A Comparative Study of Pattern Recognition Algorithms on Sales Data

Maulik Shah
K.J.Somaiya College ofEngineering

Nirali Shah
K.J.Somaiya College of Engineering

Anviksha Shetty
K.J.Somaiya Collegeof Engineering

Darshan Shah
K.J.Somaiya College of Engineering

Pradnya Gotmare
K.J. Somaiya College of Engineering

ABSTRACT

In the realm of Data Mining looking for patterns and association rules is a very critical task and has been widely studied in the past years. There exist several data mining algorithms to find Association Rules in given datasets. One of the most popular and widely used algorithm is the Apriori algorithm to find patterns and itemsets in huge datasets and getting the association rules between them. This is done to gather knowledge from otherwise unsuspecting and random data. The Fp-Growth algorithm is similarly a different algorithm which uses an extended frequent pattern prefix-tree data structure for storing critical data after compression about frequent pairs. In this paper we do a comparative analysis of the 2 most popular pattern recognition algorithms and their performance on sales data of a college canteen sales transnational database where each record consists of items purchased by customer.

General Terms

Pattern Recognition, Data Mining, Apriori, Fp-Growth, Frequent Itemsets, Item-pair.

Keywords

Comparison, Data Mining, Frequent, Itemset, Apriori, Algorithm, FP-Growth, Knowledge Discovery.

1. INTRODUCTION

With the progress of the technology and the boom of information there is an ever increasing need for extracting useful information from datasets to support business decisions [2], the process of knowledge discovery from datasets and data mining seems to help achieve this goal. Data mining is the process of finding interesting and useful correlations and patterns in massive heaps of data which are stored in data warehouse, OLAP (on line analytical process), databases and other information repositories [1].

While developing a software solution for our college canteen we came across a unique situation. The profit margins are very low so the profit is to be made by increasing the sales volume. So we tried working on existing dataset from the canteen to find patterns of items being frequently sold together and try and incentivise higher sales by introducing offers on these items sets so their sales increase thereby increasing the profits. To find these item-sets we used the principles of data mining for knowledge discovery by trying to find frequent item-sets in the historic sales data. Even if 10% of the people tend to buy 2 things together there's a higher chance of increasing the sales of those 2 items by giving an incentive to the customers thereby increasing the sales volume and subsequently the profits. For example consider, "15% of all customers who buy a pizza also buys a coke". This helps us identify customer purchase patterns and aids incentive generation and product placement.

Sometimes the sales data can reach over 1000 transactions per day with an average of over 600 transactions per day. This data piles up if we consider data over months and years and the computing becomes very expensive and time consuming. In this paper we compare and provide an overview of the two most widely used Association rule mining algorithms and provide a comparative performance analysis of the same.

2. LITERATURE SURVEY

2.1 Data Mining

Data mining is a step in the process of knowledge discovery in datasets. It has been defined as the analysis and aggregation of (mostly large) transactional or observational data sets to finds patterns and relations that aren't otherwise obviously visible. It also means summarising the data in a way that can be both understood and easily used by the user. It is also defined as a process of extraction of Implicit, previously unknown, unsuspecting and useful information from the data generally stored in a database, XML repository or a data warehouse [3][6]. In practice, data mining is an analytical process of finding patterns and cross-correlations amongst various fields in large RDMS and distributed databases than can help a business decision support system. The main goal of data mining is extracting useful information from large volumes of data. Data mining has been used in a wide array of applications such as fraud detection, CRM, market analysis, behaviour analysis and other business decision making process [5].

2.2 Knowledge Discovery Database (KDD)

Knowledge Discovery in Databases(KDD) is defined as the process of discovering useful and meaningful knowledge from datasets. Data mining is a single step in the iterative and interactive process of knowledge discovery which involves implementing a specific algorithm. The KDD process has many steps. The Fig.1 below shows the steps involved in Knowledge Discovery

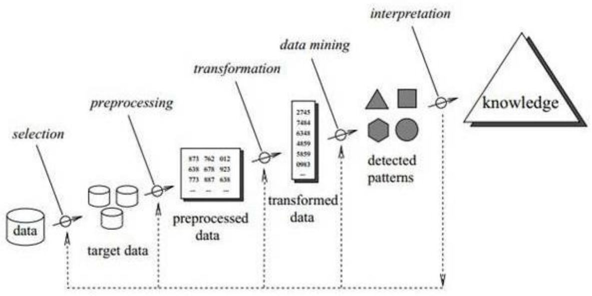

Fig.1 Knowledge Discovery Process [14]

2.3 Association Rule

This paper concentrates on association rule mining as it focuses on findings the relation between items bought by the customers from the purchase and sales data. The objective is to discover the concurrence associations amongst the data in very large sales databases, and to discover unique associations amongst the attributes. An association rule is an if-then-rule or a direct implication rule that is supported by dataset. The goal for the development of association rules is to use them market basket analysis which analyses point-of-sale(POS) transactional datasets [7].

