Market basket analysis using apriori algorithm to find consumer patterns in buying goods through transaction data (case study of Mizan computer retail stores)

M Qisman, R Rosadi, and A S Abdullah*
Department of Computer Science, Universitas Padjadjaran, Jl. Raya Bandung
Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia

*atje.setiawan@iunpad.ac.id

Abstract. Mizan Computer Shop is a shop that is engaged in the trading sector, especially in the field of selling computers and supporting accessories. Growing and increasing number of business actors in the computer sector, can makes the players challenged to be able to create unique differentiation and clear positioning. So, that consumers can differentiate from their competitors. Competitive and dynamic market conditions make every company should always observe competition in their business environment. Retail stores need to use all of available resources including data. Data processing is expected to be able to provide information that can be used to support marketing strategies. One of the data processing methods that are often used in marketing strategies is the use of data mining techniques i.e Market Basket Analysis using a priori algorithm. The application is designed using the waterfall method which starts from analyzing user needs, designing a process using UML which consists of: Use Case Diagrams, Activity Diagrams and Sequence Diagrams. This Market Basket Analysis application was built using the PHP programming language. From the results of the analysis in this research, it can be concluded that for the combination of 2 items with the highest confidence value 100% and the lift ratio value 3.39 i.e if a consumer buys a Laptop Charger, he will also buy a keyboard and for a combination of 3 items with the highest confidence value 100% and the lift ratio value 2.17 i.e if a consumer buys a Joystick and Laptop, he will also buy a mouse.

1. Introduction
Mizan Computer Store is a shop engaged in the trading sector, especially in the field of selling computers and supporting accessories. Mizan Computer Store is located in Sumber, Cirebon Regency. Mizan computer shop started its business in 2004 which initially only focused on printing. Based on the results of interviews conducted with the owner of the Mizan computer store. In 2008, he expanding his business by creating a shop that sells computers and other supporting accessories. In 2010, computer sales at the Mizan computer shop increased, this is because computers are currently a basic requirement for the community both for education and business activities. However, choosing the right computer equipment according to your needs and financial budget is not an easy thing. The large number of choices available on the market make potential buyers confused in choosing an existing computer device.

Growing and increasing business actors in the computer sector, can makes the players challenged to be able to create a unique differentiation and clear positioning so that consumers can differentiate
from their competitors. Competitive and dynamic market conditions make every company should always observe competition in their business environment. In facing an increasingly strong and fierce competitive environment, every company is required to be able to formulate a series of effective marketing strategies and always develop these marketing strategies continuously and sustainably. This is done as an effort to gain a competitive advantage against competitors.

Facing this competition, Mizan Computer Retail Store must be able to determine a more selective marketing strategy. In determining a marketing strategy, retail stores need to take advantage from all available resources, including data. Data processing is expected to be able to provide information that can later be used to support marketing strategies. One of the data processing methods that are often used in marketing strategies is the use of data mining techniques i.e Market Basket Analysis.

Based on the explanation above, it is necessary to plan a marketing strategy for Mizan Computer Retail Store in the face of increasingly fierce business competition. Mizan Computer Cirebon Retail Store has been using computers to record sales since 2017 so very large data has accumulated without being processed to analyze the sales. For this reason, it is necessary to develop an application that is able to implement Market Basket Analysis and the Apriori Algorithm for product purchasing patterns at the Cirebon Computer Mizan Retail Store so that can be used as an input in developing a marketing strategy.

2. Method

2.1. Database

Database is a collection of data that is logically related to each other, and also a description of these data, which is designed to present the information needed by an organization. In the database, there are three important terms, i.e entities, attributes, and relationships. An entity is a different object (can be a person, place, thing, concept, or event) in an organization that must be represented in a database. Attribute is a property that describes several aspects of the object to be recorded. Relationship is an association between entities [1].

2.2. Data Mining

Data mining is a computational process or what we usually call “calculations” to find interesting patterns in a large dataset by involving methods that are directly related to machine learning, statistics, and also database systems [2]. In KDD, data mining is one of the stages of the whole process that has the role for extracting interesting patterns from data by applying certain techniques and algorithms from the branches of machine learning, pattern recognition and statistics [3].

a. Association Rules

The rules of association (association rule) is one of the main techniques in data mining and its the form most commonly used in finding a pattern or patterns from a data set. Association rules were first introduced in 1993 by Agrawal which aim to extract interesting correlations, frequent patterns, associations or causal structures between sets of items in a transaction database, or other data repositories [4]. The association rule can also be defined as a process to find all associative rules that meet the minimum support requirements which can be interpreted as a supporting value indicating how often an item appears in the database, and a minimum confidence requirement that indicates the number of times an item is found together with the combination. other items at the same time.

