Multidimensional Web Access Pattern Tree (MD-WAP Tree)

Satendra K. Jain
S.A.T.I.,Vidisha

ABSTRACT
Mining frequent web access patterns from large data (web log) is one significant application of sequential pattern mining. Web access patterns are set of frequent sub sequences that are useful to know user behaviour in real time in order to make dynamic decisions. Techniques for extracting web access patterns from data available in two flavours: apriori based and non apriori based (tree based). It has been observed that extracting web access pattern with respect to multiple dimensions gives interesting results rather than considering one dimension. In this paper a very interesting data structure, multidimensional web access pattern tree (MD-WAP Tree) is presented that can discover web access pattern with respect to multiple dimensions known as multidimensional web access patterns.

Keywords
web access pattern,apriori,wap-tree, multidimension.

1. INTRODUCTION
Sequential pattern mining is an important data mining problem used in many domains, like customer relationship management (CRM), DNA analysis, health, educations, learning behaviour etc. It can be defined as to identify frequent subsequences from sequence database. Sequence database consists with multiple sequences. Web log data can be considered a kind of sequence database. It is generated at many places like web server, client machines ect., when an user surf a web site. A sequence in web log data can be one session of a user. Find frequent sequences from web log data is known as web access pattern. There are two flavours to find web access patterns (1) apriori based and (2) non apriori based (tree based). [6] proposed apriori based approach also known as candidate generation and test methodology to find frequent subsequences. It is a two step approach, in step first frequent sub sequences are generated that are used in the second step to generates the strong rules. Frequent sub sequences are generated by candidate sub sequences that satisfy user specified minimum support. The nature of apriori based technique is combinatorial due to large set of candidate generation specially when the data items are large. Apriori based techniques read sequence database multiple times. It is equal to the size of maximal length of pattern in worst case. These limitations of apriori based techniques draw the attention towards non apriori based(tree based) techniques. [4,5] has given an idea to construct a tree data structure of sequences and it is known as a wap-tree. Such techniques scan database less time as compared to apriori based. Further many modifications on wap-tree have been introduced in the literature. [2] has given an idea of PL-wap tree that removes some limitations of wap-tree.

In this paper the concept of wap-tree is inherited and modified it by incorporated the multiple dimensions and introduced a new and powerful data structure MD-WAP tree that can generate more interesting frequent web access pattern with respect to multiple dimensions. As an example, it may found from a web log that "mostly a web page B.html is accessed after A.html by students in the summer season at Delhi". This rule includes many dimensions like group of person, time, region etc.

The rest of the paper is organized as follows: In section two, web log, sequence and multidimensional sequence database is discussed in brief. Some well known existing tree based techniques to find frequent web access patterns are summarized in section three. Algorithm and implementation of MD-WAP tree is given in section four. Finally in section five we conclude this paper.

2. BASIC CONCEPTS
[3] explained in detail how interesting knowledge can be discovered from web log data. [8] suggested that raw web log should be pre processed before mine. Five steps need to apply on raw web log data to transform it into a transaction database, because existing techniques discussed in this section used transformed form of web log data.

These steps are data cleaning (elimination of irrelevant information not required in mining process), user identification (identification of unique user is must by various heuristic methods), session identification (page accesses by a user in a specified time period), path completion (construction of complete and consistent navigational path) and formatting (data need to format properly such as relational database etc.).

As an example Equivalent transaction database of web log is shown in the table 1, that may be the outcomes of these steps. Each row in this table shows a user's session also known as access sequence. Each access sequence consists with number of events. Here a web page can be considered as an event. An access sequence may contain various sub sequences. For example in table1,E.html,A.html,E.html,B.html,C.html,A.html,C.html, is one access sequence. A sub sequence A.html,B.html,,A.html,C.html, exist in this sequence. Mining web access pattern is to discover frequent sub sequence from web sequence database. Frequent sub sequences always satisfy user define minimum support value.

