Mining in Social Networks using Apriori Growth and FP Split Algorithm

Tamanna
PG Student
Department of Software Engineering
University Institute of Engineering and Technology
Kurukshetra University, 136119 Kurukshetra, INDIA

Dr. Sanjeev Dhawan
Assistant Professor
Department of Software Engineering
University Institute of Engineering and Technology
Kurukshetra University, 136119 Kurukshetra, INDIA

Abstract

Web Usage Mining the data mining techniques to extract exciting usage patterns from SNS data, in order to understand and serve the needs of users in an efficient manner. Usage data captures the origin and the identity of the user along with the browsing behavior at the social web site so that they can serve well. It is further classified depending on the type of data to be mine. This paper proposes an algorithm which takes advantages of both apriori growth algorithm and Fp split algorithm and shows the experimental results which shows the efficiency of the proposed algorithm.

Keywords: SNS, Frequent Patterns, Association mining, Apriori growth, FP-Split algorithm, Access Logs

I. INTRODUCTION

Web contains explosive data which is mostly unstructured. The amount of data kept in databases on the internet is increasing at a rapid pace due to which World Wide Web has become a dominant medium to retrieve information and mine knowledge. It increases the complexity of dealing with the information for different users, viewers, etc. The users want to have best searching tools to find information without any delay. The Web service providers want to reduce the load and designing complexity so that user can interact easily.

For on-line social networks analysis, the analysis targets are mainly focused on resources from the web, such as its content, structures and the user behaviors. Apriori algorithm for web usage mining is a best technique. Therefore, association rules can help to discover the hidden relationships between nodes in a social network or even cross networks. For e.g.: the person who read or liked X’s blog article and Y’s blog article.

- The site to be design (interface, content, structure, usability, etc.) is one of the most important aspects that wants to pay attention in the cyberspace.
- To understand the behavior of user to find the most frequently used sites and patterns so that they can be serve well.

II. RELATED WORK

Sharma and Bhartiya in their research paper shared that clustering, a method of grouping similar data into data sets known as clusters. A primary technique, cluster analysis, in clustering methods and conventional data analysis have been recognized which needs the number of clusters to be precise in advance and is dependent on the initial starting points. This paper presents a new algorithm to discover data clusters for numerical and nominal data. Apriori algorithm generates quite a large number of candidate set which is not efficient for both data. In another research which was carried out by Agrawal and Srikant considered the problem of discovering association rules between the items in a large database of sales transactions. It also shows how only the best features of the two proposed algorithms can be combined into a hybrid algorithm which they named as Apriori Hybrid. Experiments show that Apriori Hybrid scales linearly with the no. of transactions and it also has excellent scale-up properties w.r.t. the number of items in the database and the transaction size. Furthermore, Hsein Ting [3] investigated the issues around using web mining techniques for analysis of on-line social networks. Techniques and concepts of web mining and social networks analysis were introduced and reviewed along with a discussion about how to use web mining techniques for on-line social networks analysis. In addition this paper set out a process to use web mining for on-line social networks analysis, which could be treated as a general process in this research area. The authors described some initial results from ongoing work that was focused on extracting, and exploiting this structural information. We note that there is a lack of adequate data sets to fully test the new approaches. According to Suneetha and Krishnamoorthi, web usage mining is an application of data mining techniques which is put into use to discover the usage patterns from the web data so that it could serve the needs of the web based applications in a much better way. Instead of tracking the behavior of both interested or not interested users to redesign the web site to support users, the mining techniques such as association, clustering, and classification could be applied only on the group of interested regular users in order to find frequently accessed patterns which results in lesser time consumption and lesser memory utilization with higher accuracy and performance.
Ankit R Kharwar et al. in May 2011 in their research paper shared that WUM is an application of data mining technique to discover interesting usage patterns from web data carried out in order to understand and serve better, the needs of web-based applications. Web server data corresponds to the user logs that are collected at a web server. In order to produce the usage patterns and user behaviors, this paper implements the process of Web Usage Mining using basic association rules algorithm called Apriori algorithm.

Further Latheefa.V and Rohini.V presents the idea that mining web data in order to extract useful knowledge from it has become vital with the wide usage of the World Wide Web. In this paper they have proposed a custom-built Apriori algorithm for the discovery of frequent patterns in the web log data. This research also includes development of a tool to discover the frequent patterns, association rules in the web log data. The experiments that have been conducted in this study proved that custom-built Apriori algorithm is efficient than Classical Apriori algorithm as it takes lesser time.

