# Web Log Mining using Multi Item Sequential Pattern based on PLWAP

Mr. Mukund Patel

*Department of Computer Engg, SCET,Rajpur, Kadi, Gujarat, India, Mobile: 9662712400,*
*Email:mukund27p@gmail.com*

Mr. Ramesh Prajapati

*Dept. of Computer Science Engineering*
*Rai University, Center for Research & Development*
*Saroda , Dholka , India*
*Email:rtprajapati@gmail.com*

Dr . Samrat Khanna

*Dept. of Information Technology*
*Istar , Sardarpatel centre for science & Tech.*
*V.Vnagar, India*
*Email: sonukhanna@yahoo.com*

Keywords: WAP-Tree (Web Access Pattern Tree), Sequential Pattern Mining, PLWAP, Position coded feature, Web Usage Mining, Web log Mining, MPLWAP, Multi-item pattern mining

**ABSTRACT**

Sequential pattern mining serves as a key for problems in web log mining, principally for web usage mining. Research on sequence mining seeking faster algorithms.PLWAP-Tree based on WAP Tree is an algorithm of the SPM, widely used algorithm especially for web log mining and it has phenomenal performance for single item sequence. In this research, we scrutinize multi-item sequence and we present Multi-PLWAP-Tree which extends PLWAP-Tree for multi-item sequence databases. Basically we propose new algorithm MPLWAP-Tree based on PLWAP. Mining the MPLWAP-Tree system can generate more frequent and interesting pattern from web log data.

## I. Introduction

**Sequential Pattern Mining (SPM) [3]** is major as well as important mining task of uncover habits of users with respect to time in sequence databases. Sequential Pattern mining is a area of data mining related with finding statistically useful patterns between the real data where the sequence of values are important.SPM is an application of Web Log Mining(WLM).It helps recognizing user navigation pattern on web site than can guide processes like recommendation.

PLWAP-Tree (Pre Ordered linked Web Access Pattern Tree) based on WAP-Tree (Web Access Pattern-Tree) have shown phenomenal performance on single-item SPM [10].It is a hybrid algorithm of pattern growth and early pruning algorithms of SPM. Among all the algorithms of SPM PLWAP is one of the algorithms which is best used as Web Log Mining. Inspired by the result of PLWAP-Tree, we propose a new data structure MULTI-PLWAP-Tree which extends PLWAP-Tree for representing multi-item/general sequence databases. Secondly, we propose an algorithm to mine that MPLWAP-Tree algorithm.
The rest of the paper is composed of four sections. In next section, section 2, we discuss background information on existing system and proposed algorithm. We present theoretical analysis and in Section 3 and Performance Analysis and Experimental Evaluation in section 4. At last, we conclude in Section 6.

## II. Proposed System

**Existing System**: "Sequential pattern mining concerned with finding user navigational patterns on the world wide web by extracting knowledge from web logs, where ordered sequences of events in the sequence database are composed of single items and not sets of items, with the assumption that a web user can physically access only one

 

web page at any given point in time" **[9]**. The propose method is to assign multiple pages which could be accessed in a specific time interval from the same parent node but till now we could not see this consideration, so the probable approach is basically to perform sequential pattern mining using this approach.

**Proposed algorithm** constructs a tree which is based on PLWAP-Tree using multi-item sequence. We consider the multi-item sequence based on the parent page (referred url) of the particular item(web page).The concept behind capturing the referred page is to check the referred page of different items which were surfed by a user in specific time interval. If the referred pages are same for the specific user and session we put those pages in a same node to generate tree. So basically the idea is to consider multiple pages instead of single page while mining the weblog data. Instead of considering single web page as a single node we can store multiple pages in a single node which were surfed by user in a specific time interval and to find the same information. By doing this we can find the frequent pattern and also recommend more pages.

**Proposed Algorithm to generate MPLWAP tree**

**Input:** Web Access Sequence

**Output**: MPLWAP Tree

MPLWAP scans the access sequence database first time to acquire all events. Only those events will be considered whose support count is greater than or equal to minimum support.

Identify the number of events in a single node.

Each node in a MPLWAP tree registers three information: number of items, pointer to the item, position code. And Item registers information: **count**.

MPLWAP scans the database again (second) time to acquire frequent sequence S.

Build tree data structure.

--Considering the first web page (event), increment the count of the same if exist, otherwise

--Check parent page of current node and for the first event is same

---If yes same then put event into current node and increase number of items to the node and also assign count 1 to that event

--Otherwise create new child node and set count of that to one for event and make that event as current node. Also assign position code for that.

-Add current node to the sequence.

The proposed algorithm differs from existing PLWAP in following points:

**Input Structure (<abcdef> vs. <(ab)c(def)>) :** After considering all web page as a single node the input structure of our propose algorithm looks different. For example in existing system for all the web pages after user identification and session identification the input structure looks like < abcdef > and in the proposed system after considering the web pages as a group using referred url the input structure look like <(ab)c(def)>.

