Discovery of Frequent Usage Pattern for Web Data to Optimized Web based Applications

Jaswinder Kaur
Research Scholar
Dept. of Computer Science & Applications,
Kurukshetra University, Kurukshetra, Haryana, India

Kanwal Garg, PhD
Assistant Professor
Dept. of Computer Science & Applications,
Kurukshetra University, Kurukshetra, Haryana, India

ABSTRACT
The traffic on WWW is increasing at a rapid rate due to users interaction with web sites, these activities contributes to enormous information which is maintained in Web log file. Web usage mining plays an important role in discovering frequent pattern from Web data, that helps to better serve the need of Web based applications. In present research work, researcher finds out different user and their session, which help in identifying unique user’s navigational path from pre-processed Web log data. Further, researcher also proposed Modified Apriori algorithm which helps in extracting frequent usage pattern using average support.

Keywords
Frequent Pattern, Average Support, Web Usage Mining, Web data, Pre-processed

1. INTRODUCTION
Web log mining also known as Web Usage mining is application of mining techniques aimed to obtain interesting and frequent user access patterns from web log files, browser logs or proxy server logs. Web usage mining consist of three important phases which are preprocessing, pattern discovery and pattern analysis. In preprocessing, conditioning of data is carried out as real world Web data may contain anomalies which are to be removed to have data with useful information. During second phase, which is data mining, techniques are applied on pre-processed data to discover frequent usage patterns. Various techniques used for pattern discovery are stastical analysis, clustering, classification, association rule and sequential pattern. In third phase, which is termed as pattern analysis, where obtained usage patterns are analyzed to filter irrelvent information and extract the valuable information [1].

In this paper, pre-processed web access log file of size 11,625 KB is used to discover frequent pattern and is orgnized as follows: section 2 presents related work which concerns in discovering frequent pattern from web data. section 3 describes the way that helps in finding user and session. Modified apriori algorithm is explained in section 4. Experiment results is discussed in section 5, Finally section 6 concludes the paper.

2. LITERATURE REVIEW
Pattern discovery is an essential phase in Web usage mining. At this stage, data mining techniques is applied on pre-processed Web log data in order to extract frequent user pattern. Veeramalai et.al. (2010)[2] proposed Enhanced Modified Apriori hash tree with fuzzy algorithm is used to overcome Crisp boundary problem. Kiruthika et.al. (2011)[3] applied Clustering approach on processed web log data and finds strong association rule. Rahul & Abha (2012)[4] presents FP-growth model which uses Fp-tree data structure to obtain frequent pattern from Web log data. Latheefa & Rohini (2013)[5] developed a Custom-Built Apriori tool to discover interesting frequent access pattern in Web usage data. Alagesh & Ramaraj (2013)[6] discussed the usage of frequent user access pattern and analyzed the gap between existing technology and requirement. Meera & Firoz (2014)[7] proposed hybrid approach of FP-Growth algorithm and Decision Tree. FP-Growth algorithm is used to remove the unimportant information from the contents and Decision tree is used to fetch the contents from Web page. Aanum (2015)[8] analyzed two algorithms namely apriori and FP-growth to determine association rules that occur in the Web log dataset. Aarti et.al. (2015)[9] Clustering is used to find common behavioral users then Apriori algorithm is applied on clustered data to find access pattern.

3. USER AND SESSION IDENTIFICATION
User identification is the process of identifying each user accessing the WWW. Objective of identifying an user is to know the access characteristics so as personalised services are provided. Session is the collection of activities performed since an user have logged in till the instant the user is logged off. WUM techniques are employed to identify different user session from the web access log, as the amount of data to handle is huge with intricacy. This is a very complex process owing to the reason of presence of firewalls, cache and proxy server.

- There are incidents where different user have same IP address and same browser or same operating system, in such cases URI entries are taken into account for user identification.
- Time oriented approach, in which time between access request by an user is taken into consideration, if it surpass a limit of 30 minutes(default session timeout) or timestamp on two consecutive accessed pages of a user is more than 10 minutes, it implies a new user session have started.

