International Journal of Engineering & Technology, 7 (4.33) (2018) 204-207

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Implementation of Apriori Algorithm for Determining Purchase Patterns in One Transaction

Murnawan[1]*, Ardiles Sinaga[2], Ucu Nughraha[2]

[1]Information System, Faculty of Engineering Widyatama University, Bandung, Indonesia
[2]Informatics, Faculty of Engineering Widyatama University, Bandung, Indonesia
**Corresponding author E-mail: murnawan@widyatama.ac.id*

Abstract

The organization data owned is one of the assets of the organization. With the daily operational activities, the longer the data will increase. By using techniques that can do data processing, these data can be obtained important information that can be used for future developments. Association rules are one of these techniques which aims to find patterns in the form of products that are often purchased together or tend to appear together in a transaction from transaction data which is generally very large by using the concept association rules themselves derived from Market Basket Analysis terminology, namely search for relationships from several products in a purchase transaction. In designing this application will build applications that classify the data items based on the tendency to appear together in a transaction using the Apriori Algorithm. The Apriori algorithm is the first algorithm and is often used to find association rules in data mining applications with association rule techniques.

Keywords: *Apriori algorithm; association rule; market basket analysis.*

1. Introduction

Daily operational activities make data longer and more numerous. However, this data is often treated only as a record without further processing so it does not have more use value for future needs. This causes the need for techniques that can process data so that the available data can be obtained important information that can be used for future developments.
With the application of the Apriori Algorithm, it is expected to find a pattern in the form of products that are often purchased together. This pattern can be used to place products that are often purchased together into an area that is close together, design product displays in catalogs, design coupons to design package sales.
Based on the results of the search for the background of the problem and the predetermined title, the problem that can be identified is how the Apriori Algorithm can classify the item data according to the tendency level to appear together in a transaction.

1.1. Scope of Problems

The problem limitation in this research is as follows:
1. It will only analyze consumer spending habits which can later be used for knowledge about consumer habits in shopping
2. The algorithm that will be applied to the application in this study is the Apriori Algorithm which is one of the Association rules techniques

1.2. Research Objectives

Based on the description on the identification of the problems above, the purpose of this research is expected to find patterns in the form of products that are often purchased together and classify the data of goods according to the level of tendency to come out together in a transaction using the Apriori Algorithm.

2. Literature Review

2.1. Market Basket Analysis

Market Basket Analysis is one of the key techniques used by large retailers to uncover associations between items. It works by looking for combinations of items that occur together frequently in transactions. To put it another way, it allows retailers to identify relationships between the items that people buy.
A typical example of frequent itemset mining is market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customers place in their "shopping baskets". The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. For instance, if customers are buying milk, how likely are they to also buy bread (and what kind of bread) on the same trip to the supermarket? Such information can lead to increased sales by helping retailers do selective marketing and plan their shelf space.
If we think of the universe as the set of items available at the store, then each item has a Boolean variable representing the presence or absence of that item. Each basket can then be represented by a Boolean vector of values assigned to these variables. The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently associated or purchased together. These patterns can be represented in the form of association rules. For example, the information that customers who purchase computers also tend to buy antivirus software at the same time is represented in Association Rule (1) below:

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

computer⇒ antivirus[support = 2%; confidence = 60%] (1)

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for Association Rule means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the software. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts. Additional analysis can be performed to uncover interesting statistical correlations between associated items [2].

2.2. Frequent Itemsets, Closed Itemsets and Association Rules

A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset. The set {computer, antivirus software} is a 2-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known, simply, as the frequency, support count, or count of the itemset.

$$\text{support}\left(A \Rightarrow B\right) = P(A \cup B) \quad (2)$$

$$\text{confidence}\left(A \Rightarrow B\right) = P(B \mid A) \quad (3)$$

$$\text{confidence}(A \Rightarrow B) = P(B|A) \quad (4)$$

$$= \frac{\text{support}(A \cup B)}{\text{support}(A)} = \frac{\text{support_count}(A \cup B)}{\text{support_count}(A)} \quad (5)$$

Rules that satisfy both a minimum support threshold (min sup) and a minimum confidence threshold (min conf) are called Strong Association Rules.
In general, association rule mining can be viewed as a two-step process:

1. Find all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, min_sup.
2. Generate strong association rules from the frequent itemsets: By definition, these rules must satisfy minimum support and minimum confidence.

