Volume 6, Issue 1, January 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Analyzing Apriori and FP-Growth Algorithm on an Arabic Corpus

Ibrar Afzal[1], Dr. Arif Iqbal Umar[2], Dr. Kifayat Ullah[3], Dr. Ali Imran Jehangiri[4]
[1, 2, 4] IT Department Hazara University, Mansehra, KPK, Pakistan
[3] Department of Computer Science and Software Technology, University of Swat, KPK, Pakistan

Abstract— In this paper an analysis is made to find frequent patterns by using Apriori and FP-Growth algorithms for Arabic corpus .First of all frequent item sets are retrieved and then Frequent patterns will be analysed using Quran as a case study. In our experiment for Arabic corpus, FP-Growth shows better performance when compare with Apriori algorithm.

Keywords— Itemset, Frequent patterns, Corpus, Apriori algorithm and FP-Growth algorithm.

I. INTRODUCTION

Unstructured data can be transformed into structure forms by use of data mining for useful information. Market basket analysis is a well-known application and provides a benchmark for data mining for finding frequent patterns. In market basket analysis different items relationships are determined which are associated with each other through association rules. Apriori Algorithm was proposed by R.Agrawal and R. Srikant in 1994 for mining the frequent itemsets, which compute the frequent patterns through candidate generation.

The main problem for the users regarding Quran is that frequent patterns are hard to find out. Single word can be easily accessed but the verses from the Quranic chapters are difficult and almost impossible for comparison.

Considering the frequent pattern analysis, large number of efforts ,for the word search in an Arabic text corpus have been made [13]-[15]. The main difference between the already available efforts and our approach is that we use FP-Growth algorithm and Apriori algorithm and provides a comparison in Arabic text for the research community. This paper provides a difference between performance of these two algorithms for Arabic Corpus with the use of different parameters like support, lift and confidence etc. By defining the support and confidence one can easily understand the association between the verses of Quranic corpus .Quran is selected for our work due to its importance for the universe. It is equally important for Muslims and Non-Muslims scholars to explore different ideas and civilization.

The remaining of the paper is organized as follows. Section II introduces related work, Section III provides a brief description of Apriori algorithm and FP-growth algorithm .Section IV presents implementation for the discussed algorithms. Section V discusses the Results and Conclusion and future work is presented in Section VI.

II. LITERATURE SURVEY

We have selected the Quran as a case study for our work. Research literature on the application of computing to Quranic data sets includes a variety of attempts. First time Meccan and Medinan chapters from Quran were demonstrated through clustering in [9].Apriori for frequent subpaths (AFS)[11] which was based on the graph structure and discussed the frequent subpath. Support was considered a key factor for frequent subpath finding. Latter on this AFS algorithm was implemented in Quranic corpus to generate frequent subpaths [15]. "AFS used bottom-up search for frequent subpaths".The proposed AFS algorithm was not suitable for larger corpuses to extract the frequent patterns. QurAna [14] is a large corpus created from original Quranic text. For information retrieval tasks antecedents are helpful and are provided by QurAna. Cosine angle is used as a distance between query and a document in QurAna. Distance between the pronouns and its antecedent is too far, this is a challenge for this corpus. QurSim [13] linked the semantically similar or related verses together. It covers similarity and relatedness in short texts. There are a lot of challenges that are faced by QurSim. Each verse of the Quran is considered as a separate document in QurSim increasing the time complexity.

III. OUR APPROACH

Before discussing the methods some introduction of Apriori algorithm and FP-Growth algorithm is being presented for understanding and selections of the best and efficient algorithm.

