Survey on Association Rule Mining Algorithms for Opinion Mining

[1]Shalini Tripathi, [2]Maruska Mascarenhas

[1,2]Department of Computer Engineering, Goa University, Farmagudi, Ponda, Goa
Email: [1]tshalini94@gmail.com, [2]maruskha@gec.ac.in

Abstract— Nowadays many people tend to shop online for products with the easily available Internet services. The purchases made online are all based on trust, which is important in making any purchase related decision. Trust also is important for the successfully making profit for any e-commerce website or organization. After buying any product people give their reviews about the product which can largely influence other customers buying decisions.This paper covers a survey on association rule mining algorithms for opinion mining and it gives an overall idea of opinion mining. Association rule mining is mostly used in finding the frequent features from review dataset used for opinion mining. Two algorithms, Apriori and FP-growth are discussed in detail. Any of these algorithms can be used to find the frequent features.

Keywords— Association rule mining, Apriori, FP-Growth.

I. INTRODUCTION

Opinion mining is a growing field which identify the opinion of different people.It helps firms to discover what exactly people like and did not like about the product and customers can understand about product which he/she wants to buy. Opinion mining helps customers to know the opinion of a person about a particular product. Different people can feel differently about products. Due to a vast use of internet, customers are free to publicly express their opinion about a product. The reviews can be seen on various e-commerce websites and customers can read these reviews and get an opinion [1].

The reviews can be positive, negative or neutral[2]. Positive reviews are those which highlight all the good features of a product, negative reviews highlight the bad points and neutral reviews are those which do not highlight any feature of the product. Positive reviews are represented as +1, negative reviews as -1 and finally the neutral as 0. This is called as polarity of reviews which can be find out using SentiWordNet [3]. It is a lexical resource which takes the sentence as input and gives scores as output. Score can be positive, negative and neutral.

Some reviews have more than one sentence in it. These reviews can be positive, negative or neutral based on the polarity of each sentence. If the sentences in a review are separated by a 'but', the polarity of a sentence before 'but' is always opposite to the polarity of a sentence after 'but'. Similarly, if the sentences are separated by an 'and' then the polarity of the sentences will be same. Apart from positive and negative reviews there is something called sarcastic reviews [4]. Sarcastic reviews are the ones which shows sarcasm. Suppose there are some n reviews. q1 to qn. If a review q is positive but review q+1 and q+2 are negative and review q-1 and q-2 are also negative than the review q is considered as sarcastic review and instead of polarity +1 it takes the polarity -1 [5].

There are three different levels of opinion mining. Document level, sentence level and aspect level [6] [7].

1) Document Level: Atthis level the full document is

analysed and given a score as a whole. It means we find the opinion of the full document, whether the document is positive or negative. The output of document level opinion mining can be +1 if the document is positive and -1 if the document is negative.

2) Sentence Level: At this level, each sentence of the

document is analysed. The evaluation is based on sentences. The score is given to each sentence. It classifies whether a particular sentence is positive or negative.

3) Aspect Level: At this level the whole document/

sentence is classified as positive or negative depending on each feature in the document/sentence. First step is to find the feature and after that classifying whether the review is positive or negative for that feature.

Two types of aspects are found in user reviews explicit and implicit [6].

1) Explicit Aspects: are mostly noun or noun phrases

which are easily identifiable.

2) Implicit Aspects: are mostly neither nouns nor noun

phrases hence they cannot be easily identified.

In opinion mining, the association rule mining algorithms are used. Using opinion mining for reviews

which are available on e-commerce sites, the opinion of products can be find out. Sometimes, the customer is interested in knowing the frequent occurring features in reviews [8]. To find frequently occurring features, association rule mining algorithms can be used.The algorithms find the association rules between different itemsets using the frequent items [9].

The first step required while working with reviews is pre-processing. It is used to remove the unnecessary information which is not used to find the polarity. This helps to reduce the size of reviews and makes the work fast [10].

II. LITERATURE REVIEW

This section gives brief introduction towhat is association rule mining and some association rule mining algorithms. The algorithms are Apriori algorithm and FP-Growth algorithm. The most familiar association rule mining algorithm is Apriori algorithm [11].

A. Association Rule Mining

In data mining, association rule mining is a popular technique to find the relation between variables. It was first introduced by Agrawal [12]. It forms association rules which are created by analysing the given data. Rules are always created by using support. Support shows how frequently an item is occurring in a dataset. Association rules are used to predict customer behaviour [13].

B. Apriori Algorithm

There are basically two steps in Apriori algorithm:

- First is to find the frequently occurring itemset using minimum support.
- The second step is to generate association rules from the frequent itemset.