2.3. The Apriori Algorithm

Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties.
Apriori Algorithm works by several steps. First, the candidate itemsets are generated. Then, scan the database to check the support of these itemsets. This later will generate the frequent 1-itemsets. In this first scan, the 1-itemsets are generated by eliminating itemsets with support below the threshold value. Later, the passes candidates became k-itemsets that generated after k-1 of threshold founded. The iteration of database scanning and calculating support will be resulting support and confidence of each association rule that found. Clear explanation of Apriori Algorithm shown at pseudo below with **T** is for transaction database, the support threshold is τ, C_k is the candidate set for the level k, and c is the confidence of each candidates [1].
The limitations of implementing Apriori algorithm in alert classification is its efficiency to scans and gathered the results. For large dataset such as alert database, it needs large number of scans of dataset. Furthermore, the Apriori Algorithm result only explains the presence and absence of n item in transactional databases [5].

2.4. Predictive Apriori Algorithm

Predictive Apriori algorithm was proposed by [9]. This algorithm uses larger support and traded with higher confidence, and calculate the expected accuracy in Bayesian framework. The result of this algorithm maximizes the expected accuracy for future data of association rules. This algorithm generates association rules as expected number of rules by user. In [9] defines this algorithm by:
Let D be a database whose individual records r are generated by a static process P, let X→Y be an association rule. The predictive accuracy c(X→Y) = Pr(r satisfies Y|r satisfies X) is the conditional probability of Y ⊆ r given that X ⊆ r when the distribution of r is governed by P [3].
By its definition, in [3] calculates the predictive accuracy as

$$E\left(c(r)\middle|\hat{c}(r), s(X)\right) = \frac{\int cB[c, s(X)]\left(\hat{c}(r)\right)P(c)dc}{\int B[c, S(X)]\left(\hat{c}(r)\right)P(c)dc} \quad (6)$$

where $E\left(c(r)\middle|\hat{c}(r), s(X)\right)$ is the expected predictive accuracy of a rule r X→Y. The confidence denotes as $\hat{c}$, and the support of the rule denoted as s(X).

3. Methodology

Data is obtained from a retail company. Their primary focus is to:

1. We first need a list of transactions and what was purchased. This is pretty easily obtained these days from scanning cash registers.
2. Next, we choose a list of products to analyze, and tabulate how many times each was purchased with the others.
3. The diagonals of the table show how often a product is purchased in any combination, and the off-diagonals show which combinations were bought.

3.1. Data Preprocessing

Recorded transaction data in the form of receipts stored in the form of text (.txt) files. The file stores all transactions that occur within 1 day. Many transactions that occur with the items purchased are the same so that the formed datasets are sparse, that is, the types of items sold are many, but the transactions of each item do not occur very often.
Transaction data has the attribute date of the transaction, transaction code, item code, item name, item number, and item price. In some transaction data found two or more of the same item not counted into one item. This raw data is processed first so that it becomes data that can be used for data extraction process.

Table 1: Sample Raw Data

ID	Date	ID Items	Items Name	Brand	Qty
1	2017-07-01	1	Roti (RT)	Sri Roti	2
		2	Air Minum (AM)	Aquos	1
		11	Mie Instan (MI)	Indomi	2
		12	Mie Instan (MI)	Mi Sedap	1
2	2017-07-01	2	Air Minum (AM)	Aquos	1
		6	Snack (SN)	Chiki	2
		3	Roti (RT)	Prambanan	1
		5	Shampoo (SP)	Clean	1
3	2017-07-02	4	Sabun (SB)	Lifboy	1
		5	Shampoo (SP)	Clean	1
		9	Rokok (RK)	Malioboro	1
4	2017-07-02	1	Roti (RT)	Sri Roti	2
		10	Air Minum (AM)	Sprit	2
		8	Snack (SN)	Citata	3
		7	Sabun (SB)	Luks	1
		12	Mie Instan (MI)	Mi Sedap	2
5	2017-07-03	6	Snack (SN)	Chiki	1