A. Apriori Algorithm Overview

Apriori Algorithm was proposed by R.Agrawal and R. Srikant in 1994 for mining the frequent itemsets. This algorithm used the Apriori property. This property of algorithm is relevant to the subset and super set of the frequent items found during processing. If any subset of items are frequent then pattern found must be frequent as well[10]. When considering the famous market basket analysis if {beer, chips, nuts} is frequent, so is {beer, chips}, i.e., every transaction having {beer, chips, nuts} also contains {beer, chips}. Once the frequent items are found then pruning is required because too many items are searched for comparison[16]. Apriori pruning principle states that if there is any pattern which is

infrequent, its superset should not be generated and tested. Support and confidence are two metrics on the basis of which frequency of frequent patterns retrieved increases or decreases.

In our case when Arabic corpus is considered support was 0.03 and confidence was 0.9.Its mean that when frequent patterns are retrieved they must satisfied the metric criteria. In pruning process only maximal frequent were retrieved. We have used Weka 3.0 and RapidMiner 5.3 for frequent patterns. In RapidMiner, after successful installation, we have to install the Weka extension for Apriori algorithm. There are numbers of parameter to retrieve the frequent pattern in RapidMiner for Apriori algorithm.

Some of the parameter are discussed below:

- N: It retrieved required number of rules for users.
- T: Association rules are ranked through this metric type.
- C: This metric defines the minimum confidence of a rule.
- D: This metric decrease the value of minimum support in each iteration.
- I: To output the values of itemsets found.
- R: When columns have missing values, those columns must be removed through this metric

When Apriori algorithm is provided the data set in Arabic format, it is observed that Apriori is not suitable for large dataset, also demonstrated by us in the experiment section of this paper. When huge candidate sets are retrieved, it becomes bottleneck in Apriori.

B. FP-Growth Algorithm

Basically no candidate generation concept was introduced in FP-Growth algorithm.As a result time complexity is decrease and graphical representation for the association rules are also mentioned.

We have used RapidMiner in our work for the differentiation of time complexity between Apriori and FP-Growth algorithm. FP-Growth algorithm uses divide-and-conquer methodology for breakdown of large mining tasks. "FP-growth algorithm is a two-step process" .In the initial step a compact data structure is used by FP-Growth algorithm to encodes the data set. It also avoids costly database scans. Frequent itemsets are directly extracted from FP-tree. In the second phase, from frequent itemset association rules are mined. The numbers of association rules are dependent on the frequent itemset generation in the initial phase by executing the process in RapidMiner for Arabic text as shown in Figure 1.

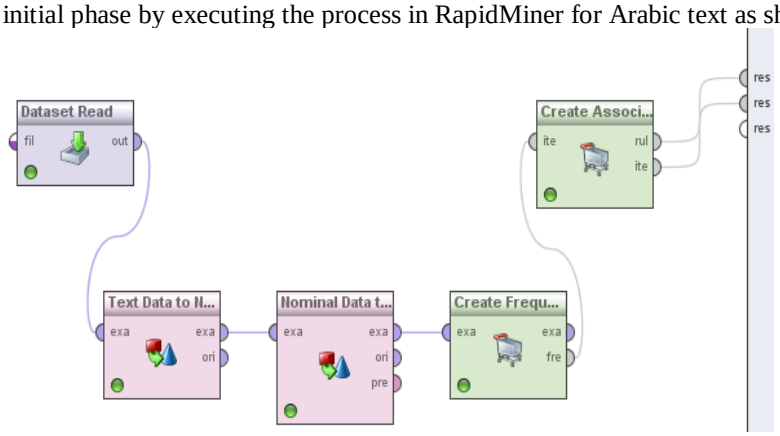

Fig. 1 Process for Frequent Itemset Generation

User can restrict the number of association rule through use of support and confidence metric as shown in Figure 3.Only 19 association rules are generated when confidence was set t0 0.9 and support was 0.05. In step two tree is built using the counter as shown in Figure 2.