The association rule will produce a pattern that consists of 2 (two) parts, namely the antecedent which is declared as if and consequent which is declared by then. If represents the item found in the data, then represents the item found in combination with the if item. The general form of association rules is [5]:

\[ A \rightarrow B \]

where:

\[ A = \text{antecedent if}, \text{and} \]
\[ B = \text{consequent then} \]
so that it can be concluded that if x buys A, then x will also buy B.

Lift is a ratio number that shows how many chances of finding an attribute appearing together with other attributes. The lift shows the power level of the rule for the events of the antecedent and consequent based on their respective supports. Calculation of lift can be seen in the following equation:

\[ \text{Lift Ratio} = \frac{\text{confidence}}{\text{Benchmark Confidence}} \]

\[ \text{Benchmark Confidence} = \frac{\text{transactions with the item in consequent}}{\text{transactions in database}} \]

A lift ratio value more than or equal to 1 indicates a benefit from this rule. Higher the lift ratio value, greater the strength of the association [6].

b. Market Basket Analysis

Association Rule also called with Market Basket Analysis because of its usefulness in modeling consumer buying behavior and analyzing consumer behavior in buying an item. Market Basket Analysis is a procedure in market basket analysis to find association rules that meet the minimum requirements for support and confidence (minimum support and minimum confidence) [7]. Market basket analysis is a data mining technique to find associative rules between a combination of items. The association rule algorithm will use training data, in accordance with the definition of Data Mining, to generate knowledge [8].

Market Basket Analysis is a method that aims to find patterns that often appear between many transactions, where each transaction consists of several items so this method supports the recommendation system through finding patterns between items in the transactions that occur. The basic methodology of market basket analysis is divided into 2 stages [9]:

1. High frequency pattern analysis.
   This stage looks for a combination of items that meet the minimum requirements of the support value in the database. The support value of an item is obtained by the following formula:

   \[ \text{Support}(A) = \frac{\text{The transactions amount contains } A}{\text{Total Transactions}} \]  

   while the support value of 2 items obtained the following formula :

   \[ \text{Support}(A, B) = \frac{\text{The transactions amount contains } A \text{ and } B}{\text{Total Transactions}} \]

2. Establishment of associative rules.
   After all high frequency patterns are found, then look for an associative rule that meets the minimum requirements for confidence by calculating the confidence of the associative rule A B The confidence value of the A_B rule is obtained by the following formula:

   \[ \text{Confidence } P (B|A) = \frac{\text{Number of transactions containing } A \text{ and } B}{\text{Number of transactions containing } A} \] 

   c. Apriori Algorithm

The Apriori algorithm is an algorithm that was introduced by R. Agrawal and R. Srikant in 1994. The apriori algorithm aims to find frequent itemsets that run on a set of data. In the k-iteration you will find all itemsets that have k-items, which are called k-itemset. The main characteristic of the apriori algorithm is that all subsets of frequent itemsets are also members of frequent itemsets [10].
Apriori algorithm is also a data mining technique to find associative rules between a combination of items. Usually used for the analysis of the purchase of goods in supermarkets with the intention to find out how likely a consumer is to buy one type of item together with other types of goods, so that companies can organize promotional strategies and placement of related goods placed close together and determine a promotional pricing strategy for certain items that are interconnected [11].

The apriori algorithm is divided into several stages called narration or pass [12].

1. The formation of the itemset candidate, the k-itemset candidate is formed from the combination (k-1) - the itemset obtained from the previous iteration. One way of the apriori algorithm is the trimming of k-itemset candidates whose subset containing k-1 items is not included in the high frequency pattern with length k-1.
2. Support calculation for each k-itemset candidate. Support from each k-itemset candidate is obtained by scanning the database to calculate the number of transactions that contain all the items in the k-itemset candidate. This is also a feature of the apriori algorithm where it is necessary to calculate the entire database as many as the longest k-itemset.
3. Set high frequency pattern. High frequency pattern that contains k-items or k-itemset that support is greater than the minimum support.