Aprioriperforms breadth first search [14][15] [16].

Algorithm

Step1: Consider all the given transactions and the minimum support.

Step2: Find the occurrence of each item (singletons) from the given transaction and discard the ones which do not satisfy the minimum support. Consider the remaining singletons

Step3: Generate pairs from the singletons and again check for minimum support. Make association set from the remaining pairs.

Step4: Similarly, generate the triplets and quadruples and add those itemsets to association set which pass the minimum support.

Step5: The remaining association set is the set of frequent items.

Let us suppose the minimum support is 50% and the transactions are:

TABLE I : TRANSACTIONS GIVEN TO APRIORI

Tid	Transactions
T1	beer, wine, bread, butter
T2	wine, diaper, bread, beer
T3	milk, butter, wine, beer
T4	beer, butter, spinach, eggs

First Step: Apriori finds the singletons i.e it counts the occurrence of each itemset

TABLE II : OCCURRENCE OF SINGLETONS

Items	Occurrences
beer	4
wine	3
bread	2
butter	3
diaper	1
milk	1
spinach	1
eggs	1

Diaper, milk, spinach and eggs do not satisfy the minimum support. So, discard them.

Second Step: now make the pairs of the remaining word which satisfies minimum support.

Pairs are: {beer, wine}, {beer, bread}, {beer, butter}, {wine, bread}, {wine, butter}, {bread, butter}

Now count the occurrence of each pair

TABLE III : OCCURRENCE OF PAIRS FORMED BY SINGLETONS

Pairs	Occurrences
{beer, wine}	3
{beer, bread}	2
{beer, butter}	3
{wine, bread}	2
{wine, butter}	2
{bread, butter}	1

Discard the pair {bread, butter} because it does not satisfy the minimum support. Make association set with the remaining pairs.

Set = {{beer, wine}, {beer, bread}, {beer, butter}, {wine, bread}, {wine, butter}}

Third Step: generate triplets and apply minimum support

TABLE IV : OCCURRENCE OF EACH TRIPLET

Triplets	Occurrences
{beer, wine, bread}	2
{beer, wine, butter}	2
{wine, bread, butter}	1

{beer, wine, bread} and {beer, wine, butter} passes the minimum support. Hence add these two triplets and remove the subsets that are inside it.

Set = {{beer, wine, bread}, {beer, wine, butter}}

Fourth Step: generate quadruples and apply minimum support{beer, wine, bread, butter}: 1

Discard this quadruple.

Set = {{beer, wine, bread}, {beer, wine, butter}}

This is the final result.

Main advantage of Apriori is that it forms more sets of frequent items. Disadvantage is that it reads a file for each iteration to count the itemsets and therefore it takes more time. It generates singletons, pairs, triplets, quadruples etc which can be slow.

C. FP-Growth Algorithm

It is the improvement of Apriori algorithm. It basically eliminates the bottlenecks of Apriori. It works in depth-first order [17]. FP-Growth usesfrequent access pattern tree(Fp-tree) and simplifies the problems of Apriori. Each node of Fp-tree represents an item and its count.

Algorithm

Step1: Consider the given transaction and minimum support.

Step2: Find the occurrence of each item in the transactions and discard the ones which do not satisfy minimum support.

Step3: Sort the remaining items in increasing order according to the number of occurrences.

Step4: Build Fp-tree for the first transaction and start inserting items of each transaction in Fp-tree. Insert in the same order as the items are in the sorted list.

Step5: Increase the count for the repeated items.

Step6: Repeat step4 till the last transaction.

Step7: Discard those branches which do not pass the minimum support.

Step8: The remaining Fp-tree is the final result and the remaining branches form an association set which is the set of frequent items.

For example, consider the same transactions as given above. The first step will remain same as Apriori algorithm. In second step consider the minimum support as 50% which means that each item should present two times in a transaction. The items which appear less than two times should be discarded.

The third step is to arrange the items in increasing order of their occurrence. Sorted transaction is:

T = {beer: 4, wine: 3, butter: 3, bread: 2}

Next step is to start building the tree. Insert each transaction in the tree and insert the items in such a way that they will be in the same order as in the sorted list.

