhule Pune University, Pune, Maharashtra, India
e-mail:ali_hyderi@yahoo.com

Abstract—Web page recommendation is the technique of web site customization to fulfil the needs of every particular user or group of users. The web has become largest world of knowledge. So it is more crucial task of the webmasters to manage the contents of the particular websites to gather the requirements of the web users. The web page recommendation systems most part based on the exploitation of the patterns of the site's visitors. Domain ontology's provide shared and regular understanding of a particular domain. Existing system uses pre-order linked WAP-tree mining (PLWAP Mine) algorithm that helps web recommendation system to recommend the interested pages but it has some drawbacks, it require more execution time and memory. To overcome the drawbacks of existing system paper utilizes PREWAP algorithm. The PREWAP algorithm recommends the interested results to web user within less time and with less memory and improves the efficiency of web page recommendation system. In work, various models are presented; the first model is Web Usage Mining which uses the web logs. The second model also utilizes web logs to represent the domain knowledge, here the domain ontology is used to solve the new page problem. Likewise the prediction model, which is a network of domain terms, which is based on the frequently viewed web-pages and represents the integrated web usage. The recommendation results have been successfully verified based on the results which are acquired from a proposed and existing web usage mining (WUM) technique.

Keywords*- Domain ontology, knowledge representation, sequential patterns, web page recommendation, web mining.*

*****

I. Introduction

Web page recommendation system is most important term in today's world. It recommends the web pages to web user according to his search term. These are some areas where recommendation system [1] used some application like music, movie, books, news, restaurants etc. As World Wide Web grows tremendously the size and complexity of many web sites increased rapidly so user faces the problem of directing the web page in their area of interest, and it is difficult and time consuming to find the information they are looking for. Users don't know initially what they required and their requirements may change that will lead to change in selection. The fast presentation of current websites has overwhelmed Web users by offering numerous choices. Consequently, Web users tend to make poor decisions when surfing the Web due to a failure to cope with enormous measures of data. Recommender systems have proved in current years to be a valuable means of helping Web users by giving useful and effective recommendations.

Recommender systems core method is the learning and prediction models which learn user's behavior and find what users might want to view in the future. Specifically, **Recommendation system** offers required results or information by studying the behavior of user and area of interests. **Recommendation systems** uses collaborative and content based filtering approaches for providing what they want by learning past behavior of user. Since a website is normally designed to present the index pages on the home page, the index pages plays important role of directing users to the recent pages on the website through Web-page joins whereas with the index pages, a user generally required to navigate a n number of Web-pages to reach the content page they are looking for.

Index of pages plays vital role, when index pages of website are not well designed then user will suffer lot for finding relevant pages they are interested in. suppose some websites have irrelevant data which is not required by user in that case user will unsatisfied. So web-page recommender systems have become important for helping the web users to discover the most interesting Web-pages on particular websites. For making attractive web-page recommendations to users it does not contain extra data that is irrelevant data so every user satisfied by getting interested content.

Nowadays the Internet has become very popular. Millions of people access the Web to search information, do online shopping, learning new things. An ontology [4] may include individuals, classes, attributes, relations, restrictions, rules, and axioms. Ontology model allows human and machine understandable content and human-machine interaction. Domain Knowledge refers to important information that expert use for recommending web pages to user in that case Ontology [4] are used. The backbone method for knowledge representation is ontology. Ontological representation of the knowledge can be machine understandable and can help in interpreting and reasoning about the Web access patterns discovery in the mining step. To improve the results PREWAP [1], [2] algorithm is used over the PLWAP Mine, the both algorithms are works together [5] to get frequent sequential patterns.

An efficient web-page recommender system can be developed to offer the N most visited web-pages to web users from the currently visited web pages based on the system. Ontology's have been constructed by system developers in recommender system with domain experts. Ontology development is a critical process [4] which is inordinate and demands a high state of expertise in the domain. It is challenging to design and develop an ontology for a website because there are high number of pages on one website.

When user wants to visit the page that is not present then system cannot offer any web page recommendationto such user, Web-page that is not present in the discovered Web access sequence that problem is called as "new-page problem"[1]. Domain ontology used to resolve such new-page problem.

Section I describes introduction about web page recommendation system, section II describes literature survey, section III includes proposed work where we see the system architecture, modules description, mathematical models, algorithms and experimental setup, section IV describes expected results , and at last section V concludes the paper.