If no high frequency pattern is found then the whole process is stopped. If it is obtained again, then k is added by one and returns to point 1.

2.3. Research methods
In this study, we use a research systematics as follows:
   a. Literature study
   b. Collecting data
   c. Need analysis
   d. Design process
   e. Coding and software testing
   f. Analysis of result and discussion

3. Result and discussion
3.1. Data Collection
The author conducts an interview which aims to obtain data which are known only by the data source or the interviewee. The interview was conducted with Mr. Mizan as the owner of the Mizan Computer Cirebon Retail Store to obtain the data needed to design the system to be made. The data collected originated from Mizan Komputer Cirebon Retail Store. The data collected is the data on the number of transactions and stock of goods from January 2019 to March 2019.

3.2. Needs Analysis
The needs analysis in this study is divided into three parts, namely analysis of user needs, analysis of software needs, and analysis of hardware needs.
   a. Analysis of User Needs
      1. Users possessing the ability to operate PC
      2. Users used to using Windows operation system
   b. Analysis of Software Needs
      1. Sublime text, as a text editor with PHP programming language to make an application
      2. Microsoft Excel 2010, as a software in preprocessing stages
c. Analysis of Hardware Needs

1. Processor: Intel® Core™ i5-6400 CPU @ 2.70GHz
2. Memori/RAM: 16 GB
3. Storage: 120GB SDD dan 1 TB HDD
4. VGA: NVIDIA GeForce GTX 1070
5. Sistem Operasi: Windows 10 Education
6. Tipe Sistem: 64-bit Operating System

3.3. Preprocessing
Preprocessing is a basic operation which involves making strategy, which is appropriate with the dataset, to deal with the missing value and noise. The author’s strategy that can be carried out is to obtain the product data which is often purchased simultaneously in the Mizan Computer Retail Store because the author aims to find the rules occurring that will be used in this study. The following are the types of products which will be used:

1. Laptop
2. Laptop charger
3. Flash disk
4. Printer Ink
5. Cartridge
6. Printer
7. Router
8. LAN Cable
9. HVS Paper
10. Mousepad
11. Mouse
12. Headphone
13. Joystick
14. Keyboard
15. Hardisk

3.4. Implementation
This application is made using the PHP programming language so the user must access this application via a web browser to run it. As explained in Chapter III, implementation is carried out by following several steps to solve the problem using the association method and apriori algorithm. The data used is transaction data in the Mizan Computer Retail Store.

| Date           | Goods                                      |
|----------------|--------------------------------------------|
| 2019-01-01     | External Hard Drive, HVS Paper, Router, Lan Cable |
| 2019-01-02     | Laptop, Flashdisk, Mousepad, Mouse, Headphones |
| 2019-01-02     | Flashdisk, HVS Paper, Printer              |
| 2019-01-02     | Printer Ink, Printer, HVS Paper            |
| 2019-01-02     | Keyboard, Joystick, External Hard Drive, Mouse |
| 2019-01-02     | HVS Paper, Cartridges, Printer Ink         |
| 2019-01-03     | Laptop, Printer, Joystick, Mouse           |
| 2019-01-03     | Router, Lan Cable, Printer                 |
Table 1 is the transaction data in the Mizan Computer Retail Store that the author has selected to remove dirty data in order to get the desired rules. From this data, calculations will be carried out using the apriori algorithm so the purchasing patterns of the products can be seen. The first interface that will appear when running this application is the login display. The login appearance on this system is as follows:

