



Association Mining in Social Networks

¹Sanjeev Dhawan, ²Tamanna

¹Faculty of Computer Science & Engineering, ²Student of M.Tech. (Software Engineering)

^{1,2}University Institute of Engineering and Technology,
Kurukshetra University, Kurukshetra, Haryana, INDIA.

Abstract: WWW opened flood gate to many new concept like emails, net meetings, online advertisement and social media. Social media has revolutionarized the world by bringing lacks of people on a single platform and sharing their contents. SNS sites like facebook, twitter, google+ etc have millions of people and billions of articles posted by these people on the social media. The performance of these sites is a critical issue in the cut throat competition. There novace association mining technique normally used in mining SNS data. We can create frequent item sets of articles which are shared amongst different set of people. The common technologies in association mining are apriori and fp tree algorithm. Different modifications of these algorithms have been proposed by different researchers each having its own merits. This research proposes technique which is hybrid of apriori and fp split tree. Fp split tree performs better than fp tree avoiding recursive creation of subtree. Apriori growth has efficient mining in comparison to fp tree or fp split tree. So, we propose itemsets to be fp split tree. So, we propose itemsets to be generated and stored in fp split tree and then uses apriori growth for mining the experimental result confirm the effectiveness of proposed technique against traditional techniques.

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

I. Introduction

Web is a vast, explosive, diverse, dynamic and mostly unstructured data repository, which supplies an incredible amount of information, and also raises the complexity of how to deal with the information from different perspectives of users, view, web service providers, business analysts.[2] The users want to have effective search tools to find relevant information easily as well as precisely. The Web service providers want to find the best way to predict the users' behaviors and personalize information to reduce the traffic load and design the Web site that is suited for different group of users.

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. 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.

- The design of the whole site (interface, content, structure, usability, etc.) is one of the most important aspects for an institution that wants to survive in the cyberspace.
- Understand the way user browses the site and find out which is the most frequently used link and pattern of using the features available in the site.

II. Related work

Recently, A.Gupta, *et al.* in 2014 described that Web usage mining focused and analyzed the useful patterns on websites. It was the technique to finding the facts and figures to better serve the need of web based applications. It was classified into three categories are web server, server and application level data. Apriori Algorithm generated the association rules to work on databases that contain lots of transactions. With the use of Apriori algorithm and Improved frequent pattern tree algorithm gave comparison of memory usage and speed of producing association rules. Drawback: In Apriori Algorithm, candidate generation sets large no. of usage patterns before each scan and it made costly.FP Tree Algorithm lacked a better candidate generation method. Jun Yang *et al.* in 2013, proposed an improved algorithm named Featured Apriori Algorithm. In the database all the transactions were equally important. Traditional Apriori Algorithm was not designed for looking a particular item. Due to mining all the association rules, it generated many redundancies. In this paper an improved traditional based algorithm was proposed. Every Transaction items had its own features, during mining the association rules of the same featured items would be scanned and computed. After analysis in a book recommendation system, it took less time and got more reasonable association rules. Lastly, Shipra Khare *et al.* in 2013 in their research stated that web usage mining is the application of data mining techniques to discover interesting usage patterns from web data, to better serve the needs of Web-based applications. This paper implemented the process of Web Usage Mining using basic association rules algorithm called as Apriori Algorithm. Web usage mining architecture divided process in two main parts- the first part included preprocessing, data integration components and transaction identification. Second part included the largely domain independent application of generic data mining and pattern matching. It also contains an efficient improved iterative FP Tree algorithm for generating frequent

patterns. This piece of research work concentrated on web usage mining and in particular focused on discovering the web usage patterns of websites from the server log files. The comparison of time and memory usage was compared using Apriori algorithm and Frequent Pattern Growth algorithm. Drawback: The work could be extended for developing a web usage mining tool with customized web log preprocessing and combined pattern analysis approaches according to different application.

Latheefa.V *et al.* in 2013 in their research suggested the idea that mining web data in order to extract useful knowledge from it had become vital with the wide usage of the World Wide Web. In this paper they proposed a custom-built Apriori Algorithm for the discovery of frequent patterns in the web log data. This research also included development of a tool to discover the frequent patterns, association rules in the web log data. The experiments that had been conducted in this study proved that custom-built Apriori Algo was efficient than Classical Apriori Algo as it involved lesser time. Drawback: This analysis could be extended by implementing other interesting measures such as lift in order to fine tune the results. The proposed work could also be implemented by the adoption of more advanced techniques/tools such as fuzzy associations using the interesting measures such as fuzzy support, fuzzy confidence, fuzzy interest and fuzzy conviction etc. Chin Feng Lee *et al.* in 2005 proposed a fast algorithm called frequent pattern split, simply FP-split, for improving the process of the FP-tree construction. The proposed FP-split algorithm contained two main steps. The first step was to scan transaction database only once for generating equivalence classes of frequent items. The second step was to sort these equivalence classes of frequent items in descending order so as to construct the FP-split tree. Through detailed experimental evaluations under various system conditions, this method showed excellent performance in terms of execution efficiency and scalability. The FP-split algorithm was superior to FP-tree construction algorithm. There were three reasons to support that the proposed method outperformed FP-tree construction algorithm in terms of tree construction. The first one was that this method scanned the database only once. The second one was that filtering out and sorting the items in each transaction record was no longer employed in this method. The third one was that the header table and links were not repeatedly searched, while adding a new node in the FP-split tree. Drawback: It needs to be optimized for counting the support of the candidates and expanded for mining larger database.