3.2. Making Frequency Table Item

Next, create a frequency table of all the items that occur in all the transactions. For this case:

Table 2: Frequency Items Transaction

ID	Items Name						
	RT	AM	SB	SP	SN	RK	MI
1	2	1	0	0	0	0	3
2	1	1	0	1	2	0	0
3	0	0	1	1	0	1	0
4	2	2	1	0	3	0	2
5	0	2	0	0	1	0	0
6	0	0	1	1	0	0	2
7	0	1	0	0	0	1	0
8	0	0	1	1	1	0	3
9	2	1	1	0	2	1	2
10	1	2	0	0	2	0	0
11	2	0	0	0	0	1	0
12	0	0	2	1	0	0	3
13	1	1	0	0	2	0	0
14	0	0	2	2	1	0	0
15	0	0	0	0	0	1	0
16	0	2	0	0	0	0	0
Num. Trans.	7	9	7	6	8	5	6

4. Results and Discussion

In this section, an analysis will be done using Apriori algorithm to find the pattern of each combination of each item purchases.

4.1. Combination One Item

The next step is to make a combination of 1 itemsets on each product and the frequency of each combination is calculated according to the tabular data in the table.

Table 3: Combination One Item

Rule	Support	Confidence	Support x Confidence
If Buy AM Then Buy AM	0.5625	1	0.5625
If Buy SN Then Buy SN	0.5	1	0.5
If Buy SB Then Buy SB	0.4375	1	0.4375
If Buy RT Then Buy RT	0.4375	1	0.4375
If Buy SP Then Buy SP	0.375	1	0.375
If Buy MI Then Buy MI	0.375	1	0.375
If Buy RK Then Buy RK	0.3125	1	0.3125

Based on the Biggest Support Value; If 1 product is placed in a rack, then most likely the best sold is drinking water with a value of 0.5625. If there are certain AM Products with certain brands less sold, then it can be put in one rack with the group then most likely to be sold too.
Based on the Biggest Support x Confidence; The biggest chance of buying AM will buy AM at a value of 0.5625. If there are certain AM Products with certain brands sold out, it can be layed side by side with certain branded AM Products that are sold, then most likely to be sold well too.

4.2. Combination Two Items

The next step is to make a combination of 2 itemsets on each product and the frequency of each combination is calculated according to the tabular data in the table.