Association Rule mining is widely used for marketing and by retailers to find relations between the products sold. Association rule mining is also used in many diverse fields ranging from CRM to health industry. When association rule mining is implemented on the sales data to find relations between the items sold it is called market basket analysis (MBA). Technically ARM is finding how a certain item or attribute related to another item and how strongly those two are correlated. Stated by Ping & Chou [9] to develop association rules, we must identify how frequently a customer buys an item Y if he has also bought an item X. Here item X is the condition while item Y is the result. To determine association rules we need to specify the confidence and minimum support value to restrict whether a rule is useful or not [6].

Support is expressed as a percentage which represents the probability of a randomly selected subset of transactions from a dataset which will have items X & Y.

The formula is mathematically expressed as:

Support (X -> Y) = P (X n Y)

It is unnecessary to check each and every association and therefore to increase efficiency, Market basket analysis prunes support below a certain value called the threshold value [20].

Confidence which is also expressed as a percentage denotes the probability that a random set of transactions from the data set will include Y given that it also has X.

Expressed mathematically, the formula can be simplified as below:

Confidence (X -> Y) = P (Y|X)

Rules are documented as associations if their confidence value is above a given threshold (minimum confidence). This means we look for association rules that have high chances of being true.

The goal association rule mining is to report all the associations whose support value and confidence value exceed their respective predefined threshold values. The association rules encapsulate the relations and associations between attributes in the database, for Example, pizza implies coke: 0.04 support; 0.60 confidence denote that in the database, 60% of the people who buy pizza also buy coke, and these people are found in 4% of the records in the database. This shows a positive (directional) relationship between people who buy pizza and coke [8]. Association rule mining is a two-step process [6]:

- Frequent Item set Generation: generate all set of items that have support greater than a certain threshold, called min support.
- Association Rule Generation: from the frequent item set, generate all association rules that have confidence greater than a certain threshold called min confidence.

The main challenges of using association rules is that there are so many possible rules and realtions between all the attributes. For example, a supermarket may contain thousands to millions of products, and between there are billions of associations and relations possible. It is clear that to process such a vast number of rules each rule must be tested one at a time. Thus we need efficient and fast algorithms that limit the search space and check only a subset of each rule based on certain parameters. Two such algorithms are Apriori algorithm and FP-Growth.

Market basket analysis:

Market Basket Analysis (MBA) is applied on sales transaction data to aid in avenues such as business decision support, cross-selling of products, customer purchase behaviour analysis, and CRM. Using rules of data mining, it is used discovering sales patterns by determining association-rules from sales transaction data. In analysing MBA association rules, we identified 3 types of rules [10]:

- Actionable rules, which provide understandable and high-quality information and suggest effective promotions. For example is the unlikely combination of beer and diapers.
- Trivial rules, which are rules that can be identified by common sense or may reflect past marketing strategies or product bundles.
- For example, if paint, then brush.
- Inexplicable rules, which seem purely coincidental and have no clear explanation and do not suggest a clear or ethical course of action. For example a combination of car spares and kitchenware.

Market Basket Analysis(MBA) is extensively used to find correlations between products that are sold and to develop marketing strategies based on these correlations. Any commodities and goods companies can benefit by using MBA. For example, in LimitedBrands, a family of different fashion brands, the outcome of an extensive Market basket analysis was the following [11]:

- When additional products are sold with the base products the revenue from the base products isn't decreasing but actually increasing to an increase in the sales volume.
- "Buy two, get three" campaigns to promote sales are very fruitful, if market basket analysis is used to

determine the right pairs of products to be pushed together.

- Using market basket analysis, product sets are identified and tried to sell together with incentives generally in the form of discount to increase the sales volume.

