Data Mining to Determine Correlation of Purchasing Cosmetics With A priori Method

Dahlan Abdullah1*, A M H Pardede2, Eli Cahayati2, Akbar Iskandar3, AyuEsteka Sari4, Muhammad Arifin5, HerlinaLatipa Sari6, Ida Nuryana7, NurintanAsyiah Siregar8, Sara Surya9, Lilian G F Apituley10, Eni Wuryani11, ArlisDewi Kuraesin12, Cut Ita Erliana13 and Didik Setiyadi14

1Department of Informatics, UniversitasMalikussaleh, Aceh, Indonesia
2STMIK Kaputama, Binjai, Indonesia
3Department of Informatics, STMIK AKBA, Makassar, Indonesia
4Department of Management, STIE Sakti AlamKerinci, Jambi, Indonesia
5Department of Information System, Faculty of Engineering, UniversitasMuria Kudus, Indonesia
6Departement of Information Technology, UniversitasDehasen Bengkulu, Indonesia
7Department of Management, Faculty of Economic, UniversitasKanjuruhan, Malang, Indonesia
8Department of Management, STIE Labuhanbatu, Sumatera Utara, Indonesia
9Department of Pharmacy, University of Dharma Andalas, Padang, Indonesia
10Departement of Law, Halmahera University, Maluku Utara
11Faculty of Economics, Universitas Negeri Surabaya, Indonesia
12STIE Muhammadiyah Jakarta, Indonesia
13Department of Industrial Engineering, UniversitasMalikussaleh, Aceh, Indonesia
14Department of Informatics, STMIK Bina Insani, Indonesia

*dahlan@unimal.ac.id

Abstract. Data mining is the process of analyzing data using software to find patterns and rules in the data set. Data mining can analyze large data to find knowledge to support decision making. In this study the Rule Association will be discussed as one of the data mining functions implemented using the A priori Algorithm. There will also be analyzed two support calculation techniques in candidate generation in the A priori Algorithm, such as: K-way and 2 Group-By in three sample datasets with transaction attributes id and item. In this study the problem of support calculation in candidate generation is the bottleneck of the A priori Algorithm where the improvement of the A priori Algorithm was emphasized on the candidate generation and the effectiveness of the A priori Algorithm. This research was lead on the Oracle RDBMS by utilizing WEKA tools to determine maximum support and confidence and to find out the correlation between products. The results shows the highest confidence value at 93% if you buy DeepClensingMilk and DeepClensingToner then you will buy Whitening Soap with confidence = 93%.

1. Introduction
Data mining is a process to supporting decision makers to find patterns of information in data. This explorer can be done by the user, for example by using a query or can be assisted with an application that automatically searches for information patterns in the database. This search is called discovery [1]. Discovery is the process of searching in a database to find hidden patterns without the ideas that
were previously obtained or hypotheses about existing patterns. In other words, the application takes the initiative to find patterns in the data without the user thinking about the relevant questions first. One form of pattern that data mining can produce is the association rule. The Association Rule can be used to find a relationship or cause and effect [2].

Function of Association Rules are often referred as "market basket analysis", which is used to find relationships or correlations between the set of items. Market Basket Analysis is an analysis of buying habits by looking for associations and correlations between different items placed by customers in their shopping charts. This function is most widely used to analyze data in the context of marketing strategy needs, catalog design, and business decision-making processes [3].

In this study used a comparison of correlation with a priori algorithm where a priori algorithm can be used to find frequent-itemset which is an iteration in the data. In the analysis with a priori algorithm, the writer is assisted with WEKA tools. WEKA is a practical package of machine learning tools.

"WEKA" is Waikato Environment for Knowledge Analysis, was created at the University of Waikato, New Zealand for research, education and various applications. WEKA is able to solve data mining problems in the real world, especially the classification that underlies machine learning approaches. This software is written in the Java class hierarchy with object-oriented methods and can run on almost all platforms [4].