First transaction: {beer, wine, bread, butter}

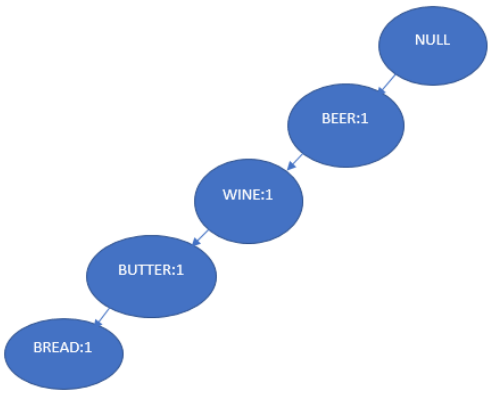

Fig.1 Fp-tree after first transaction

Second transaction: {wine, diaper, bread, beer}

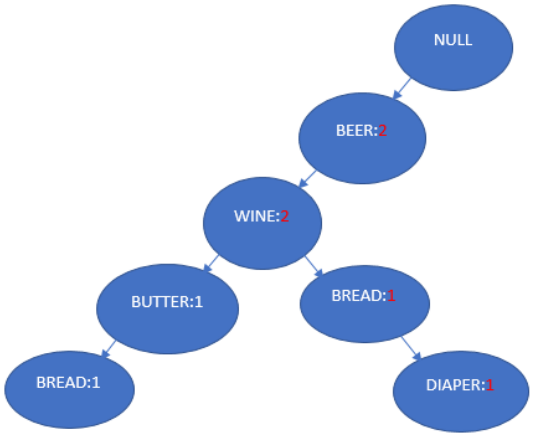

Fig.2 Fp-tree after second transaction

Third transaction: {milk, butter, wine, beer}

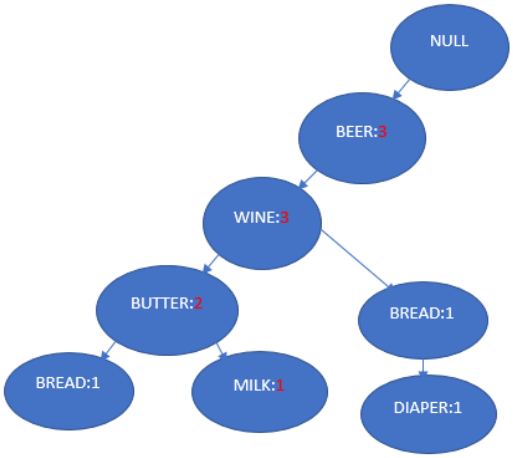

Fig.3Fp-tree after third transaction

Fourth transaction: {beer, butter, spinach, eggs}

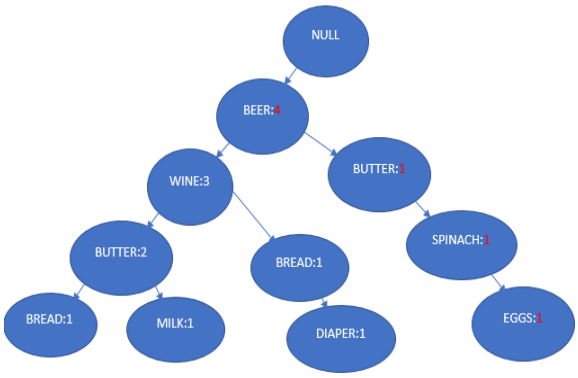

Fig.4 Fp-tree after fourth transaction

Final step: go through all the branches and consider only those which passes the minimum support.

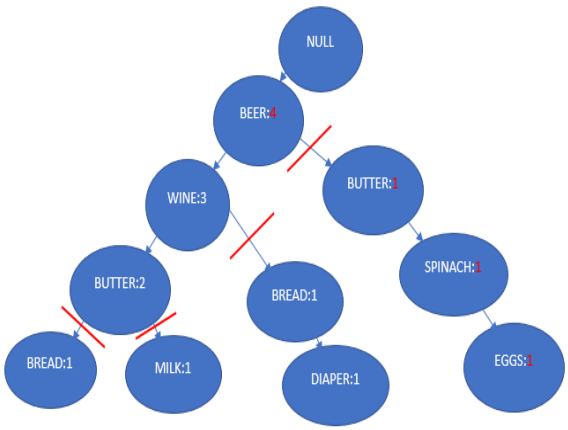

Fig.5The final Fp-tree after all the transaction

Therefore, the Association = {beer, wine, butter}

FP-Growth does not count the frequent pairs, it directly forms a F-tree and the final result is frequently occurring itemsets. Biggest advantage of FP-Growth is that it is fast in computation as compared to Apriori. Disadvantage is that the data is dependent.

III. COMPARISON

TABLE V : COMPARISON BETWEEN APRIORI AND FP-GROWTH

Algorithms	Technique	Advantages	Disadvantages	Performance Analysis
Apriori Algorithm	It generates singletons, pairs, triplets, quadruples etc to find frequent itemset.	It forms more set of frequent sets. It helps to find the frequency of each item more precisely.	It takes more space for storage. It reads the file for each iteration therefore it takes more time.	Huge memory consumption. Runtime is more.
FP-Growth Algorithm	It forms Fp-tree by inserting the items in sorted order of their occurrences.	It does not form pairs, triplets, quadruples etc therefore the storage is less. It reads the file just two times and hence it is fast.	Interdependency of data is present. No parallelization.	It has less memory usage and less runtime. It is more scalable.

