Design and Implementation Software for Mining Association Rules (Market Basket Analysis) to Design Product Layout Desicions

Tomi Listiawan

Universitas Bhinneka PGRI, Jl. Mayor Sujadi Timur No.7 Tulungagung, Indonesia

tomi@stkippgritulungagung.ac.id

**Abstract**. A provided data from a transactional database, has supported technical development which can automatically find product association or items saved in the database. This finding association rules between saved product in database known as mining association rules. There are so many theories and algorithm developed for conducting mining association rules. One of algorithm developed is Apriori algorithm. This method, has a main goal to find the maximum frequent itemset. Next, this frequent itemset will be generated into associative rules which are not shown before in the database, become valuable information for considering materials in the decision process. Apriori algorithm is a interpretation technique of mining association rules, will be implemented into a web based software. On the software test which use in some different data, it’s concluded that time for mining association rules depends on the presence of every item in every transaction, total of transaction, minimum support and minimum confidence. For smaller value of minimum support and minimum confidence that entered, program will generate more association rules, vice versa.

1. Introduction

Along with the increasing number of consumers or visitors in a supermarket, the number of transactions in the supermarket are increasing as well. So the transactions are stored in the database will be also greater. Thus, in period of time the transactions gathered are getting larger and larger. To obtain valuable information that is “hidden” in a “pile of transaction data” that exist in the database required a method called data mining.

Availability of database records of customers’ purchases in a certain supermarket has encouraged the development of techniques that can automatically find the associated products or items that are stored in the database, such as data on transactions collected from the bar-code scanners in supermarkets. This database contains very large records such as the records of the registered items purchased by a customer in a purchase transaction. From here the manager of the supermarket can determine whether a group of items are always purchased together, so that the manager of a supermarket could use the information to make the lay-out of his/her supermarket in order to prepare these items can be optimized to each other. In addition, it can be for the purposes of promotion, segmentation buyer, the manufacture of a product catalog, or see spending patterns.

## Mining Association Rules (Market Basket Analysis)

Mining association rules is data mining techniques to discover association rules between a combination of items. It is one of the most important technique of data mining [1]. For example of association rules is from the analysis of purchasing in a supermarket, it can be found how likely a customer buys bread along with milk. This will help the owner of the supermarket to adjust the placement of the goods or design a marketing campaign of a particular items combinations. Since the analysis of the association became famous because of its application to analyze the contents of the shopping cart in the supermarket, association analysis is also often referred to market basket analysis.

The idea of association rules is to examine all the possible relationships if - then (if - then) between the items and choosing only the most likely, as an indicator of dependency relationships among items [3]. Antecedent term is usually used to represent the “if” and consequent to represent the “then”

Whether an association rule is important or not can be determined by two parameters, the support (value of supporting / support) and confidence (value certainty)[6]. Support of a rule is the number of transactions containing items, both in the antecedent and consequent. It is called support because it measures how the level of data supporting the validity of the rules developed. Support is usually measured by the percentage (%).

A large number of rules that may be developed, so we need to choose the rules which have quite strong degree of dependence among items in the antecedent and consequent. In addition, there is another measurement of the support that measures the degree of certainty of the rule of “if-then”. The measurement is called confidence where it is the ratio between the number of transactions that include all of the items in the antecedent and consequence by the number of transactions that included all of the items in the antecedent.

Some formal definitions related to this association rules form the basis of which, let  is set of *m* items and let  is collection of *database* of *n* transactions, where for every . A set of items  called itemset. A transaction  contains itemset *X* if where each itemset *X* is associated with a set of transaction , that is set of transaction contains itemset *X*. let  is non empty itemset with , then *association rules* is implication , and formally :

*supp*(*X*) =, , dan *supp* = *supp*

## Apriori Algorithm

Apriori algorithm is the most famous algorithm to find a frequent item set [2]. Frequent item sets are patterns of individual items in the database that have a frequency or support which is above a certain threshold called the minimum support. Frequent item set is used to formulate the rules of the association.

Apriori algorithm is divided into several stages, called iteration or pass. Each iteration produces a frequent item set of equal length starting from the first pass generates frequent item set with a long one. In this first iteration, the support of each item is calculated by scanning the database. After gaining the support of each item, the items that have support more than the minimum support chosen as the frequent item set of length 1 or often abbreviated 1-itemset. Abbreviations k-itemset means a set of k items.