WEKA is easy to use and applied on several different levels. Available implementation of state-of-the-art learning algorithms that can be applied to datasets from the command line. WEKA contains tools for pre-processing data, classification, regression, clustering, association rules, and visualization. Users can pre-process data, include it in a learning scheme, and analyze the resulting classifier and its performance - all without writing the program code at all. An example of using WEKA to apply a learning method to the dataset and analyze the results to obtain information about the data, or apply several methods and compare their performance to be selected[4].

2. Related Works
The Apriori algorithm includes the types of association rules in data mining. Besides a priori, which is included in this class is the generalized rule induction method and hash based algorithm. The rules in state associations between several attributes are often called affinity analysis or market basketball analysis. Association analysis or association rule mining is a data mining technique to find associative rules between a combination of items. Examples of associative rules from analyzing purchases in a supermarket are knowing how likely a customer is to buy bread along with milk. With this knowledge, supermarket owners can arrange the placement of their goods or design marketing campaigns using discount coupons for certain items [5][6].

3. Research Methodology
A system analysis can be defined as the decomposition of a complete information system into its component parts with a view to identifying and evaluating problems, opportunities, obstacles that occur and the expected needs so the improvements can be proposed. Or more easily, system analysis is a research on an existing system with the aim to designing a new or updated system. The system analysis phase is a very critical and very important stage, because errors in this stage will also cause errors in the next stage. The main task of system analysis at this stage is find weaknesses in the current system so that improvements can be proposed[7].

The feasibility analysis of the system must fill a various aspects, including technological feasibility, legal feasibility, operational feasibility, and economic feasibility. Each must be fulfilled well to find out if a system is indeed feasible and needs to be designed and applied in the object of research. Before the transactional data is entered into WEKA as the tools used in this study, it is necessary to note what purchase transactions will be included in the item combination[8][9]

| Transaction | Purchased Item |
|-------------|----------------|
| 1           | Dcm,Dct,Ws,Es,Bo,Fw,Twc,Wc,Hbl,Dc |
| No | Item                        | Code |
|----|-----------------------------|------|
| 1  | Deep Clensing Milk          | DCM  |
| 2  | Deep Clensing Toner         | DCT  |
| 3  | Whitening Soap              | WS   |
| 4  | Face Cream                  | FC   |
| 5  | Eye Cream                   | EC   |
| 6  | Serum C                     | SC   |
| 7  | Eye Shadow                  | ES   |
| 8  | Blush On                    | BO   |
| 9  | Feminim Wash                | FW   |

Table 2. Representation of Transaction Data in Transactional Database
Basic methodology Association analysis is divided into two stages:

1. **Analysis of High Frequency Patterns**
   
   This stage find out for item combinations that qualified with the minimum requirements of the support value in the database.
   
   The rule used is **if x then y**, where x is antecedent and y is consequent. Based on the rule, it takes 2 items which one of them as antecedent and the rest as consequent. From the steps above, 1 Fk is obtained, namely F2. F1 is not included because it only consists of 1 item. For antecedents may be more than 1 element, while for consequent consists of 1 element.

2. **Forming of Association Rules**
   
   After all high frequencies have been found, the association rules are found that qualified with the minimum confidence requirements by calculating associative rules.

Calculate support and confidence.

\[
\text{SUPPORT} = \frac{\sum \text{Transactions purchased at once}}{\sum \text{Total Transactions}} \times 100\%
\]

\[
\text{CONFIDENCE} = \frac{\text{Transactions purchased at once}}{\sum \text{Number of Transactions in Antecedent Part}} \times 100\%
\]

For Σ items purchased at once if buy Athen buy B, there are 15 transactions. The total number of transactions is 31 transactions, so the support is:

\[
\text{SUPPORT} = \frac{15}{31} \times 100\% = 48\%
\]

For Σ items purchased at the same time If buy A then buy B, there are 3 transactions, while the number of transactions that buy A is 4 transactions, so the confidence is:

\[
\text{CONFIDENCE} = \frac{15}{20} \times 100\% = 75\%
\]