This section summarizes the invaluable researches done by different researchers in the field of frequent pattern mining and association mining. The algorithms in these works mainly use or modify the apriori and frequent pattern growth algorithm for the purpose. The FP Split tree [32] explains a novel technique which improves the performance of FP Tree which has long been considered the best prevailing technique for frequent pattern mining. The different modifications of apriori suggested by different authors highlight the shortfalls in these basic algorithms..

III. Problem formulated

On-line social networking has become a very popular Web 2.0 application. The present research work pointed out the issues around using web mining techniques for analysis of on-line social networks. This research work was initiated through a system study and analysis phase, where significant study was conducted to understand the existing algorithms on web usage mining. 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. 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 proposed algorithm which is a hybrid of a modified FP Tree creation algorithm and Apriori Growth mining algorithm, is given below in two phases.

The first phase constructs the FP Split Tree which is more efficient way to create candidate sets than FP Tree since the latter involves two complete scans of the database while the former does it once. This impacts the efficiency almost 2 times better. More over the FP Split Tree created by our algorithm involves lesser use of pointers as we don't link each node in the Tree to its predecessor and successor. Rather we maintain a header list separately which maintains a list separately for each of the pages which points to the occurrence of these items in the final tree created.

The second phase involves mining the FP Split tree created using the Apriori growth algorithm. This algorithm is more efficient than FP Growth as it does not involve recreating the FP Split trees repeatedly every time in recursion as in FP Growth algorithm thereby reducing the time involved.

Content	Count	Link_Sibling
List		
Link_Child		

Node Structure for FP Split Tree

Phase 1:- Tree construction using Fp-Split Tree Algorithm.

Step-1. Scanning the database to create equivalence class of item. Let the equivalence class of item be EC, = Tid I Tg are the identifier of transaction ti; 1 is an item of ti).

Step-2. Calculating support 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 pCLI denote the support of the equivalence class EC,. After calculating the supports of items, we delete the items whose supports are below the predefined minimum support

Step-3. After generating frequent items, the equivalence class of item is then converted into nodes for the construction of FP-split tree. To facility tree traversal, a header table is built in advanced so that each item can point to its first occurrence in the FP-split tree.

Step-4. While constructing the FP-split tree, a root will be generated, which is a dummy node.

Step-5. There are the 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 $\square$ 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 $\square$ n

else

Call Compare (p.link-child.List, n.Lisf)

End if;

Rule 3:

If (n.List n p.List = 0 and p.Link-siblings null) Then

p.Link-sibling $\square$ n

else

Call Compare (p.Link-child.List, n.LW)

End if;

Rule 4:

If (p.List n n.List $\neq$ 0 and p.List - n.List $<>$ 0) Then

Call split (n) and return two nodes n1 and n2

End if

Phase 2:- Tree Mining using Apriori Growth Algorithm.

Nodes used in the algorithm

Name
*Tablelink
Count

Data Structure of Node of Header Table

Name
*TableLink
*parent
*Child
Count

Data Structure of Node of Fp-split tree

Candidate set Algorithm on which Apriori Growth implemented.

```

Step 1.      List1=k-1 frequent item dataset from Data
Step 2.      N=size(list1)
              Initialize mylist as a blank list to contain generated frequent item dataset
Step 3.      Repeat for I=1 to n-1
Step 4.      Repeat for j=I+1 to n
Step 5.      l1=list1[I]
Step 6.      l2=list1[j]
Step 7.      Remove the last elements from l1 and l2
Step 8.      if l2 is a subset of l1 then
              Flist=append last element of l2 at end of l1
[End of while]
Step 9.      If count(flist)>=supp then
              Add flist to mylist
Else
              return null
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:-

```

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]
Step 2.      Add mylist1 to data
Step 3.      Repeat for k=2,3,4,...
Step 4.      Ck=getCandidate(k)
Step 5.      If [Ck]=0 then
              Goto step 6
            Else
              Add Ck to data
            End if
Step 6.      End
    
