Design and implementation software for mining association rules (market basket analysis) to design product layout decisions

T Listiawan¹,* and M N Hudha²

¹Universitas Bhinneka PGRI, Jl. Mayor Sujadi Timur No.7 Tulungagung, Indonesia
²Universitas PGRI Kanjuruhan Malang, Jl. S. Supriadi no 48, Malang 65148, 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
The 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 [1]. 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 [2–4].

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 [5,6]. 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.
1.1. 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 [1,4]. 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 [5]. 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) [3,7,8]. 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 \( I = \{I_1, I_2, \ldots, I_m\} \) is set of \( m \) items and let \( D = \{T_1, T_2, \ldots, T_n\} \) is collection of database of \( n \) transactions, where for every \( T \in D, T \subseteq I \). A set of items \( X \subseteq I \) called itemset. A transaction \( T \) contains itemset \( X \) if \( X \subseteq T \) where each itemset \( X \) is associated with a set of transaction \( TX = \{T \in D | T \supseteq X\} \), that is set of transaction contains itemset \( X \). Let \( X, Y \subseteq I \) is non empty itemset with \( X \cap Y = \phi \), then association rules is implication \( X \rightarrow Y \), and formally:

\[
supp(X) = \frac{|\{T \in D | X \subseteq T\}|}{|D|}, \quad conf(X \rightarrow Y) = \frac{|\{T \in D | X \cup Y \subseteq T\}|}{|\{T \in D | X \subseteq T\}|}, \quad \text{dan} \ supp(X \rightarrow Y) = supp(X) \cdot conf(X \rightarrow Y)
\]

1.2. Apriori algorithm
Apriori algorithm is the most famous algorithm to find a frequent item set [9–11]. 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 [1,12–14].

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 [15,16]. 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 set 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:

\[
L_1 = \{\text{large 1-itemset}\};
\]

\[k = 2; \quad // \ k \text{ represents the pass number}\]

\[\text{While } (L_{k-1} \neq \emptyset) \text{ do}\]

\[C_k = \text{new candidate of size } k \text{ is generated from } L_{k-1};\]

\[\text{For all transaction } t \in D \text{ do}\]

\[\text{increment the count of all candidates in } C_k \text{ that are contained in } t;\]

\[L_k = \text{all candidates in } C_k \text{ with minimum support};\]

\[k = k + 1;\]

\}\]

\[\text{answer} = \bigcup_k L_k\]

**JOIN STEP**

**Insert into** candidate \(k\)-itemset

**Select** \(p\).item\(_1\), \(p\).item\(_2\), ..., \(p\).item\(_{k-1}\), \(q\).item\(_k\)

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

**Where** \(p\).item\(_1\)=\(q\).item\(_1\), ..., \(p\).item\(_{k-2}\)=\(q\).item\(_{k-2}\), \(p\).item\(_{k-1}\) < \(q\).item\(_k\)

**PRUNING STEP**

**For all** itemset \(c\) \(\in\) candidate \(k\)-itemset **do**

**For all** \((k-1)\)-subsets \(s\) of \(c\) **do**

**If** \(s \notin \text{large}(k-1)\)-itemset **then**

**Delete** \(c\) from candidate \(k\)-itemset

2. **Methods**

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.

3. **Results 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.

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:
\[ L_1 = \text{set 1-itemset} \]
\[ k = 2 \]
\[ L_{k-1} = \emptyset \]
\[ C_k = \emptyset \]
\[ Y \]

---

**Figure 1.** Searching process frequent k-itemset flowchart.

From the flowchart 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 naïve algorithm.

Naïve algorithm pseudocode is as follows:

```
for each frequent itemset \( l \) do
  for each nonempty proper subset \( s \) of \( l \) do
    if \( \text{support}(l)/\text{support}(s) \geq \text{min}_\text{conf} \)
      output the rule \( s \rightarrow l-s \),
      with support = \( \text{support}(l) \) and
      confidence = \( \text{support}(l)/\text{support}(s) \)
```

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.

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 1.

| Table 1. The specification of data. |
|------------------------------------|
|                              | Data1 | Data2 | Data3 |
| Num. of record                 | 3676  | 2454  | 1198  |
| Num. of transaction            | 301   | 201   | 101   |
| Num of item                    | 20    | 20    | 20    |

From Table 1, 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 2.

| Table 2. T The results of trials for each data. |
|-----------------------------------------------|
| 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 Table 2, 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, that are consistent are consistent with the rules generated by Weka [17]. In other hand, there are different that Weka generates all possible rules including those for items that do not go together. 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.

4. Conclusion
From the test of this data mining software, can concluded that 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. 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. 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. 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. 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.

Acknowledgments
I would like to express my deep and sincere gratitude to my research advisor Mr. Darmawan S, for giving me the opportunity to do research and providing invaluable guidance throughout this research.
References

[1] Elisa E 2018 Market Basket Analysis Pada Mini Market Ayu Dengan Algoritma Apriori J. RESTI (Rekayasa Sist. dan Teknol. Informasi)

[2] Han J, Kamber M and Pei J 2012 Data Mining: Concepts and Techniques

[3] Han J, Pei J, Mortazavi-Asl B, Chen Q, Dayal U and Hsu M C 2000 FreeSpan: Frequent pattern-projected sequential pattern mining Proceeding of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

[4] Chen Y L, Tang K, Shen R J and Hu Y H 2005 Market basket analysis in a multiple store environment Decis. Support Syst.

[5] Brin S, Motwani R, Ullman J D and Tsur S 1997 Dynamic itemset counting and implication rules for market basket data ACM SIGMOD Rec. 26 255–64

[6] Giudici P and Passerone G 2002 Data mining of association structures to model consumer behaviour Comput. Stat. Data Anal.

[7] Kaur M and Kang S 2016 Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining Procedia Computer Science

[8] Yanrong L and Raj G 2005 Effective Sampling for Mining Association Rules Ai 2004 391–401

[9] Dhanabhakayam M and Punithavalli M 2011 A survey on Data mining algorithm for market basket analysis Glob. J. Comput. 5

[10] Kurniawan F, Umayah B, Hammad J, Nugroho S M S and Hariadi M 2017 Market Basket Analysis to Identify Customer Behaviours by Way of Transaction Data Knowl. Eng. Data Sci.

[11] Erişti H, Yildirim Ö, Erişti B and Demir Y 2013 Optimal feature selection for classification of the power quality events using wavelet transform and least squares support vector machines Int. J. Electr. Power Energy Syst.

[12] Sikumbang E D 2018 Penerapan Data Mining Penjualan Sepatu Menggunakan Metode Algoritma Apriori J. Tek. Komput. AMIK BSI

[13] Jha J and Ragha L 2013 Educational Data Mining using Improved Apriori Algorithm Int. J. Inf. Comput. Technol.

[14] Celko J 2006 Market-Basket Analysis Joe Celko’s Analytics and OLAP in SQL

[15] Li L and Zhang M 2011 The strategy of mining association rule based on cloud computing Proceedings of the 2011 International Conference on Business Computing and Global Informatization, BCGIn 2011

[16] Saabith A L S, Sundararajan E and Bakar A A 2016 Parallel implementation of Apriori algorithms on the Hadoop-MapReduce platform - An evaluation of literature J. Theor. Appl. Inf. Technol.

[17] Prasad P and Malik L 2011 Using Association Rule Mining for Extracting Product Sales Patterns in Retail Store Transactions Int. J. Comput. Sci. Eng. 3 2177–82