The second iteration, it produces 2-itemset that each set of this has two items. Firstly, make candidate 2-itemset of a combination of all 1-itemset. Then, for each candidate 2-itemset support, we need to calculate by scanning the database. The support here means the number of transactions in the database that contains the item in candidate 2-itemset. Having the support of all the candidate 2-itemset obtained, 2-itemset candidates who meet the minimum requirements of support can be setted as 2-itemset which is also a frequent item set of length 2.

Further, it is written in the form of pseudocode, a priori algorithm is as follows:

*L*1 = {large 1-itemset};

*k* = 2; // k represents the pass number

**While** () **do**

{*C*k = new candidate of size *k* is generated from *L*k-1;

**Forall** transaction  **do**

increment the count of all candidates in *C*k that are contained in *t*;

*L*k = all candidates in *C*k with minimum support;

*k* = *k* + 1;

}

answer = 

JOIN STEP

**Insert into** candidate *k*-itemset

**Select** *p*.item1, *p*.item2, …., *p*.itemk-1, q.itemk-1

**From** large (*k*-1)-itemset *p*, large (*k*-1)-itemset *q*

**Where** *p*.item1=*q.*item1, …., *p*.itemk-2=*q*.itemk-2, *p*.itemk-1 < *q*.itemk-1

PRUNING STEP

**Forall** itemset *c*candidate *k*-itemset **do**

**Forall** (*k*-1)-subsets *s* of *c* do

**If** ( *s*large(k-1)-itemset) **then**

**Delete** *c* from candidate *k*-itemset

1. Method

In association rule mode, the implemented algorithm in the making of market basket analysis application is Apriori algorithm. Algorithm is used to develop frequent itemset using 1-item first, then value support of every item would be later counted. Item whose support value was above the minimum support value was selected as 1-itemset high frequency pattern and as 2-itemset candidates. By that of 1-itemset, the development of frequent itemset into 2-itemset which would then the value of confidence be calculated next was recursively brought off.

1. Result and Discussion

The design process of this software applies a functional approach that represented by using flow charts. The flow diagram shows that the physical flows both of process and data on data mining software which is created.

Overall program flowchart is as follows:



In general, the process of data mining using association rules mining technique consists of three essential parts. The section consists of input processes minimum support and minimum confidence, the process of extracting association rules, and display output association rules. This process is shown by the following flowchart:



The process of extracting association rules has two important stages, namely k-itemset frequent that are represented by a priori algorithm and the determination of all the rules of the association. Data output from the process of finding frequent k-itemset is the formation of all frequent k-itemset. This data will be used as input data for the next process. The process of formation of all association rules meets the minimum confidence.

The flowchart of searching process frequent k-itemset is defined as follows:



From the flow chart described above, it is showed that the results obtained from the process of the formation of frequent k-itemset is then used to form association rules based on the given minimum value confidence. The algorithm used to form association rules based on frequent k-itemset is called a naive algorithm.

Naïve algorithm pseudocode is as follows:

for each frequent itemset *l* do

for each nonempty proper subset

*s* of *l* do

if (support(*l*)/support(*s*) >=

min\_conf)

output the rule *s* → *l-s* ,

with support =support(l) and

confidence=support(l)/support(s)

From the algorithm above, it shows that the process of formation of association rules is done by checking all the proper subset of frequent k-itemset. Then, it determines the confidence value by dividing the value of the support proper subset of the frequent k-itemset support. All association rules that have the confidence value above the minimum confidence value are set as the output of this process of extracting association rules.

Flowchart of the Naive algorithm is as follows:



To determine the performance of software, it is advised to conduct trials with several different data, with a relatively large number of records. The hardware used in this test is a computer with an Intel Celeron M processor 1.86 GHz with memory (RAM) 512 MB. While the operating system used is Windows XP Professional SP2.