```

V. Results Analysis

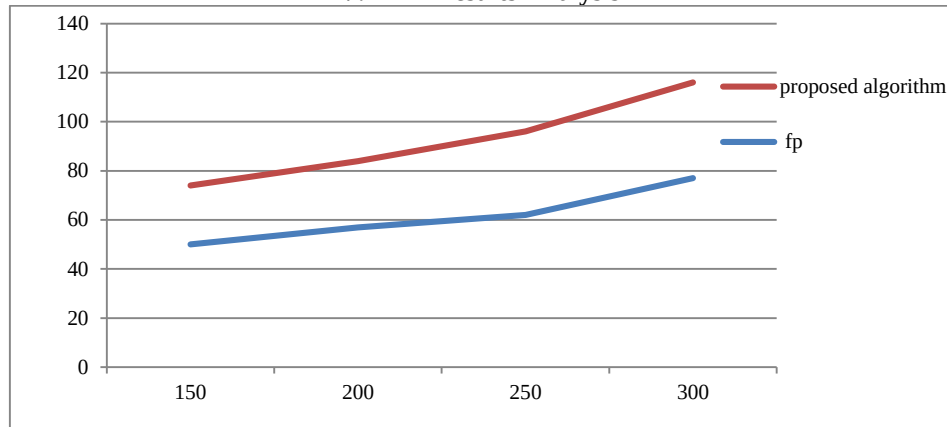


Figure 1

Table of records used for figure 1

No. of records	FP tree time	Proposed Algorithm Time
127	46	29
116	43	21
112	38	20
101	34	18
92	33	16

The above graph shows the comparison between FP Tree and FP Split tree methods for support count 3. We can see that the time taken by proposed algorithm is always better than the traditional method. The difference would be more clearly visible if we get logs of few days time for some SNS website where the incoming traffic is also more frequent. The proposed algorithm has been tested on logs received from alexa.com website for only 12 minutes. The more is the no. of records, the better visible is the difference between the efficiency of the two algorithms.

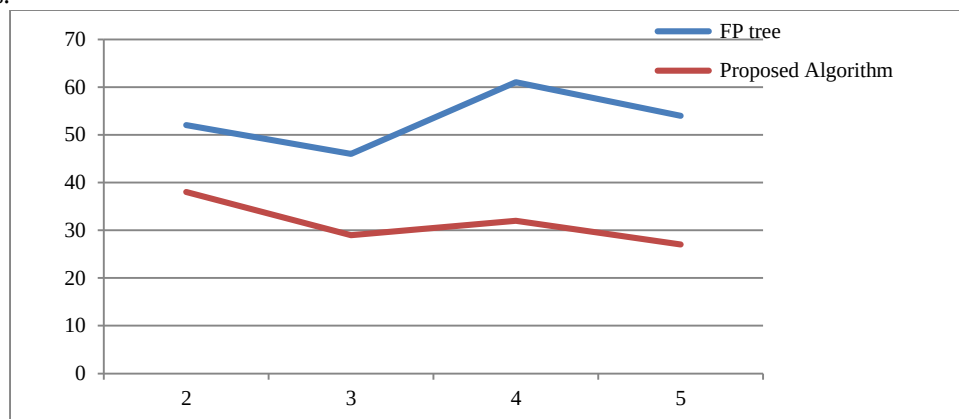


Figure 2

The above graph shows the comparison between FP Tree and FP Split tree methods for different support counts for a fixed no. of records.

Table of records used for figure 2

. Support	FP tree time	Proposed Algorithm Time
2	52	38
3	46	29
4	61	32
5	54	27

Frequent sets generated for threshold value 2:

Mining the FP split tree created

Added 1 frequent itemset

[[0], [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73]]

Added 2 frequent itemset

[[0, 39], [0, 49], [1, 39], [1, 49], [2, 39], [7, 39], [9, 39], [12, 39], [12, 49], [13, 39], [14, 39], [23, 39], [29, 39], [29, 49], [30, 39], [39, 41], [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, 49], [1, 39, 49],

VI. Conclusion

The most happening place these days is the social network platform which has brought the world closer and better livable. These sites are in a continuous pursuit to improve their services. This research performs association mining on a social network data using two of the most efficient techniques discussed over the years in Web usage mining – FP Split Algorithm[18] and Apriori growth algorithm[20]. FP Split technique performs better than its predecessor FP Tree algorithm in the way that it has reduced complexity. The recursive creation of subtrees has been replaced by splitting the node into two to create two branches in the tree. The Apriori growth has been used for mining the FP Split tree as it has better efficiency for mining the data and generating the frequent sets. This research generated a hybrid of these two algorithms by creating the candidate sets and storing them in the FP Split tree and then used Apriori growth to mine the FP Split tree. The experimental results confirm the effectiveness of the proposed technique against the traditional algorithms. The proposed method can be in future applied to different types of data like web logs or WSN data. Another direction to work with can be to reduce the complexity further by reducing the steps involved in creating the FP Split tree.

VII. References

- [1]. Goswami D.N., ChaturvediAnshu, Raghuvanshi C.S., "An Algorithm for Frequent Pattern Mining Based On Apriori", (IJCSSE) International Journal on Computer Science and Engineering, Vol. 02(4), 2010, pp. 942-947.
- [2]. MajaDimitrijevic, ZitaBosnjak, "Discovering Interesting Association Rules in the Web Log Usage Data", Interdisciplinary Journal of Information, Knowledge, and Management Volume 5, 2010, pp. 191-207.
- [3]. Ford LumbanGaul, "Exploring the Pattern of Habits of Users Using Web Log Sequential Pattern", Second International Conference on Advances in Computing, Control, and Telecommunication Technologies, 2010.