3. CLASSICAL MINING ALGORITHMS

3.1 Apriori

In [4], Agrawal proposed an algorithm called Apriori to address the problem of mining association rules. It is a breadth first, bottom up approach to mine association rules. The frequent itemsets are extended one item at a time. The main idea is to generate k-th candidate itemsets from the (k-1)-th frequent itemsets and to find the k-th frequent itemsets from the k-th candidate itemsets. The algorithm terminates when frequent itemsets can't be extended any more. But it has to generate a large amount of candidate itemsets and scans the data set as many times as the length of the longest frequent itemsets. [12]

The pseudocode for Apriori Algorithm is:

```
$C_k$: Candidate Itemset of size k
$L_k$: Frequent Itemset of size k

$L_1$={Frequent Items};
for(k=1; $L_k$!= Ø; k++) do begin
        $C_{k+1}$= Candidates generated from $L_k$;
        for each transaction t in database do increment the
        count of all cadidates in $C_{k+1}$ that are contained in t

        $L_{k+1}$=Candidates in $C_{k+1}$ with min_support end

        Return $\cup_k L_k$;
```

3.2 FP-Growth Algorithm

In [13], Han, Pei et al. proposed a data structure known as the FP-tree. It stands for frequent pattern tree. It is a compact and aggregated representation of all the related frequency information in the data set. Every path in the FP-tree represents a frequent itemset and the nodes are stored in a decreasing order of the frequency of the corresponding items. A great advantage of FP-tree over other algorithms is that overlapping itemsets share the same prefix path. Hence the information of the data set is greatly compressed and stored in a compact way. The data set needs to be scanned only twice and candidate item sets aren't required [12]

An FP-tree has a header table. The nodes in the header table link to the same nodes in its FP-tree. Single items and their counts are stored in the header table by decreasing order of their counts. The following figures shows the FP-Tree constructed by that data set with minsup (minimum support) = 30%.[12]

Consider the transaction data set:

A B C

A B C E F

D F

A B C

A C E G

B C

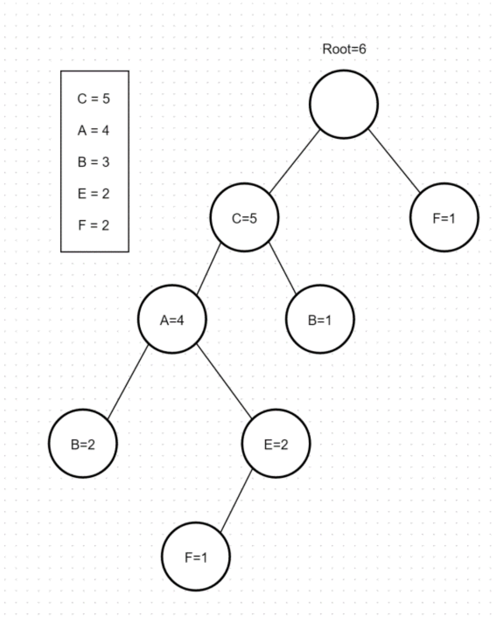

Fig.2 Frequent Pattern Tree

The FP Growth algorithm is:

Input: FP-tree constructed with the above mentioned algorithm;

D - Transactional database;

s - Minimum support threshold.

Output: The complete set of frequent patterns.

Method:

call FP-growth(FP-tree, *null).* **Procedure** FP-growth *(Tree, A)*

if *Tree* contains a single path *P*

then for each combination (denoted as *B)* of the nodes in the path *P* **do**

generate pattern B ❑ A with *support=minimum support of nodes in B*

else for each *ai* in the header of the *Tree* **do**

generate pattern *B = ai* ❑ A with *support = ai.support;* **construct** *B'*s conditional pattern base and *B's* conditional *FP-tree*

TreeB;

if *TreeB*

then call FP-growth *(TreeB, B)*

4. EXPERIMENTAL RESULTS

The contents of our test data set are the sales transaction data of our college canteen.
These algorithms were run on Intel Core i7-5500U CPU with a 8gb of RAM. Each of these algorithms are implemented in JAVA and executed on the BlueJ IDE. We compare the performance of Apriori versus Fp-Growth and the results are tabulated below. They were evaluated based on their execution times for various values of minimum support in the Table.1 and Fig. 3 below.