II. Related Work

In Paper [1] author proposes a novel method to provide better Web-page recommendation via semantic-enhancement by integrating the domain and Web usage knowledge of a website. Using three models, two new models represent domain knowledge of a website. One is semi-automatically constructed ontology-based model known as DomainOntoWP, and second is automatically constructed, semantic network of Web-pages called as TermNetWP, also conceptual prediction model is proposed to develop weighted semantic network of frequently viewed terms, known as TermNavNet.

In paper [2] author proposes Sequential Stream Mining-algorithm (SSM) based on the efficient PLWAP sequential mining algorithm, which uses three types of data structures DList, PLWAP tree and FSP-tree to handle the complexities of mining frequent sequential patterns in data streams. SSM-Algorithm supports continuous stream mining tasks suitable for such new applications as click stream data. It is a complete system that fulfills all of the requirements for mining frequent sequential patterns in data streams. SSM-Algorithm features are: DList structure for efficiently storing and maintaining support counts of all items that are passing through the streams, PLWAP tree for efficiently mining stream batch frequent patterns, and the FSP tree for maintaining batch frequent sequential patterns. So SSM algorithm produces faster execution time.

In paper [3] author proposed a novel technique to incorporate the conceptual characteristics of a website into a usage-based recommendation model. Author described a method to combine usage information and domain knowledge based on ideas from bioinformatics and information retrieval. The results are promising and are indicative of the utility of domain knowledge. Similarity is calculated by investigating the use of information content.

In paper[4] author explained the scope and purpose of ontology for "E-learning technologies" course, discussed about manual development of domain ontology, and provide a brief introduction on formalisms for knowledge representation On the ontological level. Also how ontology development works for that require Determination of the purpose and scope of the ontology, listing important concepts for capturing the domain and organizing them in taxonomical structures also Consider the other types of relations and merging of separate taxonomical structures. Defining the properties of classes and the constraints of their values and the instances and Evaluation of results, discussion and conclusions.

In paper[5] author studies the performance of two existing algorithms, the pre-order linked WAP-tree mining algorithm (PLWAP-Mine) and conditional sequence mining algorithm (CS-Mine), with respect to their sensitivity to the dataset variability, and their practicality for web recommendation. The comparison shows CS-Mine performs faster than PLWAP-Mine, but the frequent patterns generated by PLWAP-Mine are more effective than CS-Mine.

In Paper [6] author uses concept-based approach to add semantics into the mining process and to generate more semantically related results, that is result which fulfill the requirements of user. Because web usage mining uses the term and frequencies to represent a web site for the mining process it leads to poor result. Poor results means users don't get what they are require so by using concept-based approach user will get what they require or interested.

In paper [7] the Markov model is an efficient and probabilistic model to calculate the likelihood of going to Web-pages. Each Web-page is checked to as a state in the Markov model. Specifically, the N-order Markov model can be known to the next recent visited page based on the previous N-1 visited pages. The probability of the N-order Markov model is greater than the lower-order system, however, the number of steps used in a large-order Markov model will gradually increases. Because the complexity is calculated by the several of stages, the complexity of a greater-order Markov model increases when utilizing it to model a large number of Web-pages. Crossover probabilistic predictive models based on the Markov model, for example, the element of clustering-based Markov model of, have indicated improved prediction exactness over the Markov model. So, the complexity of the Markov-based models has caused about when they utilize in Web-page recommender systems reason is there are a large number of pages in a website. One efficient approach to minimize complexity of a Markov-based model is to filter out the Web-pages in the Web usage information which is not relevant.

In paper [8] author discussed Web usage mining (WUM) is a valuable technique for investigating Web usage information to get it Web client navigation practices and discover valuable Web usage learning. For an e-commerce organization, WUM can be utilized for discovering viewpoint clients who likely make a huge number of buys, or foreseeing e-commerce exchanges focused around the perception of past visitors. In the setting of web-page recommender frameworks, WUM can be utilized to find Web usage learning to help clients to settle on better choices by recommending prevalent Web-pages to the clients or a more effective approach to arrange sites for Web-based applications. Picking a successful mining algorithm assumes an imperative part in prescribing the right level of data to online users. The objective of WUM is to capture, display, and investigate the behavioral patterns and profiles of clients associating with a site.

Paper [9] states limitation of Sequential Pattern Mining technique is the crucial state of space complexity, especially for websites that have a large number of Web-pages. Recent study has evaluated that the WAP-tree based outperforms the other pattern mining method, e.g. Apriori-based, and pattern-development based technique, in terms of memory.