During the test, it needs to perform three different data namely data1, data2, and data3. The data is obtained by input into table transaction and transaction detail table randomly, with a range of data that is determined by using the commands in PHP. To obtain data1, the first step is making a transaction table with the 301 records number of transactions, and create a table with records of 20 amount of goods. Then the transaction code and goods code that have been formed, randomized as much as 3676 times to be put into the column and the column kd\_barang Table kd\_transaksi transaction details . To make the suitable data for the needs of the software input, the final step is to combine the details of the transaction and the transaction table based on the same transaction code. Similarly, the steps are used to create data2 and data3.

The specification of data used are shown in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Data1** | **Data2** | **Data3** |
| Num. of record | 3676 | 2454 | 1198 |
| Num. of transaction | 301 | 201 | 101 |
| Num of item | 20 | 20 | 20 |

In this trial, which will be done is to run the software with the same parameters on different tables. There are 3 tables used and they are called data1, data2, and data3. While the parameters used are the minimum support and minimum confidence. The minimum value of the support includes 25%, 35%, and 40%. While the minimum confidence value entered is 65%.

The results of trials for each table can be seen in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Minimum Support** | **25%** | **35%** | **40%** |
| processing time (second) | 8,80 | 5,86 | 1,187 |
| Num of candidat | 4 | 3 | 3 |
| Num of frequent | 3 | 2 | 2 |
| Num of rules | 286 | 38 | 24 |

From the test results, there are several interesting findings, the number of transactions is not the only factor that affects the time in the process of extracting association rules. It means the greater the number of transactions which do not necessarily require, the greater computing time as well. Some factors that affect the computation process time are the number of occurrences of each item in the transaction, the number of transactions, the input minimum support and minimum confidence.

1. Conclusion

From the test of this data mining software, some conclusions can be drawn as follows:

1. Association rules generated by the software, can be used as a consideration in the decision-making process in a shop or supermarket, such as the layout of goods, inventory, customer segmentation, and a marketing campaign design for a good combination.
2. The application is able to process the transaction data to discover candidate and frequent itemset, then generate association rules to be displayed in the form of text and graphics.
3. In the analysis of some data, it was found that the smaller the minimum support and confidence are determined, the more association rules that can be generated by the application.
4. From the test, it is found that the process which takes the longest time, usually occurs in the process of making the candidate 2-itemset. The more the number of 1-itemset candidates who meet the minimum support, the greater the number of candidate 2-itemset to be generated.
5. The number of 1-itemset candidates is difficult to be confirmed, because it completely depends on the number of occurrences of each item in the transaction, the number of transactions, and the minimum input support. So that the process to generate candidate 2-itemset was also difficult to be confirmed. It is caused by the time of the whole process which is difficult to be estimated or inferred .

References

1. Agrawal, R., Imielinski, T., and Swami, A. N. 2013. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216.
2. Agrawal, R. and Srikant, R. 2014. Fast algorithms for mining association rules. In Proc.20th Int. Conf. Very Large Data Bases, 487-499.
3. Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. 2017. Dynamic itemset counting and implication rules for market basket data. In SIGMOD 1997, Proceedings ACM SIGMOD
4. Cheung, D., Han, J., Ng, V., Fu, A. and Fu, Y. 1996. A fast distributed algorithm for mining association rules, in `Proc. of 1996 Int'l. Conf. on Parallel and Distributed Information Systems', Miami Beach, Florida, pp. 31 - 44.
5. Das, A., Ng, W.-K., and Woon, Y.-K. 2001. Rapid association rule mining. In Proceedings of the tenth international conference on Information and knowledge management. ACM Press, 474-481.
6. Li, Y., Gopalan, R., Effective Sampling for Mining Association Rules, Lecture Notes in Computer Science, Volume 3339, Jan 2004, Pages 391 - 401
7. Santosa, Budi.2007.Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis.Yogyakarta:Graha Ilmu.
8. Techapichetvanich, K., Datta, A., Visual Mining of Market Basket Association Rules, Lecture Notes in Computer Science, Volume 3046, Jan 2004, Pages 479 - 488
9. Wu, X., Zhang, C., Zhang, S.: Efficient Mining of Both Positive and Negative Association Rules, ACM Transactions on Information Systems, Vol. 22, No. 3, July 2004, Pages 381–405.
10. Zaki, M. J., Parallel and distributed association mining: A survey. IEEE Concurrency, Special Issue on Parallel Mechanisms for Data Mining, 7(4):14--25, December 1999.