Table 1. Web access sequence database

Transaction Id	Web Access Sequences
100	A.html,B.html,D.html,A.html,C.html
200	E.html,A.html,E.html,B.html,C.html,A.html,C.html,
300	B.html,A.html,B.html,F.html,A.html,E.html,C.html,
400	C.html,A.html,B.html,F.html,A.html,E.html,C.html,

Similarly a multidimensional web access sequence database can be created by adding up the dimensions along with sequences. It is shown in table2.

Table 2. Multidimensional web access sequence database

Transaction Id	Dimension1 (city)	Dimension 2 (profession)	Dimension3 (season)	Web access sequences
100	Delhi	Student	Summer	A.html,B.html,D.html ,A.html,C.html
200	Mumbai	Manager	Winter	E.html,A.html,E.html ,B.html,C.html,A.html,C.html,
300	Delhi	Student	Summer	B.html,A.html,B.html ,F.html,A.html,E.html ,C.html,
400	Kolakata	Teacher	Summer	C.html,A.html,B.html ,F.html,A.html,E.html ,C.html,

3. TREE BASED TECHNIQUES FOR WEB ACCESS PATTERNS

Numerous existing tree based techniques are summarized in this section that efficiently store and compress web sequence database in order to discover web access patterns.

[4] suggested an efficient method to find web access patterns from web log data that is completely different from apriori like algorithm. In this paper a new data structure WAP-tree (Web Access Pattern tree) of web log is created that required only two scans of database. In first scan, events that are not frequent are identified and removed from each sequence of sequence database because super sequence of any infrequent sequence can not be frequent. WAP-tree is constructed in the second scan by the sequence database that is modified in first scan. Web access patterns can be discovered by WAP-tree without further scanning of database by conditional search. Conditional search reduce the search space by looking for patterns with the same suffix and find frequent events in the set of prefixes with respect to the suffix. Unlike level wise searching as in apriori algorithm it is divide and conquer based approach. Authors proved experimentally that WAP-tree algorithm works efficiently than Generalized Sequential Pattern (GSP) mining algorithm given by [7] that is based on candidate generation and test approach(apriori heuristic). Figure 1 demonstrates the WAP-tree of sequence database shown in table 1.

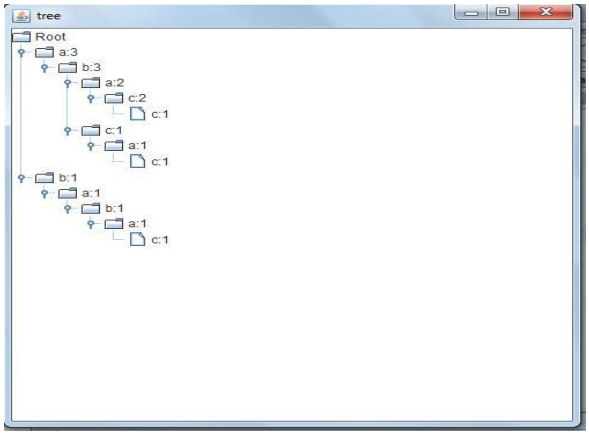

Figure 1. **WAP-tree**

Major drawback of WAP-tree algorithm is recursively reconstruction of WAP-tree during mining of web access patterns that is time as well as memory consuming process. [2] motivated by drawbacks of WAP-tree and suggested a new data structure known as PLWAP-tree (Preorder Linked WAP-tree). Like WAP-tree, creation of PLWAP-tree of web log is also required two database scans. PLWAP-tree stores the web access sequence in pre order linked along with position code of nodes. Web access patterns can be discovered by PLWAP-tree without further scanning of database. Instead of searching common suffix patterns like in WAP-tree, PLWAP-tree search common prefix patterns and avoid reconstruction of tree during mining process. PLWAP-tree work on the principle that to analyze the suffix tree of last frequent event in n-prefix sequence and extend it to n+1 prefix sequence by adding any frequent event in the suffix tree. Authors proved experimentally that PLWAP-tree mine web access pattern efficiently than WAP-tree and less time and memory. Figure 2 shows the PLWAP-tree of sequence data shown in table 1.