| Date       | Goods                                                   |
|------------|---------------------------------------------------------|
| 2019-01-03 | Laptop Charger, External Hard Drive, Mouse, Keyboard    |
| 2019-01-03 | Headphone, Flashdisk, Mousepad                          |
| 2019-01-04 | Lan Cables, Routers, Laptops, Printers                  |
| 2019-01-04 | Flashdisk, Printer Ink, Printer                         |
| 2019-01-04 | Mousepad, Mouse, Headphone, Keyboard                    |
| 2019-01-04 | External Hard Drive, Flashdisk, Printer, Printer Ink    |
| 2019-01-05 | Joystick, Mousepad, Mouse, Laptop, Flashdisk            |
| 2019-01-05 | HVS Paper, Printer, Flashdisk                           |
| 2019-01-05 | HVS Paper, Printer Ink, External Hard Drive             |
| 2019-01-05 | Laptop Charger, Keyboard, Joystick                      |
| 2019-01-05 | Router, Lan Cable, Printer                              |
| 2019-01-05 | Keyboard, Laptop Chargers, Printer Ink, Printers        |
| 2019-01-06 | Keyboard, Router, Lan Cable, External Hard Drive        |
| 2019-01-06 | Laptop, External Hard Drive, Flashdisk                  |
| 2019-01-06 | Router, Printer, Printer Ink                           |
| ...        | ...                                                     |
| 2019-01-30 | External Hard Drive, HVS Paper, Router, Lan Cable       |

Figure 1 shows the login page of the application that has been created, this page contains the username and password for the user to fill in so they can enter the next page.
Figure 2 shows the main page of the Apriori Algorithm Application. Furthermore, the system has been created uses apriori algorithm. Users will see some navigation or pages contained in this system. The navigation menu is made in accordance with the use of the priori algorithm application itself, the first is the navigation on the system consists of Transaction Data, then navigation of the Priory Process, Rule Results, and Logout. The first step a user can take is to view the Transaction Data page to see all transaction data that has been imported or added by the user. The display of the Transaction Data page can be seen in Figure 3 as follows.

This page aims in the program to display the data that has been entered by importing data from Excel, then click the Upload Data button. The next navigation is the Apriori Process page, this page will display the steps to perform apriori algorithm calculation process by selecting the time range you want to determine to select the desired data, then the user enters the minimum support and minimum confidence values desired. Below is a display that displays the Priori Process Page.
Figure 4. Apriori process display

The next page on the system is the Rule Results page, on this page the user can see the Log Process which contains information about the results that we have previously processed on the Priory Process page.

Figure 5. Display of rule results

System testing is done to determine the trial process that is run on the application market basket analysis with the association rule method using apriori algorithm. In this study, a discussion of the testing has been carried out and an evaluation of the results issued by the system. From this test, it will be known the rules that are formed. This test is performed on 200 transaction data by entering the initial values and we have the transaction data as shown in Table 2.

| Absolute minimum support | 30 |
|--------------------------|----|
| Minimum Confidence       | 75%|

Table 2. Transaction data

| Date     | Goods                                      |
|----------|--------------------------------------------|
| 2019-01-01 | External Hard Drive, HVS Paper, Router, Lan Cable |
| 2019-01-02 | Laptop, Flashdisk, Mousepad, Mouse, Headphones |
| 2019-01-02 | Flashdisk, HVS Paper, Printer               |
Table 2 is the transaction data used to perform calculations. The first step is to calculate the minimum support value:

$$Minimum \ support = \frac{Absolute \ Minimum \ Support}{Total \ Transactions} \times 100\% = \frac{30}{200} \times 100\% = 15\% \quad (4)$$

After getting the minimum support value, then calculate the support value of each item, if the support value is greater than the specified minimum support. Then the item passes to follow the next calculation process.

$$Support = \frac{Number \ of \ item}{Total \ transactions} \times 100\% \quad (5)$$

| No | Item                              | Total | Support % |
|----|-----------------------------------|-------|-----------|
| 1  | Laptop                            | 53    | 26.5      |
| 2  | Joystick                          | 36    | 18        |
| 3  | External Hard Drive               | 64    | 32        |
| 4  | Mousepad                          | 35    | 17.5      |
| 5  | Laptop Charger                    | 41    | 20.5      |
Table 3 is itemset 1 which has a support value greater than the predetermined minimum support value. The next calculation is to calculate the support value of the two items and also the confidence value to see how much consumer confidence in buying goods that have a combination of 2 items in 1 transaction.

\[
\text{Support } (A \cup B) = \frac{\text{The transactions amount contains } A \text{ and } B}{\text{Total Transactions}} \times 100\% \quad (6)
\]

\[
\text{Confidence } P(A) = \frac{\text{The transactions amount contains } A \text{ and } B}{\text{Total Transactions amount contains } A} \times 100\% \quad (7)
\]