**Node Structure: After considering the input structure** to create tree in the existing system each node itself a single item and it has three label namely node: label, count, and position code. While in the proposed system because the input structure is different so the node structure of tree is also different form existing system. The node

 

structure of proposed system registers three information: number of items, pointer to the item, position code. And Item(s) in the node registers information: count of item.

**Insertion of Web page:** As above mentioned the different input structure and different node structure, the insertion of new web page (item) also differs from the existing system. When new item comes in the PLWAP tree first it checks that it exist or not .It increments the count by 1 if it exists else it will be inserted as a child node according to the proper sequence and assign position code and set count to 1.And in proposed system the node contains multiple items, while in MPLWAP it checks the node for the specific item and if exists it increments the count of item otherwise it will be inserted as a child and set the position code according to the rule. And so for the recommendation point of view we can propose more relevant pattern to the user.

## III. Theoretical Analysis

MPLWAP scans the access sequence database first time to acquire all events in the given event set, E. Only those events will be considered whose support count is greater than or equal to minimum support. Each node in a PLWAP-tree registers three pieces of information: node: label, count and position code, denoted as label: count: position [10]. While the node structure of proposed system registers three information: number of items, pointer to the item, position code. And Item(s) in the node registers information: count of item. Then MPWAP scans the database a second time to acquire the frequent sequences in each transaction. The non-frequent events in each transaction are deleted from the sequence. After the scans the dataset is mentioned in table 2. But here in proposed system we add one more step that to put all the web page (item) whose parent id (referred url) is same. After doing this step for proposed algorithm the final sequence is mentioned in table 2

**Table: 1 Frequent 1 item set (Support Count: 3)**

| TID | Web Access Sequence |
|---|---|
| 1 | 11,92,92 |
| 2 | 29,11,92 |
| 3 | 29,11,92 |
| 4 | 29 |

**Table: 2 For MPLWAP TREE data sequence using referred url**

| TID | Web Access Sequence |
|---|---|
| 1 | (11,92 ) , 92 |
| 2 | (29,11) ,92 |
| 3 | 29 , (11,92) |
| 4 | 29 |

 

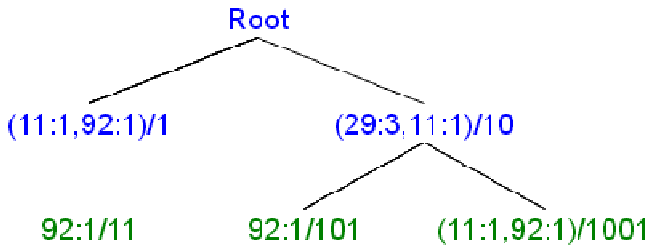


**Fig: 1 Manual analysis for proposed MPLWAP Tree**

## IV. Performance Examination and Experimental Assessment

This section compares the performance of both the algorithms, PLWAP and MPLWAP. We are using synthetic datasets. This datasets are produced using the easily available synthetic data generation program of the IBM Quest data mining project at http://www.almaden.ibm.com/cs/quest/, which data has been widely used in projects of SPM and some other mining research work.
We are using another dataset which is real time web log file of an E-Commerce website.

**Experiment 1: No of Patterns for different support on synthetic dataset**

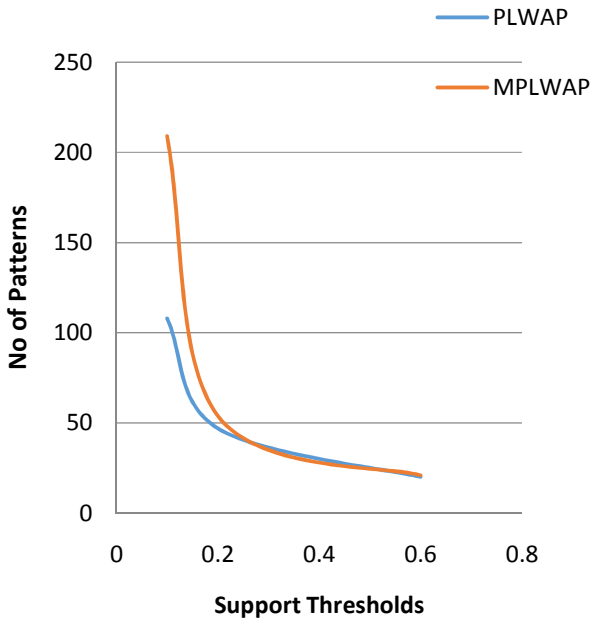


**Fig: 2 No of patterns with different threshold**

This experiment is based on the fixed size synthetic dataset with different threshold value. The algorithms are tested with different threshold value.