Table 4: Combination Two Items

Rule	Support	Confidence	Support x Confidence
If Buy AM Then Buy RT	0.375	0.857142857	0.321428571
If Buy AM Then Buy SN	0.375	0.75	0.28125
If Buy SB Then Buy SP	0.3125	0.833333333	0.260416667
If Buy SB Then Buy MI	0.3125	0.833333333	0.260416667
If Buy RT Then Buy AM	0.375	0.666666667	0.25
If Buy SN Then Buy AM	0.375	0.666666667	0.25
If Buy SP Then Buy SB	0.3125	0.714285714	0.223214286
If Buy SN Then Buy RT	0.3125	0.714285714	0.223214286
If Buy MI Then Buy SB	0.3125	0.714285714	0.223214286
If Buy RT Then Buy SN	0.3125	0.625	0.1953125
If Buy SN Then Buy SB	0.25	0.571428571	0.142857143
If Buy SB Then Buy SN	0.25	0.5	0.125
If Buy RT Then Buy MI	0.1875	0.5	0.09375
If Buy AM Then Buy MI	0.1875	0.5	0.09375
If Buy SN Then Buy SP	0.1875	0.5	0.09375
If Buy SP Then Buy MI	0.1875	0.5	0.09375
If Buy MI Then Buy SP	0.1875	0.5	0.09375
If Buy SN Then Buy MI	0.1875	0.5	0.09375
If Buy MI Then Buy RT	0.1875	0.428571429	0.080357143
If Buy SP Then Buy SN	0.1875	0.375	0.0703125
If Buy MI Then Buy SN	0.1875	0.375	0.0703125
If Buy MI Then Buy AM	0.1875	0.333333333	0.0625
If Buy SB Then Buy RK	0.125	0.4	0.05
If Buy RT Then Buy RK	0.125	0.4	0.05
If Buy AM Then Buy RK	0.125	0.4	0.05
If Buy SB Then Buy RT	0.125	0.285714286	0.035714286
If Buy RK Then Buy SB	0.125	0.285714286	0.035714286
If Buy RT Then Buy SB	0.125	0.285714286	0.035714286
If Buy AM Then Buy SB	0.125	0.285714286	0.035714286
If Buy RK Then Buy RT	0.125	0.285714286	0.035714286
If Buy SB Then Buy AM	0.125	0.222222222	0.027777778
If Buy RK Then Buy AM	0.125	0.222222222	0.027777778
If Buy SP Then Buy RK	0.0625	0.2	0.0125
If Buy SN Then Buy RK	0.0625	0.2	0.0125
If Buy MI Then Buy RK	0.0625	0.2	0.0125
If Buy AM Then Buy SP	0.0625	0.166666667	0.010416667
If Buy RT Then Buy SP	0.0625	0.166666667	0.010416667
If Buy RK Then Buy MI	0.0625	0.166666667	0.010416667
If Buy RK Then Buy SP	0.0625	0.166666667	0.010416667
If Buy SP Then Buy RT	0.0625	0.142857143	0.008928571
If Buy RK Then Buy SN	0.0625	0.125	0.0078125
If Buy SP Then Buy AM	0.0625	0.111111111	0.006944444

Based on the Biggest Support Value; If two products are placed in one rack, then most likely the best sold is AM and RT with a value of 0.375. If there is an AM product or RT with a certain brand that is not sold well, then it can be put in one rack with the group then the big possibility will be sold too.
If two products are placed in one rack, then most likely the best is drinking water and snack with a value of 0.375. If there are AM products or SN with certain brands inadequate, then it can be put in one rack with the group then the most possible will be sold too.
Based on the Biggest Support x Confidence; The biggest chance of buying AM then will buy RT with a value of 0.5625. If there are bread products with certain brands lack of sold, then can be placed opposite with certain branded AM products that are sold, the big chance will be sold too.

4.3. Combination Three Items

The next step is to make a combination of 3 itemsets on each product and the frequency of each combination is calculated according to the tabular data in the table.

Table 5: Combination Three Items

Rule	Support	Confidence	Support x Confidence
If Buy AM And Buy SN Then Buy RT	0.3125	0.714285714	0.223214286
If Buy RT And Buy AM Then Buy SN	0.3125	0.625	0.1953125
If Buy SN And Buy RT Then Buy AM	0.3125	0.555555556	0.173611111
If Buy RT And Buy	0.1875	0.5	0.09375