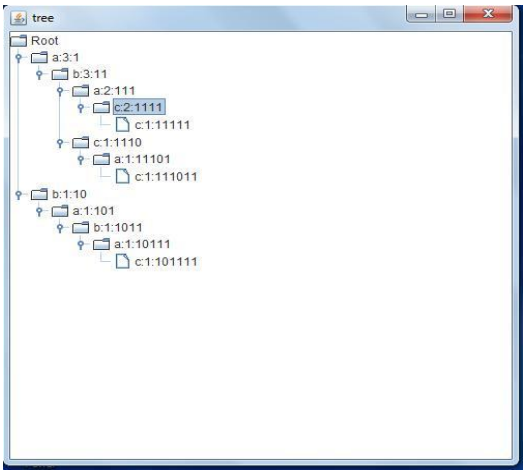

Figure2. **PL-WAP Tree**

[9] proposed doubly linked tree in order to find web access patterns. In doubly linked tree each node contains a pointer to parent as well as pointer to child node. This makes easy traversing of branches in backward and forward direction. Mining web access patterns from doubly linked tree is much similar to WAP-tree. It is experimentally proved that it work efficiently than Generalized Sequential Pattern (GSP) mining algorithm.

[1] explained that only a single minimum support assume that

all items in the database have the same nature that can not be the case in the real life applications usually. Some times it is important to capture the rules involving less frequent items along with rules having frequent items. Consider an example that in the super market people purchase Microwave Oven along with Cooking Pan much less frequently than Bread and Milk. In general Microwave Oven and Cooking Pan may be durable and/or expensive but generate more profit to store. [1] explained the algorithm MS-GSP, that use multiple minimum supports (MMSs) to address above said problem. Due to candidate generation and test nature of MS-GSP it is costly and time consuming.

[10] introduce a new data structure known as PLMS-tree(Pre order Linked Multiple Support tree). It is an extension of PLWAP-tree proposed by (C. Ezeife, et al., 2005). It store and compress all necessary information from a sequence database. Once the PLMS-tree is created an efficient a PLMS-tree based mining algorithm MSCP-growth (Multiple Supports – Conditional Pattern growth) is applied on it in order to find complete set of sequential patterns with multiple minimum supports. Experimentally it is proved that it outperforms MS-GSP.

4. MD-WAP TREE

MD-WAP tree is constructed in two phases. The first phase is called an embedding phase. In the embedding phase dimensions are incorporated in the sequences as an element. The second phase is called the construction phase. In the construction phase an MD-WAP tree is constructed. The process of this two phase method is demonstrated in the block diagram which is shown in figure 3. As an example the embedding phase is applied on the multidimensional sequence database that is given in table 2, that produce a modified database which is shown in table 3. The overall process is summarized in the algorithm which is given below. The algorithm is implemented in java Swing by JTree classes.