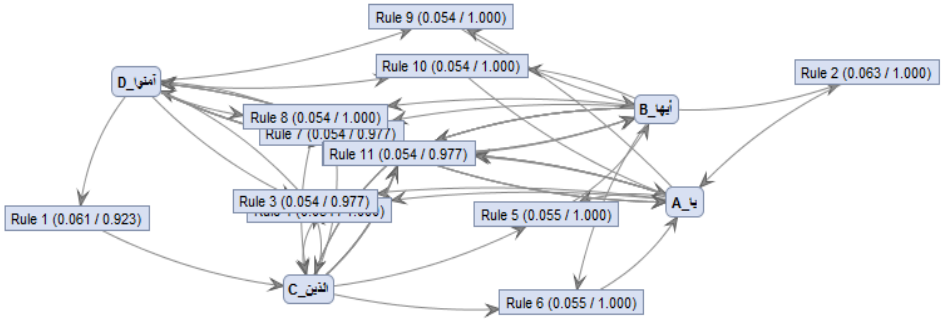

Fig. 2 Process for Frequent Itemset Generation

Counter is incremented when same prefix is appeared. If multiple transactions share an identical frequent item set, they can be merged into one with the number of occurrences registered as count. Graphs can be described(as shown in Figure 2) with the help of association rules in which support and confidence are main factors and demonstrated in the experiment section of this paper.

No. of Sets: 19
Total Max. Size: 4
Min. Size: 1
Max. Size: 4
Contains Item:
Update View

Size	Support	Item 1	Item 2	Item 3	Item 4
1	0.082	C_الذين			
1	0.077	A_يا			
1	0.072	B_الذين			
1	0.067	D_آمنوا			
1	0.065	A_إن			
1	0.063	B_أيها			
1	0.059	C_الله			
1	0.056	C_من			
2	0.055	C_الذين	A_يا		
2	0.061	C_الذين	D_آمنوا		
2	0.055	C_الذين	B_أيها		
2	0.054	A_يا	D_آمنوا		
2	0.063	A_يا	B_أيها		
2	0.054	D_آمنوا	B_أيها		
3	0.054	C_الذين	A_يا	D_آمنوا	
3	0.055	C_الذين	A_يا	B_أيها	
3	0.054	C_الذين	D_آمنوا	B_أيها	
3	0.054	A_يا	D_آمنوا	B_أيها	
4	0.054	C_الذين	A_يا	D_آمنوا	B_أيها

Fig. 3 Frequent itemset generation based on support.

The parameter for FP-Growth which are used in RapidMiner process, are too much. Some are defined for clarity are discussed below:

- max_number_of_retries: This parameter determines how many times the operator should lower the minimal support to find the minimal number of item sets.
- min_support: The minimum support criteria is specified by this parameter
- max_items: This parameter specifies the upper bound for the length of the itemsets i.e. the maximum number of items in an itemset.

Running the process for FP-Growth with given support=0.05 and confidence=0.9 in RapidMiner gives us the result with 19 frequent itemsets as shown in the Figure 2.From these frequent itemsets rules are formed and designed to produce useful graphs though FP-Growth algorithm.

TABLE I COMPARISON OF APRIORI AND FP-GROWTH ALGORITHM

Support	**Confidence**	**Apriori Run time in Seconds**	**FP-Growth Run time in Seconds**
0.01	0.8	16	2.5
0.02	0.8	15	3
0.03	0.8	15	2.5
0.01	0.9	14	3.5
0.02	0.9	14	3.5
0.03	0.9	14	2.5

IV. IMPLEMENTATION

To implement Apriori and FP-Growth algorithm, Weka 3.0 and RapidMiner 5.3 is used. Support and confidence were the two main parameters for testing the test-bed. Two different test-bed were used for the comparison of the algorithms.

V. RESULTS

Our proposed method use FP-Growth algorithm and Apriori algorithm .In this paper we are just analyzing the performance through time complexity of the two proposed algorithms.

Initially when excel file is read by RapidMiner and minimum confidence was set to 0.001 and support was 0.5 then 529 frequent itemset was found. With that large set of frequent item association rules retrieved was above 7500.So it was too much difficult to analyze the result. When the minimum confidence was increased to 0.05 and support to 0.7 then above 200 frequent item sets were retrieved. For better understanding confidence was set to 0.9 and support 0.7 then we has 45 frequent itemset and 18 rules for association. In Apriori algorithm generates large item set in which number of frequent itemset are shown ,also demonstrated in the Figure 2.