Jun Yang et al. in 2013, proposed Featured Apriori Algorithm. In the database all the transactions were given equally importance. Traditional Apriori Algorithm was not designed for scanning a particular item. Due to mining all the association rules, it lead to redundancies.

### III. PROBLEM FORMULATED

On-line social network has become a very popular Web 2.0 application. This research work study the existing algorithms on web usage mining. For the analysis, resources from the web, such as its content, context and the behaviors of users were targeted. Using Apriori algorithm for web usage mining is a novel technique. Therefore, association rules can help to discover the hidden relationships between nodes in a social network or even cross networks. For e.g.: the person who read or liked A’s blog article and B’s blog article.

All this information is available online but is hidden from the users. Presently, there is no technique that can analyze this hidden information and this research work uses web usage mining (WUM) using FP-split tree and Apriori growth based approach for analyzing the browsing behavior of the users. Therefore, a proper mining should be there in order to have: (a) Enhance server performance, (b) Better user interaction, (c) Improve web site navigation, (d) Improve system design of web applications, and (e) Identify potential prime advertisement locations. Therefore, there is an urgent need to design and develop an efficient approach to analyze web usage behaviour to accommodate best possible results.

This Research work will concentrates on
- Web usage mining and in particular focuses on discovering the web usage patterns of SNS websites from the server database.
- The comparison of memory usage and time usage is compared using FP split tree and Apriori growth algorithm.

The previous works in web usage mining in Apriori algorithm and FP Tree mining algorithm had their disadvantages. Here we create a hybrid of FP-split tree and Apriori growth mining algorithm to take advantage of positives of both schemes. The Apriori algorithm performs repeated scans of the database while generating candidates while the FP tree mining algorithm is a time consuming, complicated algorithm. So we create a hybrid by combining FP Split tree for candidate generation and Apriori growth for mining.

### IV. PROPOSED METHODOLOGY

The algorithm which is proposed in this paper is the hybrid of the apriori growth algorithm and FP split algorithm and this work in two phases

In the first phase, candidate sets are created using fp split tree which is better than the fp tree algorithm. In the latter algorithm, two scans are involved which takes for time. In fp split tree, one scan is there. Moreover, it involves lesser pointer, as there is no link to its successor and predecessor. There is a header list which keep track of occurrence of the items in the final tree.

In the second phase apriori growth algorithm is used for mining as it is best and efficient algorithm. So, our algorithm takes advantages of both these algorithm in order to reduce the time of mining. There is no repeatedly recursive tree is constructed.

<table>
<thead>
<tr>
<th>Content</th>
<th>Count</th>
<th>Link_Sibling</th>
</tr>
</thead>
<tbody>
<tr>
<td>List</td>
<td></td>
<td>Link_Child</td>
</tr>
</tbody>
</table>

Node Structure for FP Split Tree

**A. Phase 1: construction of tree using Fp-Split Tree Algorithm:**
1) Step-1: equivalence class of item is created by scanning the database. Let the equivalence class of item be EC, \( \text{Tid} = (\text{Tid I Tg} \text{are the identifier of transaction ti; 1 is an item of ti).} \)
2) Step-2: Support factor is calculated in order to filter out non-frequent items. The support of each item I- refers to the number of records contained in the equivalence class EC,. Let the support of the equivalence class EC is denoted by pCLI. Then the items below minimum support factor are removed.
3) Step-3: From the previous step, frequent item sets are generated. Then equivalence class of item is converted into nodes which help in building fp split tree. Header table is maintained which helps in tree traversal.
4) Step-4: While constructing the FP-split tree, which is a dummy root node will be generated.
5) Step-5: Four rules for the constructing of FP tree, where p stands for a specific node in the FP tree.
   Rule 1:
   If ( p is root and p.Link-child = null ) Then
   p.link-child n
   Else
   Call Compare (p.Link-childList, n.list )
   End if
   Rule 2:
   If ( n.list c p.list and p.Link-child = null ) Then
   p.Link-child n
   else
   Call Compare (p.Link-childList, n.List )
   End if;
   Rule 3:
   If ( n.List n.p.List = 0 and p.Link-siblings null) Then
   p.Link-sibling n
   else
   Call Compare (p.Link-childList, n.LW )
   End if;
   Rule 4:
   If (p.List n.p.List # 0 and p.List n.List<>0 ) Then
   Call split ( n ) and return two nodes n1 and n2
   End if

