Literature Review on Web Mining

Bibu Skaria, Dr Eldhose T John and P.X. Shajan

Abstract--- Web is a platform for information exchange, as it is simple and easy to publish documents. Searching for information becomes a difficult and time-consuming process as the web grows. Web mining uses various data mining techniques to discover useful knowledge from usage log file from the web. The mining tools are used to scan the HTML documents, images, and text, the results is provided for the search engines. It can assist search engines in providing productive results of each search in order of their relevance. In this paper, we brief introduction to the concepts related to web mining and then an overview of different Web usage mining.

Keywords--- Apriori, FP Growth

I. INTRODUCTION

DATA MINING is an interdisciplinary subfield of computer science[8][9]. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Organizations have large amount of data that are growing rapidly they may be in different file formats and different databases. Data mining is applied in various domains like statistical data analysis, pattern recognition, database systems, information retrieval, WWW etc, has made tremendous growth during the last fifteen years[1][2]. Web mining is the application of data mining techniques to extract knowledge from web data, they may be from the web content, web structure or from the web usage data. In this paper we are reviewing the features of a few web usage mining algorithms.

II. WEB MINING TAXONOMY

Web mining is broadly classified into three distinct categories, depending on the kinds of data to be mined.

A. Web Content Mining

It is basically a process of extracting useful data or information from the contents in a web documents. Content data is the collection of facts embedded in a web page it may contain text content, images, audio, video etc. Application of text mining is widely researched area in web content.

B. Web Structure Mining

Is a tool used to identify the relationship between Web pages linked by information or direct link connection. This structure data is discoverable by the provision of web structure schema through database techniques for Web pages[10].

C. Web Usage Mining

Web Usage Mining is the application of Data Mining to discover and analyze patterns from click streams, user transactions and other logged user interactions with a website[4].

III. FP GROWTH ALGORITHM

The FP Growth algorithm operates in the following four modules.

- Preprocessing module
- FP Tree an FP Growth Module
- Association Rule Generation
- Results

Preprocessing Module

This module converts the ASCII format log file into a database like format, which can be processed by the FP Growth algorithm.

FP Tree an FP Growth Module

This module is performed in two steps.

- FP Tree generation
- Applying the FP Growth algorithm to generate association rules

FP tree is a compact data structure that stores important, crucial and quantitative information about frequent patterns.

Algorithm

Input: A database DB, represented by FP-tree constructed and a minimum support threshold.

Output: The complete set of frequent patterns.

Method: call FP-growth(FP-tree, null).

Procedure FP-growth(Tree, a) {
1) if Tree contains a single prefix path then // Mining single prefix-path FP-tree {
2) let P be the single prefix-path part of Tree;
3) let Q be the multipath part with the top branching node replaced by a null root;
4) for each combination (denoted as β) of the nodes in the path P do
5) generate pattern β ∪ a with support = minimum support of nodes in β;
6) ) let freq pattern set(P) be the set of patterns so generated;
7) else let Q be Tree;
8) for each item ai in Q do { // Mining multipath FP-tree
9) generate pattern β = ai∪ a with support = ai .support;

____________________________________________________________

Bibu Skaria, Dept of Computer Science, BPC College, Piravom, Kerala, India. E-mail:bibuvelyathu@gmail.com
Dr Eldhose T John, Dept of Computer Science, BPC College, Piravom, Kerala, India. E-mail:eldhosejohn@yahoo.com
P.X. Shajan, Dept of Computer Science, BPC College, Piravom, Kerala, India. E-mail:shajansx@gmail.com
DOI: 10.9756/BIJDM.8127

ISSN 2277 - 5048 | © 2016 Bonfring
10) construct β’s conditional pattern-base and then β’s conditional
FP-tree Tree β;
11) if Tree β ≠ Ø then
12) call FP-growth(Tree β , β);
13) let freq pattern set(Q) be the set of patterns so generated;
14) return(freq pattern set(P) ∪ freq pattern set(Q) ∪ (freq pattern set(P) × freq pattern set(Q)))

Advantages
1. It uses Compact data structure.
2. It eliminates repeated database scan.
3. It is faster than Apriori algorithm.
4. It reduces the total number of candidate item sets by producing a compressed version of the database in terms of an FP tree.