IV. CONCLUSION

This paper covers the overview of opinion mining and the uses of it. Two main algorithms of association rule mining are explained in detail.The algorithms are Apriori and FP-Growth.These algorithms are used to find the frequent itemsets from the given transactions. Apriori generates association rules after finding the frequent items. These rules are used to find the relation between items. FP-Growth forms Fp-tree based on the sorted list of items.

REFERENCES

[1] MuktaPatkar, Pooja Pawar, Mony Singh and Ashwini Save, A New way for Semi Supervised Learning Based on Data Mining for Product Reviews, 2nd IEEE International Conference on Engineering and Technology (ICETECH), 17th & 18th March 2016, Coimbatore, TN, India, 2016.

[2] Pankaj Kumar, KashikaManocha, and Harshita Gupta. "Enterprise analysis through opinion mining." Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on. IEEE, 2016.

[3] Apoorv Agarwal, Vivek Sharma, GeetaSikka and RenuDhir, Opinion Mining of News Headlines using SentiWordNet, 2016 Symposium on Colossal Data Analysis and Networking (CDAN), 2016.

[4] Ellen Riloff, AshequlQadir, PrafullaSurve, Lalindra De Silva, Nathan Gilbert and Ruihong Huang, Sarcasm as Contrast between a Positive Sentiment and Negative Situation**,** Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 704–714, Seattle, Washington, USA, 18-21 October 2013.

[5] Roberto González-Ibáñez, SmarandaMuresan and Nina Wacholder, Identifying Sarcasm in Twitter: A Closer Look, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: shortpapers, pages 581–586, Portland, Oregon, June 19-24, 2011.

[6] Richa Sharma, Shweta Nigam and Rekha Jain, Mining of Product Reviews at Aspect Level, International Journal in Foundations of Computer Science & Technology (IJFCST), Vol.4, No.3, May 2014.

[7] T.C. Chinsha and Shibily Joseph, A Syntactic Approach for Aspect Based Opinion Mining, in Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing, 2015.

[8] J. Jin and P. Ji, Mining Online Product Reviews to Identify Consumers Fine-Grained Concerns, in Proceedings of the ISORA, 2015.

[9] Hemant Kumar Soni, Sanjiv Sharma, and Manisha Jain. "Frequent pattern generation algorithms for Association Rule Mining: Strength and challenges." Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on. IEEE, 2016.

[10] C. Fiarni, H. Maharani, and R. Pratamal, Sentiment Analysis System for Indonesia Online Retail Shop Review Using Hierarchy Naive Bayes Technique, inProceedings of the Fourth International Conference on Information and Communication Technologies, 2016.

[11] Harikumar, Sandhya, and Divya Usha Dilipkumar. "Apriori algorithm for association rule mining in high dimensional data." Data Science and Engineering (ICDSE), 2016 International Conference on. IEEE, 2016.

[12] Rakesh Agrawal, Tomasz Imielinski andArun Swami, Mining Association Rules between Sets of Items in Large Databases, Proceedings of the 1993 ACM SIGMOD Conference Washington DC, USA, May 1993.

[13] Peddi Kishor and Dr.SammulalPorika, An Efficient Approach for Mining Positive and Negative Association Rules from Large Transactional Databases, Inventive Computation Technologies (ICICT), International Conference on. Vol. 1. IEEE, 2016.

[14] Paresh Tanna andDr. Yogesh Ghodasara., Using Apriori with WEKA for Frequent Pattern Mining, International Journal of Engineering Trends and Technology (IJETT) – Volume 12 Number 3 - Jun 2014.

[15] Avadh Kishor Singh, Ajeet Kumar and Ashish K. Maurya, An Empirical Analysis and Comparison of Apriori and FP- Growth Algorithm for Frequent Pattern Mining, 2014 IEEE International Conference on Advanced Communication Control and Computing Technologies (lCACCCT).

[16] Ajay Kumar Shrivastava and R. N. Panda, Implementation of Apriori Algorithm using WEKA, KIET International Journal of Intelligent Computing and Informatics, Vol. 1, Issue 1, January 2014.

[17] Dr.S.Vijayarani and Ms. S.Sharmila, Comparative Analysis of Association Rule Mining Algorithms, Inventive Computation Technologies (ICICT), International Conference on. Vol. 3. IEEE, 2016.