**Table 3. Final Association Rules**

| If antecedent then consequent | Support | Confidence | Support x confidence |
|-------------------------------|---------|------------|---------------------|
| If buy Dcm then buy Dct       | 48%     | 75%        | 0.36                |
| If buy Dcm then buy Ws        | 48%     | 75%        | 0.36                |
| If buy Dcm then buy Fw        | 48%     | 75%        | 0.36                |
| If buy Dcm then buy Wc        | 48%     | 75%        | 0.36                |
| If buy Ec then buy Dct        | 48%     | 75%        | 0.36                |
| If buy Bo then buy Dct        | 48%     | 71%        | 0.3408              |
| If buy Ec then buy Ws         | 48%     | 75%        | 0.36                |
| If buy ES then buy Ws         | 48%     | 75%        | 0.36                |
| If buy Bo then buy Ws         | 48%     | 71%        | 0.3408              |
| If buy Ec then buy Twc        | 48%     | 75%        | 0.36                |
4. Result and Discussion

Data Mining to determine the correlation of cosmetic purchase with Algorithm A Priori method of data processing made with WEKA and for inputting cosmetic purchase transactions using PHP and Database programming languages using MySQL. Shopping cart analysis is an analysis of data mining in cosmetic purchase transactions. Analysis of the shopping cart data can be used to get association rules from transaction data.

Table 4. Transaction Data of PureGlow

| Invoice Number | Item                | Total |
|----------------|---------------------|-------|
| 170210         | Dcm,Det,Ws,Es      | 4     |
| 170211         | Bo,Fw,Twc,Wc,Hbl,De| 6     |
| 170212         | Dcm,Fc,Ec,Sc,Fw    | 5     |
| 170213         | Lip,Wc,Nc,De       | 4     |
| 170214         | Det,Ws,Fc,Ec,Es,Bo| 6     |
| 170215         | Fw,Lip,Twc,Hbl,Nc  | 5     |
| 170216         | Dcm,Det,Ws,Ec     | 4     |
| 170217         | Es,Bo,Twc         | 3     |
| 170218         | Wc,Hbl,De         | 3     |
| 170219         | Dcm,Det,Ws,Fe,Ec  | 5     |
| 170220         | Sc,Es,Bo,Fw,Lip   | 5     |
| 170221         | Twc,Wc,Hbl,Nc,De  | 5     |
| 170222         | Dcm,Ws,Fc,Ec      | 4     |
| 170223         | Sc,Bo,Fw,Twc,Wc,Nc| 6     |
| 170224         | Dcm,Det,Ws,Ec,Sc,Es| 6    |
| 170225         | Bo,Lip,Wc,De      | 4     |
| 170226         | Dcm,Det,Ws,Fe     | 4     |
| 170227         | Ec,Bo,Fw,Twc,We   | 5     |
| 170228         | Hbl,Nc            | 2     |
| 170229         | Det,Fc,Ec,Es,Bo   | 5     |
| 170230         | Fw,Lip,Twc,Hbl,Nc | 5     |
| 170231         | Ws,Sc             | 2     |
| 170232         | Es,Bo,Fw,Lip,Hbl,De| 6    |
| 170233         | Dcm,Det,Ws,Sc,Fw,Lip| 6  |
| 170234         | Twc,Wc,Nc,De      | 4     |
| 170235         | Ws,Fc,Es          | 3     |
| 170236         | Bo,Fw,Lip,Hbl     | 4     |
| 170237         | Det,Fc,Bo,Fw      | 4     |
| 170238         | Twc,Wc,Nc,De      | 4     |
| 170239         | Dcm,Det,Ws       | 3     |
| 170240         | Fc,Ec,Sc,Es,Fw,Wc | 6     |
| 170241         | Hbl,De            | 2     |
| 170242         | Dcm,Ws,Ec,Sc,Bo,Twc| 6    |
| 170243         | Nc,De             | 2     |
| 170244         | Dcm,Ws,Fc        | 3     |
| 170245         | Sc,Es,Fw,Lip     | 4     |
After inputting the new transaction data into the database and matching the rules that qualified, user asked to enter transaction data, such as: id_transaction, name_transaction and transaction amount. Trial to matching transaction input with the association rule has been analyzed previously.