**Table 4. Support itemset 2**

| No | Item 1   | Item 2           | Total | Support % |
|----|----------|------------------|-------|-----------|
| 1  | Flashdisk| Mouse            | 43    | 21.5      |
| 2  | Flashdisk| External Hard Drive | 33    | 16.5      |
| 3  | Mouse    | Joystick         | 35    | 17.5      |
| 4  | Flashdisk| Printer Ink      | 39    | 19.5      |
| 5  | Mouse    | Mousepad         | 30    | 15        |
| 6  | Laptop   | Joystick         | 33    | 16.5      |
| 7  | Printer Ink | Printer     | 44    | 22        |
| 8  | Flashdisk| Printer          | 31    | 15.5      |
| 9  | Keyboard | Laptop Charger  | 41    | 20.5      |
| 10 | External Hard Drive | Printer | 30    | 15        |
| 11 | Laptop   | Mouse            | 35    | 17.5      |

**Table 5. Confidence of itemset 2**

| No | A=>B         | Support A U B | Support A | Confidence |
|----|--------------|---------------|-----------|------------|
| 1  | Joystick=>Laptop | 16.5         | 18        | 91.67      |
| 2  | Laptop=>Joystick   | 16.5         | 26.5      | 62.26      |
| 3  | Mouse=>Laptop     | 17.5         | 46        | 38.04      |
| 4  | Laptop=>Mouse     | 17.5         | 26.5      | 66.04      |
| 5  | Flashdisk=>Printer Ink | 19.5   | 46        | 48.75      |
| 6  | Printer Ink=>Flashdisk | 19.5 | 30        | 65        |
| 7  | Flashdisk=>Mouse | 21.5         | 40        | 53.75      |
| 8  | Mouse=>Flashdisk  | 21.5         | 46        | 46.74      |
| 9  | Flashdisk=>External Hard Drive | 16.5  | 40        | 41.25      |
| 10 | External Hard Drive => Flashdisk | 16.5  | 32        | 51.56      |
| 11 | Flashdisk=>Printer | 15.5       | 30        | 38.75      |
Table 5 shows that the support value of the two items is greater than the predetermined minimum support value is 11 transactions, from table 4,5 we can see that it is the confidence value of a combination of 11 existing transactions. Then the calculation process that will be carried out is to calculate the lift test value to help see the amount of consumer confidence in buying the product and look for the support value and the confidence value of the 3 existing itemset. The following is the formula for finding the Lift Ratio value:

\[
\text{Lift Ratio} = \frac{\text{Number of Items Appear}}{\text{Number of Transactions}} \times \frac{\text{Number of appearances A, Number of appearances B}}{\text{Number of appearances A} + \text{Number of appearances B}}
\]

\[ (8) \]

| Itemset | Support Value | Confidence Value |
|---------|---------------|------------------|
| 1       | Laptop, Mouse 16.5 | 62.62           |
| 2       | Laptop, Mouse 16.5 | 38.04           |
| 3       | Laptop, Mouse 16.5 | 91.67           |
| 4       | Laptop, Mouse 16.5 | 100             |
| 5       | Laptop, Mouse 16.5 | 94.29           |
| 6       | Laptop, Mouse 16.5 | 94.29           |

Table 6. Support itemset 3

| No | Item 1 | Item 2 | Item 3 | Total | Support |
|----|--------|--------|--------|-------|---------|
| 1  | Laptop | Mouse | Joystick | 33 | 16.5    |

Table 7. Confidence itemset 3

| No | A=>B | Support A U B | Support A | Confidence |
|----|------|---------------|-----------|------------|
| 1  | Joystick=>Mouse, Laptop | 16.5 | 18 | 91.67 |
| 2  | Mouse=>Laptop, Joystick | 16.5 | 46 | 35.87 |
| 3  | Laptop=>Joystick, Mouse | 16.5 | 26.5 | 62.26 |
| 4  | Joystick, Laptop=>Mouse | 16.5 | 16.5 | 100 |
| 5  | Mouse, Joystick=>Laptop | 16.5 | 17.5 | 94.29 |
| 6  | Laptop, Mouse=>Joystick | 16.5 | 17.5 | 94.29 |
Table 7 shows that the support value of the two items is greater than the predetermined minimum support value is 1 transaction, from table 7 it can be seen that it is the confidence value of the existing transaction combination. It obtained, the results of rules that occur when using 200 transactions with a minimum absolute support of 30 and a minimum confidence of 75% as below.