**Experiment 2: No of Patterns for different support on Web log data**

 

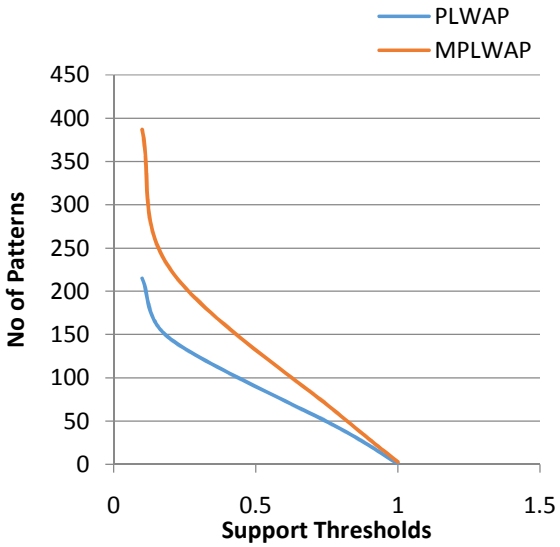


**Fig: 3 No of patterns with different threshold**

This experiment is based on web log data with different threshold value.

**Experiment 3: Execution time for different support**

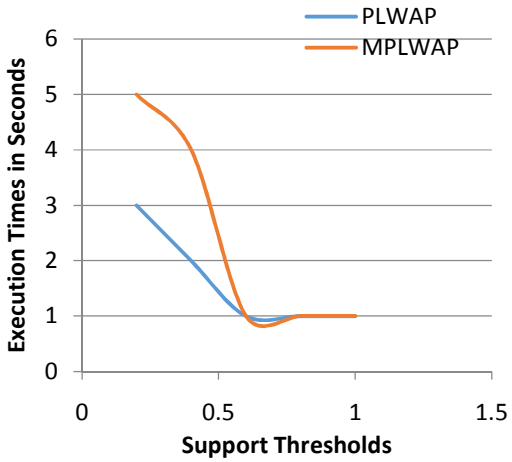


**Fig: 4 Execution time with different support**

This evalution is based on the fixed size synthetic dataset with different threshold value. The algorithms are tested with different threshold value between 0 to 1.

 

## V. Conclusions and Future Enhancements

MPLWAP – a proposed algorithm in this paper, improves on finding frequent patterns by accommodating multiple pages in a single node instead of single node as done by PLWAP mining. Even though the execution time of MPLWAP is higher than PLWAP, the patterns generated from MPLWAP are more than PLWAP mining algorithm. Experiments show that mining of MPLWAP tree gives more patterns than PLWAP tree.
Future work should consider applying MPLWAP-tree mining techniques to distributed mining as well as to incremental mining of web logs and sequential patterns

Thus if we consider multi-item sequence, we can extract useful patterns from the web log data and it can be useful for web recommendation and personalization.

## References

[1]http://en.wikipedia.org/wiki/Web_mining

[2]http://en.wikipedia.org/wiki/Sequential_Pattern_Mining

[3] Review on sequential pattern mining Algorithms
Sushila S. Shelke, Suhasini A. Itkar 2015 IJEECE.
[4] Web usage mining using improved frequent pattern tree algorithm Ashika Gupta , Rakhi arora, Ranjana sikarwar , Neha Saxena.IEEE-2014
[5] A survey on improving the efficiency of prefix span sequential pattern mining algorithm.K Suneetha , Dr. M Usha Rani. IJCCIT-2014
[6] PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth .Jian Pei,Jiawei Han, Behzad Mortazavi-Asl,Helen Pinto.IEEE-2013.
[7] Sequential Pattern Mining Methods: A Snap Shot.Niti Desai, Amit Ganatra IOSR-JCE-2013
[8] Sequence Pattern Mining:Surveyand current research challenges.Chetna Chand ,Amit Thakkar ,Amit Ganatra.IJSCE-2012
[9]A Taxonomy of Sequential Pattern Mining Algorithms.Nizar R.Mabroukeh , C.I. Ezeife.ACM-2010.
[10] Position Coded Pre-order Linked WAP-Tree for Web Log Sequential Pattern Mining Yi Lu and C.I. Ezeife2003.
[11] M. Zaki, „SPADE: An Efficient Algorithm for Mining Frequent Sequences‟, Machine Learning, vol. 40, pp. 31-60, 2001.[12] Han J., Dong G., Mortazavi-Asl B., Chen Q., Dayal U., Hsu M.-C.,‟ Freespan: Frequent pattern-projected sequential pattern mining‟, 2000, pp. 355-359.
[13] M. Garofalakis, R. Rastogi, and K. Shim, "SPIRIT: Sequential pattern mining with regular expression constraints", VLDB'99, 1999

[14] Srikant R. and Agrawal R.,‟‟Mining sequential patterns: Generalizations and performance improvements‟‟, Proceedings of the 5th International Conference Extending Database Technology, 1996, 1057, 3-17..

 