III. Proposed System

A. *System Overview*

1) Web Log: It consist of raw data that is surfing history of web users. Web log file contain the web access sequences and the URL's set which are divided in the preprocessing phase to pass the content.
2) Preprocessing: Used to extract useful information from non-useful data that is raw data. Web log having data in the form of URL’s and web access sequence. And then passed to further modules.

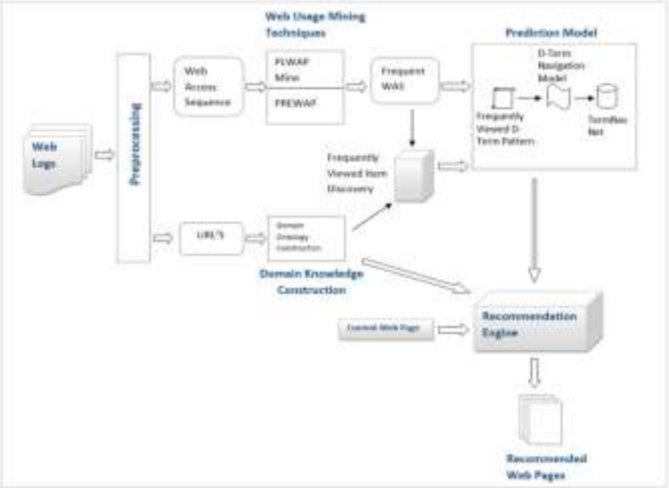

Fig 1. System Architecture

3) Web Usage Mining: This technique used to analyze web access pattern that are frequently used by web users. Because of analysis of web access pattern recommendation system will perform better then usage mining techniques are applied on that is PLWAP-Mine or PREWAP that will generate frequent view term discovery.
4) Domain Ontology Construction: Extract the important data from web logs process it and produce URL’s. Of web pages that are accessed by web users. That will generate frequent view term discovery.
5) Prediction Model: From frequent view term discovery prediction model will generate frequently viewed DTerm Patterns, and D-TermNavigation model is used to generate weighted semantic network of frequently viewed terms known as TermNavNet from frequently viewed D-Term Patterns.
6) Recommendation Engine: Recommendation Engine will recommend the web pages to the user.

B. *Algorithms*

Algorithm 1: Construction of PREWAP-Tree

Input: a database WASD, minimum support threshold λ (0 <λ≤|WASD|)

Output: a PREWAP-tree

```
Begin
(1) /* produce the root */
Add a ‘root’ node as the root node of PREWAP-tree T;

(2) /* produce ‘vent name’ and ‘occur’ values of node*/
For each Web access sequence S in WASD, do /* doing (a) and (b)*/

(a) Delete all the events in S which don’t meet the support $\lambda$, and gain frequent subsequence $S'$ ($e_1$, $e_2$...$e_n$). Set currentNode to the leftmost child of root in $T$;

(b) For $i$=1 to $n$(the length of $S'$) do /* doing (A) and (B)*/

        (A) If currentNode is NULL
                Create a new child node ($e_i$: 1);
        Else if currentNode is labeled $e_i$
                Set NodeExist to true;
        Else
                Set currentNode to currentNode’s sibling
                until $e_i$will be found or the sibling is NULL;

        (B) If NodeExist
                Increase count of $e_i$ by 1 and set
                currentNode; /*namely, for the node,
                Occur: = occur +1*/
        Else
                Create a new child node ($e_i$: 1), set
                currentNode.
                Update descendant links of $e_i$ ‘s ancestors;

(3) /* Produce ‘PREID’ and ‘desPREID’ values of node*/
Construct a header table used to store frequent events;
Then pre-visit T from root: ‘root left subtree right subtree’,
and at the same time add all nodes with the same event to a
linkage queue and record the serial number when one node is
visited;
(4) End.
```

Algorithm 2: Judging the relationship among two nodes in PREWAP-tree

Input: Two nodes α and β of PREWAP-tree

Output:

Case 1: α is $\beta's$ ancestor, return 0;
Case 2: α is $\beta's$ descendant, return 3;
Case 3: α is on $\beta's$ left tree, return 1;
Case 4: α is on $\beta's$ right tree, return 2.

```
Begin
If ($\alpha$. PREID <$\beta$. PREID and $\alpha$. decPREID ≥ $\beta$. decPREID)
        Return 0;
```

```
Else if ($\alpha$. PREID >$\beta$. PREID and $\alpha$. decPREID ≤ $\beta$.decPREID)
        Return 3;
Else if ($\alpha$. PREID <$\beta$. PREID and $\alpha$. decPREID<$\beta$.decPREID)
        Return 1;
Else
        Return 2;
End.
```

Algorithm 3:Mining on PREWAP-tree

Input: PREWAP-Tree T, the header table H (e_i represents its node), the minimum support λ ($0 < \lambda \leq$ |WASD|), ε-frequent pattern *F* (firstly *F* is empty), the a set of root of suffix tree *R* (*R* is equal to the root initially and $v_j$represents its node)

Output: ε+1-frequent sequence *F*

The main variable: C is used to store the count of the first node set of e_i.

```
Begin
(1)If R is empty
Return;

(2)Foreache$_i$ in H, find the suffix tree of $v_j$in T, do
        (a)While $e_i$ is not empty and $v_j$is not end
                Judge the relationship of ($v_j$->lSon, $e_i$) by
        calling Algorithm 2:
        Case 0: If $e_i$is not descendant of e visited in H
                        $C = C + e_i$.occur;
                If $e_i$ ->lSon is not empty
                        Push $e_i$ to R as R';
                        Next $e_i$ in H;
        Case 1: Next $v_j$in R;
        Case2, 3: Next $e_i$ in H;

        (b) If C is greater than or equal to $\lambda$
                Add $e_i$ after F as Patter F';
                Calling Algorithm 3 recursively but
        updating R and F with R' and F'
(3)End.
```

Algorithm 4: PLWAP-Mine

Input:WASD web access sequence database, minimum support.

Output:Complete set of frequent patterns.

Begin:

(1) The PLWAP algorithm computes frequent 1-items from the database transactions as F1 = {a: 5, b: 5, c: 3}, listing each event with its occurrence. It generates frequent sequences from each transaction.

(2) Using the frequent sequences, it builds the PLWAP tree by inserting each sequence from Root to leaf node.

(3) Mining the PLWAP tree to generate frequent pattern, by following the header linkage of the first frequent item.

C. *Mathematical Model*

System S is represented as,

S = (P, Σ, δ, q_0 , F)

P - A finite set of states

Σ - A finite set of input symbols called the alphabet

δ - A transition function (δ: P Σ P)

q_0 - Initial state (q_0 Є P)

F- Final state (F⊆ P)

Where,

P = {S_1, S_2}

Σ= {0, 1}

q_0 = S_1,

F = S_1, and

P = {S_1, S_2, S_3, S_4}

S_1: Pre-processing

S_2: Ontology

S_3: CPM

S_4: Recommended Web Pages

Σ= 0, 1, Input: Web log Dataset.

q_0 = S_1, Initial State: Pre-processing.

F = {S_4} Final output: Recommended web pages.

Prediction Model:

N = {(t_X, δ_X) | t_x Є *T}*: a set of term along with the corresponding occurrences counts,

Ø= {(t_x, t_y, $\delta_{x,y}$, $p_{x,y}$) | t_x, t_y Є *T }*: a set of transitions fromt_x to t_y, along with their transition weight *(*$\delta_{x,y}$*)*, and first- order transition probabilities *(*$p_{x,y}$*)*.

M = *{*t_x, t_y, t_z,$\delta_{x,y,z}$ ...$p_{x,y,z}$ *)* | t_x, t_y, t_zЄ *T}* : a set of transition fromt_x, t_y, t_z, along with their transition weights *(*$\delta_{x,y,z}$*)* and second order transition probabilities ($p_{x,y,z}$), If M is non empty, the CPM is considered as the second order conceptual predication model .

First Order Transaction Probability:

CPM states= {S, t1... tp,E}

N =|F| is the no of term pattern in F

Ps = First Order Transaction Probability

S and t_x are the two states

The first-order transition probabilities are estimated according to the following expressions:

1. The first-order transition probability from the starting state S to state t_x,
$$\rho s, x = \delta s, \frac{x}{\sum_{y=1}^{N} \delta s, y}$$
2. The first-order transition probability from state t_x to t_y,
$$\rho x, y = \delta x, y / \delta x$$
3. The first-order transition probability from state t_x to the final state E.
$$\rho x, E = \delta x, E / \delta x$$
4. The second-order probabilities are estimated as follows:
$$\rho x, y, z = \delta x, y, z / \delta x$$

D. Experimental Setup

The system is built using Java framework (version Jdk 1.8) on Windows platform. The NetBeans (version 8.0) is used as a development tool.