Table 7. Final results

| No | A=>B                        | Confidence | Lift Ratio |
|----|-----------------------------|------------|------------|
| 1  | Joystick => Mouse, Laptop   | 91.67      | 5.24       |
| 2  | Joystick, Laptop => Mouse   | 100        | 2.17       |
| 3  | Mouse, Joystick => Laptop   | 94.29      | 3.56       |
| 4  | Laptop, Mouse => Joystick   | 94.29      | 5.24       |
| 5  | Joystick => Laptop          | 91.67      | 3.46       |
| 6  | Laptop Charger => Keyboard  | 100        | 3.39       |
| 7  | Mousepad => Mouse           | 85.71      | 1.86       |
| 8  | Joystick => Mouse           | 97.22      | 2.11       |

Table 8 is the final result of the calculations that have been carried out. The results for the combination of 2 items with the highest confidence value of 100% and the lift ratio value of 3.39, if a consumer buys a Laptop Charger, he will also buy a keyboard and for a combination of 3 items with the highest confidence value of 100 % and the value of the lift ratio is 2.17, if a consumer buys a joystick and laptop, he will also buy a mouse.

4. Conclusion
The priori algorithm can analyze transaction data at Mizan Computer Retail Store by finding association rules that meet the minimum support and minimum confidence requirements, based on the items in the transaction. The application is designed using the waterfall method which starts from analyzing user needs, designing a process using UML which consists of: Use Case Diagrams, Activity Diagrams and Sequence Diagrams.

This Market Basket Analysis application was built using the PHP programming language. There are 7 interface pages, namely, the login page, the main page, the transaction data page, the priori process page, the results page, the results process log page and the view rule page.

From the results of the analysis in this research, it can be concluded that for the combination of 2 items with the highest confidence value of 100% and the lift ratio value of 3.39, if a consumer buys a Laptop Charger, he will also buy a keyboard and for a combination of 3 items with The highest confidence value is 100% and the lift ratio value is 2.17, if a consumer buys a joystick and Laptop, he will also buy a mouse.

Acknowledgments
We thanks to Rector Universitas Padjadjaran. Financial support was received from ALG Unpad contract: 1427/UN6.3.1/LT/2020.

References
[1] Connolly T and Begg C 2015 Database Systems ‘A Practical Approach to Design, Implementation, and Management (Paisley University: Pearson)
[2] Clifton C 2014 Data Mining Encyclopedia Britannica [Online] Available: http://www.britannica.com/technology/data-mining [Accessed 24 April 2016]
[3] Fayyad U, Piatetsky-Shapiro and Smyth, 1996 From Data Mining to Knowledge Discovery in Database in AI Magazine p. 37-54
[4] Kantardzic M 2011 *Data Mining: Concepts, Models, Methods, and Algorithms* (USA: Wiley)
[5] Zaki M 1997 Fast Mining of Sequential Patterns in Very Large Databases, Semantic Scholar
[6] Santosa B 2007 *Data Mining: Data Utilization Techniques for Business Purposes* (Yogyakarta: Graha Ilmu)
[7] Jananto A 2012 The use of market basket analysis to determine student competency patterns
*DYNAMIC Journal of Information Technology* 17
[8] Masnur A 2015 Data mining analysis using market basket analysis to determine consumer purchase patterns *Science and Information Technology* 1(2)
[9] Fadlina 2014 Data mining untuk analisa tingkat kejahatan jalanan dengan algoritma association rule metode apriori *Scientific Information and Technology* (INTI)
[10] Haryati R, Natalis and Purnamasari Y 2017 The application of the CBA Method on Association Rule using the Priori Algorithm for classification of ARI *semanTIK* 3 31-38
[11] Informaticalogy 2017 Apriori Algorithm (Association Rule) [Online] Available: https://informatikalagi.com/algorith-apriori-association-rule/
[12] Setiawati D D 2009 Using the Priori Method for market basket analysis on minimarket sales transaction data using PHP & MySQL *Jurnal Informatika Merdeka Pasuruan* 3(2)
[13] Han J and Kamber M 2006 *Data Mining: Concepts and Techniques*, 2nd Ed (San Francisco: Morgan Kaufmann Publisher)