Tree based Techniques for web access patterns – a survey

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

Abstract

Mining frequent web access patterns from web log data is one important application of sequential pattern mining. Web access patterns are useful to anticipate user behaviour in real time. Techniques for extracting web access patterns from web log data available in two flavours: apriori based and non apriori based. In this paper we present a survey of non apriori based (compact prefix tree) techniques to generate interesting web access patterns. The aim of this paper to show the efficient tree based techniques for web access patterns along with their modifications and benefits over apriori based approaches.

Keywords: web access pattern, apriori, compact prefix tree, wap-tree.

1. Introduction

Sequential pattern mining is an important data mining problem used in many domain, like customer relationship management (CRM), DNA analysis, 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 web server when a 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 (compact prefix tree). [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 sequential pattern is large. Apriori based techniques read sequence database multiple times. In the worst case it is equal to the size of maximal length of pattern. These shortage of apriori based techniques draw the attention towards non apriori based techniques. Such techniques scan database less as compared to previous one. Web access patterns, frequent sequence pattern or frequent access patterns are used interchangeably in this paper.

The rest of the paper is organized as follows: In section two, web log data is defined. Existing tree based techniques to find frequent access patterns are summarized in section three. Finally in section four we conclude this paper.

2. Basics of Web Access Patterns

[3] explained in detail how interesting knowledge can be discovered from web log data. Various data mining techniques are defined on web log data in this paper. Sequential pattern mining is one important techniques which can be applied on web log data to find sequential navigational patterns that appears in user’s session frequently. Log data are generated at web server when a user surf a web site. It is usally available in unstructured form. A web log is a sequence of events. [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 section three 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.). Equivalent transaction database of web log is shown in the table 1, that will 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 is considered as an event. An access sequence may contain various sub sequences. For example in table 1, eaebcac is one access sequence. A sub sequence abac 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

<table>
<thead>
<tr>
<th>Transaction Id</th>
<th>Web access sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>abdac</td>
</tr>
<tr>
<td>200</td>
<td>eaebcac</td>
</tr>
<tr>
<td>300</td>
<td>babfaec</td>
</tr>
<tr>
<td>400</td>
<td>cabfaec</td>
</tr>
</tbody>
</table>

3. TREE BASED TECHNIQUES TO ACCESS WEB PATTERNS

Numerous 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 demonstrate the WAP-tree of sequence database shown in table 1.

![Figure 1. WAP-Tree](image)

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.
[15] proposed a new data structure known as STMFP (Sequence Tree for Mining Frequent Patterns) in which each tree node may contain more than one data item. STMFP is constructed by scanning database one time. In STMFP, frequent access pattern can be mined by combining nodes from leaf to root. It experimentally proved that this method works efficiently than apriori based techniques.

[13] introduced a novel data structure called IFP-tree (Incremental Frequent Pattern tree) for interactive mining with multiple support. It is an extension of FP-tree[5]. It handled the rapidly growing web log data. Outdated instance can be removed efficiently by efficient insertion and deletion operations.

[9] proposed a 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 works efficiently than Generalized Sequential Pattern (GSP) mining algorithm.

[10] proposed a MAP-tree (Mining Access Pattern) structure in order to generate frequent patterns. Creation and mining of MAP-tree is much similar to WAP-tree. MAP-tree is also applied on multilevel data and experimentally proved that it works 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 cannot 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.

[14] introduced 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.

[11] proposed a new data mining approach that generate frequent tree-like patterns. Tree-like patterns show the relationship for the pair of items in a sequence. The theme of this data mining approach is inherited from apriori based approach. In this new approach frequent tree-like patterns are discover by candidate generation and test approach. A novel method is also discussed in this paper to count supports of tree-like pattern efficiently using queue structure.

[12] proposed an improved version of traditional WAP-tree known as IWAP-tree (Improve WAP tree). IWAP-tree generates web access patterns efficiently by avoiding the constructions of Conditional WAP three that is a need of traditional WAP-tree. IWAP shows better stability as the length of patterns vary.
4. CONCLUSION

The main aim of this paper is to represent the tree-based techniques to discover the web access patterns that work efficiently than apriori-based approaches. It is found after going through various research papers that conventional WAP-tree is used and modified to discover web access patterns efficiently. This method can also be applied to discover multidimensional web access patterns in future.

REFERENCES