B. Phase 2: Tree Mining using Apriori Growth Algorithm
Candidate set Algorithm on which Apriori Growth implemented.
1) Step 1: List1=k-1 frequent item dataset from Data
2) Step 2: N=size(list1)
Initialize mylist as a blank list to contain generated frequent item dataset
3) Step 3: Repeat for i=1 to n-1
4) Step 4: Repeat for j=i+1 to n
5) Step 5: l1=list1[i]
6) Step 6: l2=list1[j]
7) Step 7: Remove the last elements from l1 and l2
8) Step 8: if l2 is a subset of l1 then
   Flist=append last element of l2 at end of l1
   [end of while]
9) Step 9: If count(flist)>=supp then
   Add flist to mylist
   Else
   return null
10) Step 10: end
New Apriori Growth Algorithm
- Input: pagelist2: list of pages with equivalence class satisfying support count
- Tree: Nodes of tree as a list.
- Sup: minimum support
- Output: frequent item sets ‘data’
The algorithm is implemented with the following steps:
1) Step 1: Repeat following step while scanning pagelist2 till end
   - Create list containing single item from pagelist2
   - Add this list to mylist1
   [End of repeat]
   [where mylist1 is 1 frequent itemset]
2) Step 2: Add mylist1 to data
3) Step 3: Repeat for k=2,3,4,….
4) Step 4: C_k=getCandidate(k)
5) Step 5: If |C_k|=0 then
Goto step 6
Else
Add \( C_k \) to data
End if

6) Step 6: End

V. RESULTS ANALYSIS

Figure 5.1 shows the screenshot where logs are transferred to the database and cleaning is done to remove unnecessary data which is not required for our mining process called filtering the log.

Figure 5.2 shows the successful transmission of log entries into the database. Click "OK".

Figure 5.3:
The above graph shows the comparison between FP Tree and FP Split tree methods for different values of support.

A. Frequent sets generated for threshold value 2
Mining the FP split tree created
Added 1 frequent itemset
[[0, 39], [1, 39], [1, 49], [2, 39], [7, 39], [9, 39], [12, 39], [14, 39], [14, 49], [23, 39], [29, 39], [29, 49], [30, 39], [39, 48], [39, 49], [39, 55], [39, 65], [39, 70], [39, 72], [41, 49], [48, 49], [49, 65], [49, 72]]

Added 2 frequent itemset
[[0, 39], [0, 49], [1, 39], [1, 49], [2, 39], [7, 39], [9, 39], [12, 39], [14, 39], [14, 49], [23, 39], [29, 39], [29, 49], [30, 39], [39, 48], [39, 49], [39, 55], [39, 65], [39, 70], [39, 72], [41, 49], [48, 49], [49, 65], [49, 72]]

Added 3 frequent itemset
[[0, 39], [0, 49], [1, 39], [2, 39], [7, 39], [8, 39], [9, 39], [10, 39], [11, 39], [12, 39], [13, 39], [14, 39], [15, 39], [16, 39], [17, 39], [18, 39], [19, 39], [20, 39], [21, 39], [22, 39], [23, 39], [24, 39], [25, 39], [26, 39], [27, 39], [28, 39], [29, 39], [30, 39], [31, 39], [32, 39], [33, 39], [34, 39], [35, 39], [36, 39], [37, 39], [38, 39], [39, 39], [40, 39], [41, 39], [42, 39], [43, 39], [44, 39], [45, 39], [46, 39], [47, 39], [48, 39], [49, 39], [50, 39], [51, 39], [52, 39], [53, 39], [54, 39], [55, 39], [56, 39], [57, 39], [58, 39], [59, 39], [60, 39], [61, 39], [62, 39], [63, 39], [64, 39], [65, 39], [66, 39], [67, 39], [68, 39], [69, 39], [70, 39], [71, 39], [72, 39]]

VI. CONCLUSION

Social network brought our world closer. The sites need continuous improvement so that they give better services. This research work on two basic algorithms i.e, FP Split Algorithm[18] and Apriori growth algorithm[20]. FP Split technique is far better than its predecessor FP Tree algorithm as it reduces complexity. FP Split Algorithm split the node and form two branches inspite of fptree which recursively create subtrees. The Apriori growth algorithm is used for mining the FP Split tree as it efficiently mine the sets created. The graphs plotted above confirmed the effectiveness of the proposed technique against the previous algorithms.

This algorithm can be used for effectively mine SNS data, web logs or WSN data.

ACKNOWLEDGMENT

I owe this work done by me to my guide Mr. Dr Sanjeev Dhawan, without whose support and guidance, this work would never have been possible. I also thank my parents and family for their support and patience.

REFERENCES