Table 1. Apriori Vs FP-Growth

Support	Apriori (in secs)	FP-Growth (in secs)
10%	1.052	0.8972
7.5%	3.858	1.3468
5%	7.419	1.8908
4%	12.329	1.8972
3%	12.462	2.1086
2%	13.221	2.1098
1%	17.862	2.1108
0.5%	24.655	2.1221
0.1%	223.272	2.1269

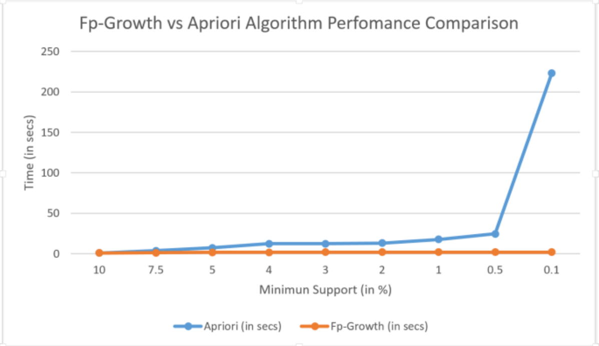

Fig 3. FP-Growth vs Apriori Perfomance Comparison Graph

5. COMPARISON

Table 2. Comparison Chart

	Apriori	FP-Growth
Technique	Generate singletons, pairs, triplets, etc.	Insert sorted items by frequency into a pattern tree
Runtime	Candidate generation is extremely slow. Runtime increases exponentially depending on number of different items.	Runtime increases linearly, depending on the number of items and transactions.
Memory Usage	Saves singletons, pairs, triplets, etc.	Stores a compact version of the database in the Fp-tree.
Parallelizability	Candidate generation is very parallelizable	Data is very inter dependent, each node needs the root.

6. CONCLUSION

Association rules and market basket analysis play an important role in data mining applications. Apriori and Fp-Growth algorithm are the most common algorithms for finding frequent patterns in data sets. The results show that for canteen data set FP-growth provides a much more consistent and quicker performance. This is due to the fact that Fp-growth uses the strategy of divide and conquer and it needs to scan the data set only twice and therefore is much more scalable.

7. REFERENCES

[1] M. Halkidi, “Quality assessment and uncertainty handling in data mining process,” in Proc, EDBT Conference, Konstanz, Germany, 2000.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” AI magazine, vol. 17, no. 3, p. 37, 1996.

[3] Fayyad, U. (1997). Data Mining and Knowledge Discovery in Database: Implications from Scientific Database (pp.2-11), Washington USA.

[4] R. Agrawal, R. Srikant, “Fast algorithms for mining association rules”, Proceedings of the 20th Very Large DataBases Conference (VLDB’94), Santiago de Chile, Chile, 1994, pp. 487-499.

[5] Chen, Y.L ,K.Tang.R.J. Shen and Y.H.Hu (2005). Market Basket Analysis in a multiple store environment, Decision Support Systems 40(2):339-54. Retrived on April 2, 2012 from http://cqx.sagepub.com/centent/51/4/492

[6] Kamber, Jiawei Han (2006). Data Mining Concepts and Techniques, 2nd edition Elsevier publications. Retrieved on May 14, 2012 from http://www.elsevier.com/locate/dsw/pdf

[7] Srikant,R. (1994). Fast Algorithm for Mining Association Rules in Large Database. Proc Int. Conf. on Very Large Database (pp.478-499). Santiago, Chile.

[8] Teng, C.M. (2003). A Compariosn of Standard and Interval Association Rules. In Proceedings of the Sixteenth International FLAIRS Conference (pp. 371-375).

[9] Ping Ho Ting, Steve Pan & Shou Shiung Chou. (2010). Finding Ideal Menu Items Assortments. An Empirical Application of Market Basket Analysis.SAGE.

[10] Pardoe, I. (2008). Data mining techniques:Market basket analysis rules.

[11] LimitedBrands (2004). Achieving Greater Efficiencies with Market Basket Analysis, Microstrategy World 2004 Conference, Miami.

[12] Bo Wu; Defu Zhang; Qihua Lan; Jiemin Zheng, "An Efficient Frequent Patterns Mining Algorithm Based on Apriori Algorithm and the FP-Tree Structure," in *Convergence and Hybrid Information Technology, 2008. ICCIT '08. Third International Conference on* , vol.1, no., pp.1099-1102, 11-13 Nov. 008 doi: 10.1109/ICCIT.2008.109

[13] J.Han, J.Pei and Y.Yin., “Mining frequent patterns without candidate Generation”, in: Proceeding of ACM SIGMOD International Conference Management of Data, 2000, pp. 1-12.

[14] http://www.sqldatamining.com/wpcontent/uploads/2012/11/Steps-of-the-Knowledge-Discovery-in-Databases-Process.jpg