II. Dataset Description: The data was created by sampling and processing the www.microsoft.com logs. The data records the use of www.microsoft.com by 38000 anonymous, randomly-selected users. For each user, the data lists all the areas of the web site (Vroots) that the user visited in a one week timeframe.

Users are identified only by a sequential number, for example, User #14988, User #14989, etc. The file contains no personally identifiable information. The 294 Vroots are identified by their title (e.g. "NetShow for PowerPoint") and URL (e.g. "/stream"). The data comes from one week in February, 1998. Each instance represents an anonymous, randomly selected user of the web site. Each attribute is an area ("vroot") of the www.microsoft.com web site. Missing Attribute Values: The data is very sparse, so vroot visits are explicit, non-visits are implicit (missing).

IV. Results

This section presents the performance of the PLWAP Mine and PREWAP algorithms.

Fig 2 shows the time comparison of PLWAP Mine and PREWAP algorithms for various threshold size. The X-axis shows Threshold Size and Y- axis shows Time in ms. The PREWAP takes less time than PLWAP Mine when threshold gets increased.

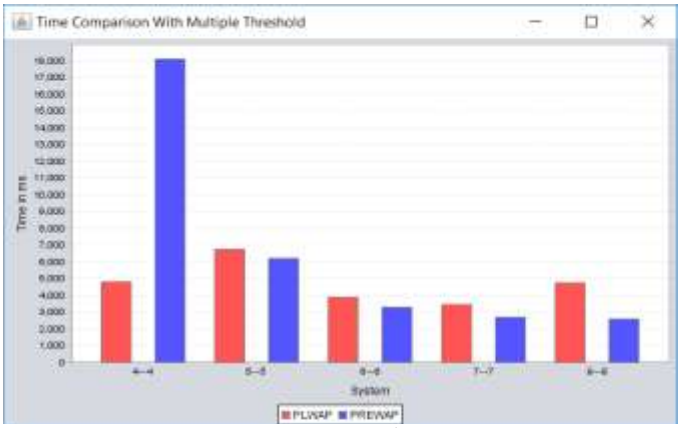

Fig. 2: Time Comparison Graph for various threshold

Fig 3 Shows Memory Comparison of PLWAP Mine and PREWAP algorithms for various Threshold. X-axis shows Algorithm & Y-axis shows Memory in bytes. PLWAP Mine require more Memory than PREWAP.

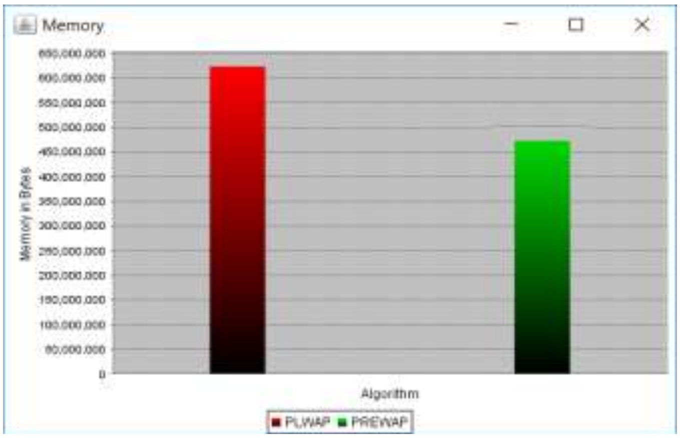

Fig. 3: Memory Comparison Graph

Fig. 4 Shows Time Comparison of PLWAP Mine and PREWAP algorithms for various dataset sizes. X-axis shows Dataset Size in kb and Y-axis shows Time in ms. PLWAP Mine require more time than PREWAP. So PREWAP is better than PLWAP Mine because it requires less time to execute when dataset gets updated than PLWAP Mine.