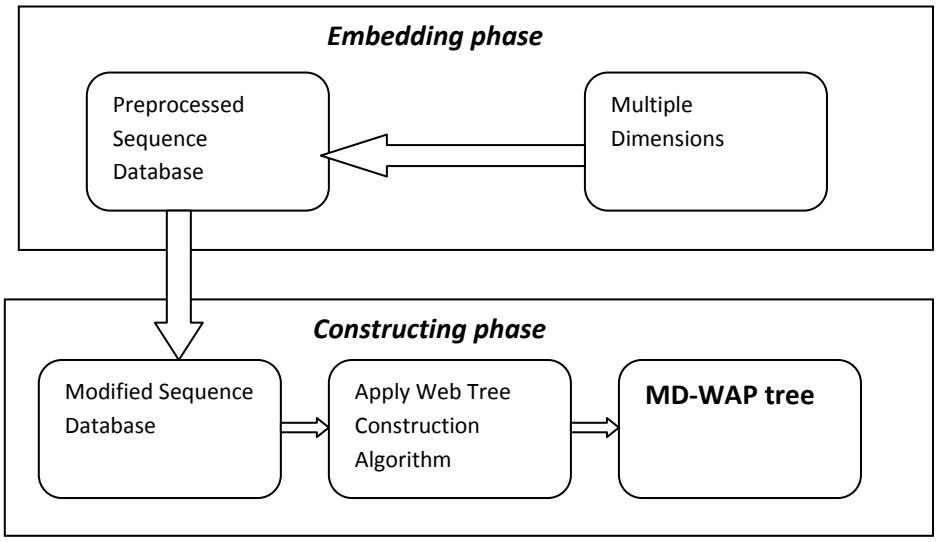

Figure 3. Block diagram to construct MD-WAP tree

Table 3. Modified Sequences with dimensions

Transection Id	Sequences with dimensions
100	<delhi, student, summer, A.html,B.html,D.html,A.html,C.html>
200	<Mumbai, manager, winter, E.html,A.html,E.html,B.html,C.html,A.html,C.html>
300	<delhi, student, summer, B.html,A.html,B.html,F.html,A.html,E.html,C.html,>
400	<Kolkata, teacher, summer, C.html,A.html,B.html,F.html,A.html,E.html,C.html,>

Algorithm: MD-WAP tree

Input: Web access sequence database, Multiple dimensions.

Output: MD-WAP tree.

Method:

1. Scan web access sequence database.
2. Embedded dimensions as an elements in sequences database.
3. 3. events
4. Construct a tree of sequences recursively.

5. CONCLUSION

The main aim of this paper to represents a new and interesting data structure MD-WAP-tree. It will be very efficient to discover multidimensional web access patterns by MD-WAP tree as compared to apriori based method. In future instead of single support value, MD-WAP tree may be used for multiple minimum support values to discover interesting and perfect strong rules. The same is also proposed to implement in the distributed environment.

6. REFERENCES

[1] Bing Liu, Web data mining: exploring hyperlinks, contents and usage data, 2nd ed. springer, 2008.

[2] C. Ezeife, Y. Lu, Mining web log sequential patterns with position coded preorder linked WAP-tree, Data

mining and knowledge discovery, 10, pp5-38, 2005.

[3] F. M. Facca, P. L. Lanzi, Mining interesting knowledge from weblogs: a survey, Data knowledge engineering, 53, pp 225-241.

[4] J. Pei, J. Han, B. Mortazavi-asl, H. Zhu, Mining access patterns efficiently from web log, Lecture note in computer science, 1805, pp 396-407, 2000.

[5] J. Han, J. Pei, Y. Yin, R. Mao, Mining frequent patters without candidate generation: A frequent pattern tree approach, Data mining and knowledge discovery, 8, pp 53-87, 2004.

[6] R. Agrawal and R. Shrikant, Fast algorithm for mining association rules. in proc. of int. conf. on very large databases(VLDB'94), Santiago Chile, pp 487-499, 1994.

[7] R. Agrawal and R. Shrikant, Mining sequential patterns: generalization and performance improvement, in proc. of 5th int. conf. on extending database technology(EDBT), Avignon france, pp 3-17, 1996.

[8] R. Cooley, R. Mobasher, J. Shrivastava, Data preparation for mining world wide web browsing pattern, Knowledge and information systems 1(1), pp 5-32, 1999.

[9] R. K. Jain, R. S. Kasana, S. Jain, Efficient web log mining using doubly linked tree, International journal of computer science and information security, vol. 3, no. 1, 2009.

[10] Ya-Han Hu, F. Wu, Yi-Jiun Liao, An efficient tree based algorithm for mining sequential patterns with multiple minimum support, The journal of systems and software, 86, pp 1224-1238, 2013.