Apriori algorithm by changing the minimum confidence, there is effect of “number of cycles performed”, in most cases value is 20 but when confidence is increased then “number of cycles performed “ also decreases. It is interesting that size of large item found never changes. Regarding to time comparison for Apriori and FP-Growth ,FP-Growth is found to be best as shown in table 1. Frequent pattern extraction from the text corpus from the graphic point of view is too much beneficial due to use of Rule number in the graphs.

We have observed that association rules are obtained by applying Apriori algorithm and FP-growth algorithm. By analyzing the data, and giving different support and confidence values, we can obtain different number of rules. During analysis it is found that FP-growth is much faster for large number of transactions as compare to Apriori. It takes less time to generate frequent itemsets for FP-Growth algorithm as shown in Table1.

VI. CONCLUSION AND FUTURE WORK

In this paper we analyse the Apriori and FP-growth algorithm. It is our finding that Apriori algorithm takes more time to compute association rules as compare to FP-Growth algorithm for the same dataset for the Arabic text. We will analyze these two algorithms for the Arabic commentaries in future.

REFERENCES

[1] Agrawal , R., Imielinski, T. and Swami, A, "Mining Association Rules between Sets of Items in Large Databases," in Proc. of SIGMOD 1993.

[2] Agrawal, R., Srikant, R, " Fast Algorithms for Mining Association Rules,". in Proc. of VLDB 1993.

[3] Agrawal, R. ,Srikant, R, " Mining Sequential Patterns," in Proc. of IDCE 1995.

[4] S. Brin,S. Motwani R. and C. Silverstein, C, " Beyond Market Basket: Generalizing Association Rules to Correlations," in Proc. of SIGMOD 1997.

[5] Agrawal, R., Gehrke, J., Gunopulos, D.and Raghavan, P, " Automatic subspace clustering of high dimensional data for data mining applications," in Pro. of the ACM-SIGMOD 1998.

[6] Han, J., Pei,J., and Yin, Y. 2000 Mining frequent patterns without candidate generation. in Proc. of ACM SIGMOD Conference.

[7] Grahne, G., Zhu, J, " Efficiently using prefix-trees in mining frequent itemsets," in Proc. of the ICDM'03, Melbourne, FL, pp 123–132,2003.

[8] B. Goethals, B. and Zaki, M, "Frequent Itemset Mining Implementations," in Proc. of the ICDM 2003.

[9] Thabet, N. 2005 Understanding the thematic structure of the qur'an: An exploratory multivariate approach. In Meeting of the Association for Computational Linguistics.

[10] Jiawei et al. 2007 Frequent pattern mining: current status and future directions. Springer Science+Business Media, LLC 2007

[11] Sumanta, Guha, "Efficiently Mining Frequent Subpaths," in Proc. of Eigth Australasian Data Mining Conference, Melbourne, Australia.2008.

[12] *Data mining, concepts and techniques*," Morgan Kaufman: 2009

[13]] Sharaf, M., and Atwell, Eric., 2012 QurSim: A corpus for evaluation of relatedness in short texts. LREC, 2012.

[14] Sharaf, M., and Atwell, Eric., 2012 QurAna: Corpus of the Quran annotated with Pronominal Anaphora. LREC. 2012.

[15] Imran, Ali, "Application of a Mining Algorithm to Finding Frequent Patterns in a Text Corpus: A Case Study of the Arabic, " International Journal of Software Engineering and Its Applications vol. 6, No. 3. 2012.

[16] Mehay, Ankur., Singh, Khawljeet,"AnalyzeMarket Basket Data using FP-Growth and Apriori Algorithm, " in International Journal on Recent and Innovation Trends in Computing and Communication Volume: 1 Issue: 9,2013.

