Implementation Of Data Mining On Glasses Sales Using The Apriori Algorithm

Primasatria Edastama[1], Ankur Singh Bist[2], Ari Prambudi[3]
University of Esa Unggul[1,3], Graphic Era Hill University[2]
Jl. Arjuna Utara No.9, Kb. Jeruk, Kota Jakarta Barat[1,3]
Road Society Area, Oghal Bhatta, Clement Town, Dehradun, Uttarakhand[2]
Indonesia[1,3], India[2]
e-mail: primasatria@esaunggul.ac.id[1] , ankur1990bist@gmail.com[2] ,
ari.pambudi@esaunggul.ac.id[3]

Abstract

The Optik Nasional is one of the most important optik in Tasikmalaya. Many of his everyday sales transactions, even longer sales, will increase as a result of his activities through the process of purchasing and selling glasses with diverse brands. If left to its own devices, sales transaction data devolves into useless junk. With the help of technological advancements, the ability to collect and process data is improving. Data mining is the process of extracting information and expertise from large amounts of data. Transaction data may be used and analysed to generate helpful information for improved sales and promotional goods, in addition to serving as an archive for the firm. The author utilises the Apriori algorithm to obtain information on the most popular and in-demand goods in a database transaction at a national optical firm, so that the results may be utilised to improve the growth of eyeglass sales and marketing.

Keywords : Apriori Algorithm, Data Mining, Sales.

1. Introduction

National Optics is one of Tasikmalaya's opticians. National Optics' operations include purchasing and selling glasses from a variety of brands; the number of daily sales transactions and sales data will continue to grow over time. If left unattended, sales transaction data will devolve into a useless waste. The ability to collect and process data is improving as a result of technological advancements. Data mining is the process of extracting information and expertise from large amounts of data. Data mining is a collection of procedures for determining the added value of a data set in the form of information not previously known from the data. In in the database world, data mining is also known as knowledge discovery (KDD). KKD is a

process that involves the gathering, analysis, and interpretation of data, as well as the use of history to uncover regularities and patterns in big datasets.

The growing amount of competition in the corporate world, particularly in firms that specialise on selling items, necessitates developers to devise a plan for expanding product sales and marketing, one of which is to use product sales data. With daily sales activity, the data will grow over time. The data may be used and turned into helpful information for improving sales and product promotion, in addition to serving as an archive for the firm.

The author utilises an a priori algorithm to extract information from a transaction database about the items that are most sold and in demand at a national optical firm, so that the results may subsequently be utilised to create sales and marketing of eyewear products.

The following is the problem formulation that will be used:

a. How can I use the Apriori algorithm to determine the most popular Glasses Brand's sales?
b. How can I use an a priori algorithm to evaluate Glasses Sales data and create a pattern of item combinations?
c. Is it possible for a priori algorithms to aid in the development of marketing strategies?

Limitations of the problem are:
a. To create rules, an a priori data mining method is used to discover patterns of combinations of itemsets and association rules.
b. This implementation is based on sales data for Glasses at National Optics for the previous year, which spans the 2016 calendar year.
c. Monthly sales data was used in this study.
d. Brand-based determination of the most popular items.

Data mining is a technique for automatically analysing and extracting knowledge that uses one or more computer learning techniques (machine learning). Induction-based learning, for example, is the process of establishing general concept definitions by watching specific examples of the ideas to be studied. The use of scientific techniques to data mining is known as Knowledge Discovery in Databases (KDD). Data mining is one phase of the KDD process in this context (Hermawati, 2013).

Data mining is described as the act of sifting very large data sets kept in storage using pattern recognition techniques such as statistical and mathematical approaches in order to discover new relevant correlations, patterns, and trends. (Emha taufiq Luthfi, Kusrini, 2009).
Aside from these meanings, there are a few others that are stated below (Kusrini, Emha taufiq Luthfi, 2009).
a. Data mining is a collection of procedures for extracting additional value from a data source in the form of previously unknown information.
b. Data mining is the automated examination of big or complicated data with the goal of identifying key patterns or trends that are often overlooked.
There are numerous phases to data mining. These stages are interactive, involving the user directly or via a knowledge base. Data mining is divided into six steps (Handoko, Suryadi, 2010):
a. Data cleaning
   The process of eliminating noise, inconsistencies, and unnecessary data is known as data cleaning. In general, data acquired from a company's database and experimental

findings contain flaws such as missing data, incorrect data, or a simple error. Additionally, there are data characteristics that have no bearing on the data mining concept. It's also a good idea to get rid of any data that isn't relevant. Because the amount and complexity of data handled will be decreased, data cleaning will have an impact on the performance of data mining techniques.

b.   Data integration
The combining of data from different databases into a single new database is known as integration data. Frequently, the data required for data mining originates from several databases or text files, rather than from a single database. Data integration is done on characteristics that identify distinct entities, such as names, product kinds, numbers, customers, and so on. Data integration must be done carefully since mistakes in integration data might create skewed findings and potentially lead to future action being taken in the wrong direction. For example, if data integration based on product type results in the combination of items from various categories, a link between products that do not exist will be discovered.

c.   Data Selection
Because not all of the data in the database is used, just the data that is acceptable for analysis will be extracted from the database. Consider a scenario in which people's predisposition variables are investigated.

The a priori algorithm is the most famous algorithm for finding high frequency patterns. A priori algorithm is divided into several stages called narration or pass (Setiawati, 2009).

a.   Candidate k-itemset is created from the combination (k-1)-itemset acquired in the previous phase. The subset comprising k-1 items does not fit into the high-frequency pattern of length k-1 in one way.

b.   Support from each candidate k-itemset is calculated. Each candidate k-support itemset is acquired by scanning the database and counting the number of transactions that include all of the candidate k-items. item sets This is also a characteristic of the a priori approach, which requires computing as many k-itemsets as possible using the full database..

c.   Create a pattern with a high frequency. The candidate k-itemsets whose support is larger than the minimal support are used to determine the high-frequency pattern comprising k items or k-itemsets. The procedure is terminated if no new high frequency patterns are discovered. If not, k is increased by one and part 1 is returned.