4. MODIFIED APRIORI ALGORITHM
It searches for all frequent itemset using candidate generation from Web data. This algorithm follows a level-wise searching or breath first search using frequent Web pages. Those itemset above the average support are called frequent itemset.

Modified Apriori algorithm follows 2 phase:
• Generate Phase: In this phase candidate \((k+1)\)-itemset is generated using \(k\)-itemset, this phase creates \(C_{k+1}\) candidate set.

• Prune Phase: In this phase candidate set is pruned to generate large frequent itemset using “average support” as the pruning parameter. This phase creates \(L_k\) large itemset [10].

5. EXPERIMENTAL RESULTS

On completion of data cleaning in SQL developer, processed Web data is as shown in Fig. 1. Which is obtained by removing data containing various types of anomalies.

In Fig 2, user identification is done by using the navigational path for each user, different approach in user identification is done as proxy servers may assign same IP address to different users and these users may have similar user agent.

User identification approach result is as shown in Fig 3. Where the approach enables us to identify different user with similar IP address and same user agent, in the example below row (3-4-5) indicates a single user where the IP address and user agent were identical, row 8 and row 9 in spite of having similar IP address and User agents indicates two unique users. Furthermore in identifying different users, session of a user is also shown in Fig 3. Where the first user is having one session row (3-4-5), second user row 8 with one session and the third user row 9 also have one session.
For easy understanding, only the top three frequent pattern in each path whose support value is greater than the average support is taken into consideration and is as shown in Table 1. Which clearly elaborates that when the navigational path is increased the corresponding average support decreases.

**Table 1. Frequent Pattern Report**

<table>
<thead>
<tr>
<th>TOP THREE FREQUENT PATTERN IN EACH PATH</th>
<th>SUPPORT &gt; AVG SUPPORT</th>
</tr>
</thead>
<tbody>
<tr>
<td>FREQUENT MONO PATH</td>
<td></td>
</tr>
<tr>
<td>/docs</td>
<td>29729</td>
</tr>
<tr>
<td>/normal</td>
<td>25537</td>
</tr>
<tr>
<td>/login</td>
<td>20029</td>
</tr>
<tr>
<td>FREQUENT DI PATH</td>
<td></td>
</tr>
<tr>
<td>/normal -&gt; /docs</td>
<td>12971</td>
</tr>
<tr>
<td>/normal -&gt; /root</td>
<td>7644</td>
</tr>
<tr>
<td>/docs -&gt; /docs</td>
<td>6207</td>
</tr>
<tr>
<td>FREQUENT TRIA PATH</td>
<td></td>
</tr>
<tr>
<td>/normal -&gt; /docs -&gt; /docs</td>
<td>3169</td>
</tr>
<tr>
<td>/normal -&gt; /root -&gt; /root</td>
<td>2530</td>
</tr>
<tr>
<td>/journal -&gt; /includes -&gt; /content</td>
<td>2385</td>
</tr>
<tr>
<td>FREQUENT TETTARA PATH</td>
<td></td>
</tr>
<tr>
<td>/normal -&gt; /journal -&gt; /includes -&gt; /content</td>
<td>1450</td>
</tr>
<tr>
<td>/journal -&gt; /includes -&gt; /includes -&gt; /content</td>
<td>1190</td>
</tr>
<tr>
<td>/journal, -&gt; /includes -&gt; /content</td>
<td>1182</td>
</tr>
</tbody>
</table>

Comparison chart between average support approach and minimum support approach, helps in understanding that in average support approach when the pattern length increases the number of pairs keeps on increasing. Whereas in minimum support approach when the pattern length is increased the number of pairs decreases.
6. CONCLUSION

The concept of web usage mining is easy to identify user’s behaviour and interest, which helps web designer to optimized accessibility and usability of their websites. To need this idea researcher proposed Modified Apriori algorithm for discovering frequent pattern from Web log data by introducing average support value approach. This approach proves to be more efficient than already existing Apriori algorithm, as proposed algorithm generate frequent pattern with large pattern length, especially if large number of users exist in web log file. Proposed work can be implemented with more advance techniques/tools of web usage mining which will further assist in research field.

7. REFERENCES


Fig 4: Comparison of Minimum Support And Average Support