- [4]. Ankit R Kharwar, Viral Kapadia, NileshPrajapati, Premal Patel, "Implementing APRIORI Algorithm on Web server log", National Conference on Recent Trends in Engineering & Technology, 2011.
- [5]. Suneetha K R, Krishnamoorti R, "Web Log Mining using Improved Version of Apriori Algorithm", International Journal of Computer Applications, Volume 29 (6), September 2011, pp. 23-27.
- [6]. R. Suguna, D. Sharmila, "An Overview of Web Usage Mining", International Journal of Computer Applications, Volume 39 (13), February 2012.
- [7]. Sudhamathy and JothiVenkateswaran, "Hierarchical frequent pattern analysis of web logs for efficient interestingness prediction", Indian Journal of Education Info Management, Vol. 1 (5), May 2012, pp. 29-35.
- [8]. Mr. Rahul Mishra, Ms. AbhaChoubey, "Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2(9), September 2012.
- [9]. T. Sunil Kumar, Dr. K. Suvarchala, "A Study: Web Data Mining Challenges and Application for Information Extraction", IOSR Journal of Computer Engineering (IOSRJCE), Vol. 7(3), 2012, pp 24-29.
- [10]. BinaKotiyal, Ankit Kumar, Bhaskar Pant, R.H. Goudar, ShivaliChauhan and SonamJune, "User Behavior Analysis in Web Log through Comparative Study of Eclat and Apriori", Proceedings of 7th International Conference on Intelligent Systems and Control 2013.
- [11]. Kirti S. Patil, Sandip S. Patil, "Sequential Pattern Mining Using Apriori Algorithm & Frequent Pattern Tree Algorithm", IOSR Journal of Engineering (IOSRJEN), Vol. 3, Issue 1, Jan. 2013, pp. 26-30.
- [12]. Latheefa.V, Rohini.V, "Web Mining Patterns Discovery and Analysis Using Custom-Built Apriori Algorithm", International Journal of Engineering Inventions Volume 2, Issue 5 (March 2013), pp. 16-21.
- [13]. Harendra Singh, Ashish Kumar Srivastava, SitendraTamrakar, "A Modified FP-Tree Algorithm for Generating Frequent Access Patterns", JECET June-August 2013. Vol.2.No.3, pp. 730-740.
- [14]. ShipraKhare, Vivek Jain, ManojRamaiya, "Implementation of Web Usage Mining with Customized Web Log Using FP Growth Algorithms", International Journal of Engineering & Managerial Innovations (IJEMI), Volume I (II), September 2013, pp. 1642-1647 .
- [15]. Manish Gupta, Rui Li and Kevin Chen Chuan Chang, "Towards a Social Media analytics Platform: Event Detection and User Profiling for Twitter", ACM WWW'2014, pp. 193-194.
- [16]. Ruben P. Albuquerque, Jonice Oliveira, Fabrício F. Faria, Rafael Monclar and Jano M. de Souza, "Studying Group Dynamics through Social Networks Analysis in a Medical Community", Journal of Social Networking, 2014, Vol. 3, pp. 134-141.
- [17]. SanjeevPippal, LakshayBatra, Akhila Krishna, Hina Gupta and KunalArora, "Data mining in social networking sites: A social media mining approach to generate effective business strategies", International Journal of Innovations & Advancement in Computer Science (IJACS), Volume 3(2), April 2014.
- [18]. Chin Fewng Lee and Tsung-HsienShen, "An FP-split method for fast association rules mining", IEEE 2005, pp.459-464.
- [19]. Omer Adel Nasser & Dr. Nedhal A.AL Saiyd, The Integrating Between Web Usage Mining and Data Mining Techniques, 5th Conference on CSIT, IEEE(2013).
- [20]. Jun Yang, Z. Li, Wei Xiang, An Improved Apriori Algorithm Based on Features, IEEE (2013).
- [21]. A.Gupta, R.Arora, R.Sikarwar and N.Saxena, Web Usage Mining Using Improved Frequent Pattern Tree Algorithm, IEEE(2014).