In the system, data cleaning to dispose of inconsistent data and integration data noise to combining data from multiple transformation sources data. Data was converted into a suitable form for mining DM evaluation technique application patterns to find interesting / valuable knowledge presentations with visualization techniques A priori algorithm will be used when analyzing data. When entering transaction data that has through the pre-processing stage, then the a priori algorithm starts working on the two major stages, the Frequent Itemsets information process and the Rule Association formation process.

### PRE-PROCESSING

```
Start
```

Figure 1. Pre-Processing Flow Chart

The first is determine minimum support and confidence in the system. After determining the minimum support, the system will begin on searching for and forming Frequent Itemsets. Then from the Frequent Itemsets collection, each of the Frequent Itemsets will be searched for nonempty subsets for the confidence calculation process. After the confidence calculation process is complete, it will be selected rules, if do not match with confidence will be rejected. The system will take the library from WEKA to convey the discovery of these rules.

5. Conclusion

From the observations during the process and evaluation, it can be accepted as follows:
1. The highest confidence value at 93% if you buy DeepClensingMilk and DeepClensingToner then buy Whitening Soap with confidence = 93%.
2. With this method of A Priorti Provider, you can help other company of Pure Glow in stocking items.
3. Based on the results of the trial, it was concluded that the system would not find the rules if the confidence value was larger than the value of support.
4. With Data Mining to determine the cost of purchasing cosmetics using the A Priori method, the inputting of data on PureGlow Medan cosmetic will be faster, precise, accurate, easy and better.

References
[1] M. Karismariyanti, D. Darmantoro, and D. S. Kusumo, “Aturan Asosiasi Spasial pada Basis Data Spasial,” in Seminar Nasional Aplikasi Teknologi Informasi (SNATI), 2007.
[2] D. Haryanto, Y. Oslan, and D. Dwiyana, “Implementasi Analisis Keranjang Belanja Dengan Aturan Asosiasi Menggunakan Algoritma A Priori Pada Penjualan Suku Cadang Motor,” Jurnal Buana Informatika, vol. 2, no. 2, 2011.
[3] F. Fatihatul, A. Setiawan, and R. Rosadi, “Asosiasi Data Mining Menggunakan Algoritma FP-Growth untuk Market Basket Analysis,” Jatinagor: Universitas Padjadjaran, 2011.
[4] D. Purnamasari, J. Henharta, Y. P. Sasmita, F. Ihsani, and I. Wicaksana, “Get Easy Using WEKA,” Dapur Buku, 2013.
[5] E. T. L. Kusrini, “Algoritma Data Mining,” Yogyakarta: Andi Offset, 2009.
[6] T. Mantoro, C. W. Johnson, and M. A. Ayu, “A framework in ubiquitous computing environment for providing intelligent responses,” presented at the 3rd International Conference on Mobile Ubiquitous Computing, Systems, Services, and Technologies, UBICOMM 2009, 2009, pp. 289–294.

[7] D. Abdullah, Hartono, and A. Prasetya, “STOCK PARTS FORECASTING USING LEAST SQUARE IN PT. DUNIA BARUSA LHOKSEUMAWE,” ICOMSET, Oct. 2015.

[8] A. Arifin, D. Abdullah, S. Mohamed Berhan, and R. Budiarto, “An Economical IPv4-to-IPv6 Transition Model: A Case study for University Network,” Oct. 2018.

[9] D. Abdullah, Tulus, S. Suwilo, S. Effendi, and Hartono, “DEA Optimization with Neural Network in Benchmarking Process,” IOP Conf. Ser.: Mater. Sci. Eng., vol. 288, no. 1, p. 012041, 2018.