Disadvantages
1. It takes more time for recursive calls.
2. It is good only when user access paths are common.
3. It utilizes more memory
4. FP-Tree expensive to build

IV. A PRIORI ALGORITHM[6]

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases and confines large data sets during its initial database passes and uses this result as the base for discovering other large datasets during subsequent passes. It proceeds by identifying the frequently occurring individual items in the database and extending them to a larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database. Item sets having a support level above the minimum are called frequent item sets and those below are called small item sets. The algorithm is based on the large item set property which states: Any subset of a large item set is large and any subset of frequent item set must be frequent. The traditional algorithm for mining all frequent item sets and strong association rules was the AIS algorithm later renamed as Apriori.

Apriori algorithm uses Breadth-first search method and a Hash tree structure to count the number of candidate item sets efficiently. It generates the candidate item set of length k from item sets of length k-1. Then it trims the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets[6].

Apriori Algorithm
L1=find_frequent_1-itemsets(D);
for(k=2; Lk-1≠Φ; k++)
{ 
Ck=apriori_gen(Lk-1, min_sup);
for each transaction t@D
{
Ct=subset(Ck,t);
for each candidate c@Ct
c.count++;
}
Lk={c@cK | c.count≥min_sup }
}
Answer=UkLk;
Procedure apriori_gen(Lk-1:frequent(k-1)-itemsets)
for each itemset l1 @Lk-1
{
for each itemset l2 @ Lk-1
{ 
if(l1 [1]= l2 [1]) ∧ (l1 [2]= l2 [2]) ∧…∧(l1 [k-2]= l2 [k-2]) ∧(l1 [k-1]= l2 [k-1]) then

{c=l1 l2;
if infrequent_subset(c, Lk-1) then
delete c;
else add c to Ck;
}
}
return Ck;
Procédure infrequent_subset(c: candidate k-itemset; Lk-1:frequent(k-1)-itemsets)
for each(k-1)-subset s of c
{
if s @ Lk-1 then
return true;
}
return false;
where
D=database,
minsup=user defined minimum support

Apriori algorithm which is significant in web usage mining, suffers from a number of inefficiencies or trade-offs, which have been spawned from other algorithms.

Advantages
• It is very easy and simple algorithm.
• Its implementation is easy.

Disadvantages
• It does multiple scan over the database to generate candidate set.
• The number of database passes are equal to the max length of frequent item set.
Table 1: Features of Algorithm

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Features</th>
</tr>
</thead>
</table>
| FP-Growth | • It uses Compact data structure.  
            • Efficient and scaling method for mining the complete set of frequent pattern by pattern fragment growth, using extended FP Tree structure.  
            • It eliminates repeated database scan.  
            • It is faster than Apriori algorithm.  
            • Preprocessing modules convert the log file, which normally is in ASCII format, into a database like format. |
| Apriori   | • for mining frequent itemsets.  
            • Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently.  
            • Algorithm for mining frequent itemsets for Boolean association rules. |

V. CONCLUSION

Web usage mining is one of the prominent application of data mining, this techniques is used to discover usage patterns from Web data. FP Growth and Apriori algorithm meets the need of various web service providers and various viewers, users, business analysts, etc by providing the data they need. As a result it improves the techniques of Web Usage Mining by first discovering the log files of individual users at one place. This collective information therefore can be used to design business strategies so as to boom revenue. One of the main disadvantage of Apriori algorithm is that the candidate set generation is expensive process and that of FP-growth algorithm is the explosive quantity of lacks a good candidate generation method. In near future the algorithm can be extended to web content mining. In this paper authors have made an attempt to discuss these two algorithms, their advantages and disadvantages.

REFERENCE