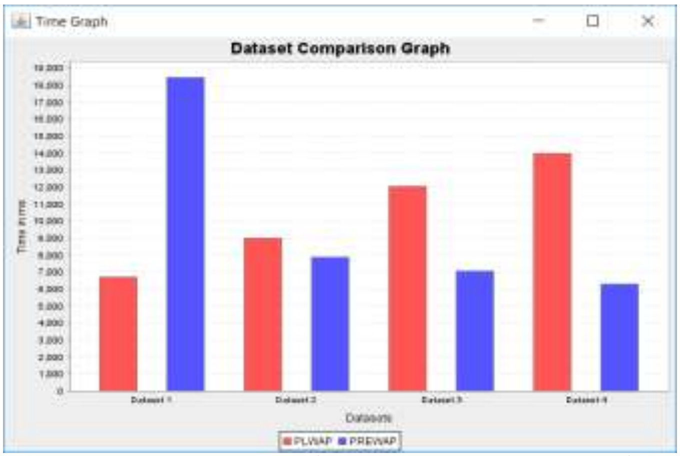

Fig 4. Time Comparison Graph of PLWAP and PREWAP algorithms for various dataset sizes.

V. Conclusion And Future Scope

Hence we conclude that PLWAP Mine and PREWAP Web usage mining techniques are used in Web-Page Recommendation System that will provide interested web pages to web user. When dataset gets updated and there are too many small frequent item sets generated in such case PLWAP Mine will not work properly in terms of execution time and it requires more memory. PREWAP take less time and less memory to execute than PLWAP Mine. So PREWAP is better than PLWAP Mine. Frequent viewed terms discovery and frequently viewed D-Term patterns are used by Recommendation system to recommend the interested web pages. Web usage mining techniques such as PREWAP and PLWAP Mine and Domain Ontology are used to generate Frequent patterns also new page problem is resolved by constructing domain ontology on URL's obtained from web logs.

Given the time constraints of this study, the scope of the system development was limited to applying to a single website, in future the further development can be focused on the support for Multi-site.

Acknowledgment

I take this opportunity to express my profound gratitude and deep regards to my guide Prof. H. A. Hingoliwalafor his exemplary guidance, monitoring and constant encouragement which helped me in completing this task through various stages. The blessings, help and guidance given by him time to time shall carry me a long way in the journey of life which I am about to embark. We are thankful to the authorities of SavitribaiPhule University of Pune.

References

[1] Thi Thanh Sang Nguyen, Hai Yan Lu, and Jie Lu OCTOBER 2014," Web-Page Recommendation Based on Web Usage and Domain Knowledge" IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 10

[2] ChristieI. EZEIFE 1, Monwar Mostafa, "a PLWAP-based algorithm for mining frequent sequential stream patterns", Springer.

[3] Amit Bose, Kalyan Beemanapalli, Jaideep Srivastava, Sigal Sahar, "Incorporating Concept Hierarchies into Usage Mining Based Recommendations", 2006 ACM 1-59593-444-8.

[4] Dale Dzemydiene, Lina Tankeleviciene, "on the development of domain ontology for distance learning course", ISBN 978-9955-28-283-9.

[5] Thi Thanh Sang Nguyen* and Hai Yan Lu, "Investigation of sequential pattern mining techniques for web recommendation", Int. J. Information and Decision Sciences, Vol. 4, No. 4, 2012.

[6] Sebastian A. Rıos Juan D. Velasquez, "Semantic Web Usage Mining by a Concept-based Approach for Off-line Web Site Enhancements" IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 2008.

[7] Borges, J. & Levene, M. 2004, A Dynamic Clustering-Based Markov Model for Web Usage Mining, Available online at http://xxx.arxiv.org/abs/cs.IR/0406032.

[8] Liu, B., Mobasher, B. & Nasraoui, O. 2011, 'Web Usage Mining', in B. Liu (ed.), Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Springer- Verlag Berlin Heidelberg, pp.527-603.

[9] Mabroukeh, N.R. & Ezeife, C.I. 2010, 'A Taxonomy of Sequential Pattern Mining Algorithms', ACM Comput. Surv. vol. 43, no. 1, pp. 1-41.

[10] Ezeife, C. & Liu, Y. 2009, 'Fast Incremental Mining of Web Sequential Patterns with PLWAP Tree', Data Mining and Knowledge Discovery, vol. 19, no. 3, pp. 376-416.

[11] Henze, N., Dolog, P. & Nejdl, W. 2004, 'Reasoning and Ontologies for Personalized E-Learning inthe Semantic Web', Educational Technology & Society, vol. 7, no. 4, pp. 82-97.

[12] Khalil, F. 2008, 'Combining Web Data Mining Techniques for Web Page Access Prediction', Doctoral thesis, University of Southern Queensland.