AM Then Buy MI			
If Buy SB And Buy SP Then Buy MI	0.1875	0.5	0.09375
If Buy MI And Buy SB Then Buy SP	0.1875	0.5	0.09375
If Buy SB And Buy SN Then Buy MI	0.1875	0.5	0.09375
If Buy AM And Buy MI Then Buy RT	0.1875	0.428571429	0.080357143
If Buy SP And Buy MI Then Buy SB	0.1875	0.428571429	0.080357143
If Buy SN And Buy MI Then Buy SB	0.1875	0.428571429	0.080357143
If Buy MI And Buy SB Then Buy SN	0.1875	0.375	0.0703125
If Buy MI And Buy RT Then Buy AM	0.1875	0.333333333	0.0625
If Buy RT And Buy SN Then Buy MI	0.125	0.333333333	0.041666667
If Buy AM And Buy SB Then Buy MI	0.125	0.333333333	0.041666667
If Buy AM And Buy SN Then Buy MI	0.125	0.333333333	0.041666667
If Buy SN And Buy SB Then Buy SP	0.125	0.333333333	0.041666667
If Buy RT And Buy SB Then Buy MI	0.125	0.333333333	0.041666667
If Buy MI And Buy AM Then Buy SB	0.125	0.285714286	0.035714286
If Buy SN And Buy RT Then Buy SB	0.125	0.285714286	0.035714286
If Buy SP And Buy SN Then Buy SB	0.125	0.285714286	0.035714286
If Buy RT And Buy AM Then Buy SB	0.125	0.285714286	0.035714286
If Buy SB And Buy MI Then Buy RT	0.125	0.285714286	0.035714286
If Buy MI And Buy RT Then Buy SB	0.125	0.285714286	0.035714286
If Buy AM And Buy SB Then Buy RT	0.125	0.285714286	0.035714286
If Buy SN And Buy MI Then Buy RT	0.125	0.285714286	0.035714286
If Buy SN And Buy AM Then Buy SB	0.125	0.285714286	0.035714286
If Buy SB And Buy SN Then Buy RT	0.125	0.285714286	0.035714286
If Buy MI And Buy AM Then Buy SN	0.125	0.25	0.03125
If Buy SB And Buy SP Then Buy SN	0.125	0.25	0.03125
If Buy MI And Buy RT Then Buy SN	0.125	0.25	0.03125
If Buy AM And Buy SB Then Buy SN	0.125	0.25	0.03125
If Buy RT And Buy SB Then Buy SN	0.125	0.25	0.03125
If Buy SB And Buy SN Then Buy AM	0.125	0.222222222	0.027777778
If Buy SN And Buy MI Then Buy AM	0.125	0.222222222	0.027777778
If Buy SB And Buy MI Then Buy AM	0.125	0.222222222	0.027777778
If Buy SB And Buy RT Then Buy AM	0.125	0.222222222	0.027777778

Based on the Biggest Support Value; If the 3 products are placed in a rack, then most likely the best sold is AM, SN and RT with a value of 0.3125. If there is a AM product or SN or RT with certain brands lack of sold, then it can be put in one rack with the group.
Based on the Biggest Support x Confidence; the biggest chance of buying AM and SN will buy RT with a value of 0.2232. If there are RT products with certain brands lack of sold, then it can be placed opposite with AM products and SN with certain brands that are sold.

5. Conclusion

The final table of the association rule explains the support and confidence of each combination of 2 itemsets and 3 itemsets. The results of the calculation of support in the final table association rule is obtained from the number of transactions with a combination of items A and B divided by the total transactions in item A. Whereas the confidence is obtained from the number of combined transactions A and B divided by the total transactions. The results of the multiplication of support and confidence are the final results of the Apriori algorithm.

References

[1] Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Databases, pp. 487-499.
[2] Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. ACM Sigmod Record, 26(2), 255-264.
[3] Scheffer, T. (2005). Finding association rules that trade support optimally against confidence. Intelligent Data Analysis, 9(4), 381-395.
[4] Sharma, M., Choudhary, J., & Sharma, G. (2012). Evaluating the performance of Apriori and predictive Apriori algorithm to find new association rules based on the statistical measures of datasets. International Journal of Engineering Research and Technology, 1(6), 1-6.
[5] Shweta, M., & Garg, D. K. (2013). Mining efficient association rules through Apriori algorithm using attributes and comparative analysis of various association rule algorithms. International Journal of Advanced Research in Computer Science and Software Engineering, 3(6), 306-314.
[6] Vijayarani, S., & Sharmila, S. (2016). Comparative analysis of association rule mining algorithms. Proceedings of the IEEE International Conference on Inventive Computation Technologies, pp. 1-6.
[7] Said, A. M., Dominic, P. D. D., & Abdullah, A. B. (2009). A comparative study of fp-growth variations. International Journal of Computer Science and Network Security, 9(5), 266-272.
[8] Tan, P. N., Kumar, V., & Srivastava, J. (2004). Selecting the right objective measure for association analysis. Information Systems, 29(4), 293-313.
[9] Tan, P. N. (2007). Introduction to data mining. Pearson Education India.