Agrawal & Srikant introduced the Apriori method in 1994 for determining frequent itemsets for boolean association rules. In data mining, the a priori algorithm is a sort of association rule. Affinity analysis or market basket analysis are terms used to describe rules that express the relationship between many qualities. The data mining approach of association analysis, also known as association rule mining, is used to discover the rules of a group of elements. High-frequency pattern mining is one of the stages of association analysis that has caught the attention of many academics in order to develop efficient algorithms. Two benchmarks, called support and confidence, can be used to assess the significance of a relationship. The proportion of the combination of these items in the database is called support (support value), while the strength of the link between items in the association rules is called confidence (certainty value).
Steps in the Apriori Algorithm:

a. High Frequency Pattern Analysis with Apriori Algorithm

This step looks for a combination of objects in the database that match the support value's minimal criteria. The following formula is used to calculate an item's support value:

Support A = $\frac{\sum \text{Transactions Containing A}}{\sum \text{Transaksi}} \times 100\%$

b. Combination of 2 Itemsets

While, the support value of 2 items is obtained using the formula:

$$\text{Support (A,B)} = P(A \cap B)$$

SupportA,B = $\frac{\sum \text{Transactions Containing A and B}}{\sum \text{Transaction}} \times 100\%$

The frequent itemset displays items with a frequency of occurrence greater than the provided minimum value (). If we set the value to 2, then any item sets whose frequency of occurrence is more than or equal to 2 times are considered common. Fk denotes the set of frequent k-itemsets.

c. Formation of Association Rules After you've discovered all of the high-frequency patterns, calculate the confidence of the associative rule A->B to find association rules that fulfil the minimal standards for confidence. The following formula is used to calculate the score confidence of the rule A->B::

Confidence = $\frac{\sum \text{Transactions Containing A and B}}{\sum \text{Transaction A}} \times 100\%$

The association rules must be ranked by Support Confidence to decide which ones should be chosen. As many as n rules with the biggest outcome are chosen..

Table 1. Previous Studies

No	Researcher	Years	Difference
1	Dewi Kartika Pane	2013	The use of the Apriori Algorithm in determining the most popular electronic items is highlighted. (Pane, 2013)
2	Nugroho Wandi, Rully A. Hendrawan dan Ahmad Mukhlason	2012	When a user views the contents of a selected book or wishes to borrow it, apriori techniques and algorithms produce book lending transactions with strong associations between books in transactions that are used as book lending suggestions to assist users receive additional book recommendations. Nugroho Wandi (Nugroho Wandi, 2012)
3	Denny Haryanto, Yetli Oslan, Djoni Dwiyana	2011	The Apriori Algorithm's usage in the search for attachment patterns for product promotion is intended to reduce the advertising of low-selling items. Consumers will be less upset by the advertising of items that have not been purchased if the promotion of commodities that do not have a pattern of attachment is minimised (Denny Haryanto, Yetli Oslan, and Djoni Dwiyana, 2011).
4	Hapsari Dita	2013	The Apriori Algorithm is used to create an application that

	Anggraeni, Ragil Saputra, S.Si,M.Cs, Beta Noranita, S.Si, M.Kom	analyses the market basket of medication sales transaction data using data mining, a data analysis approach that may assist pharmacies in gaining information in the form of sales trends over a given month. (Hapsari Dita Anggraeni, Ragil Saputra, S.Si, M.Cs, Beta Noranita, S.Si, M.Kom, Hapsari Dita Anggraeni, Ragil Saputra, S.Si, M.Cs, Hapsari Dita Anggraeni, Ragil Saputra, S.Si, M.Cs

2. Research Method

The most well-known technique for detecting high frequency patterns is the a priori algorithm. A priori is broken down into many parts, known as narrative and pass. Candidate k-itemsets are created by combining the (k-1)-itemsets acquired in the preceding round. Pruning of potential k-itemsets whose subsets comprising k-1 items are not included in the high-frequency pattern with length k-1 is one method of the a priori algorithm. For identifying high frequency patterns, the a priori method is well-known. Data analysis is the first step in creating a mix of itemset patterns and rules. The data utilised is data on glass sales transactions, which is then followed by the development of a combination of itemsets pattern, and finally, association rules are generated from an intriguing combination of itemsets pattern. In this situation, the Apriori algorithm is well suited to the task of analysing sales data in order to determine the most popular items, which will have a significant impact on the company's marketing efforts.

The researcher's research was done at an optical firm in Tasikmalaya, and the data utilised was 2016 sales data, which is evaluated every month to determine which brand of glasses sold the most in 2016, so that the results may be used as a marketing reference. The firm produces eyeglasses. The brands of eyeglasses available at Optik Nasional Tasikmalaya are listed below.

Table 2. 2016 National Optical Glasses Brand List

NO	GLASSES BRAND
1	ADIDAS
2	BILLABONG
3	BURBERRY
4	BVLGARI
5	CALVIN KLEIN
6	CARRERA
7	CARTIER
8	CELINE

9	CHANEL
10	D&G
11	DIOR
12	DUNHILL
13	EMPORIO ARMANI
14	FERRARI
15	GIORGIO ARMANI

The stages carried out during the Research

Literature Study and Problem Formulation

Grouping of Glasses Sales Data in National Optics in 2016

Creating Sales Transaction Patterns Glasses at National Optics 2016

Apriori Algorithm

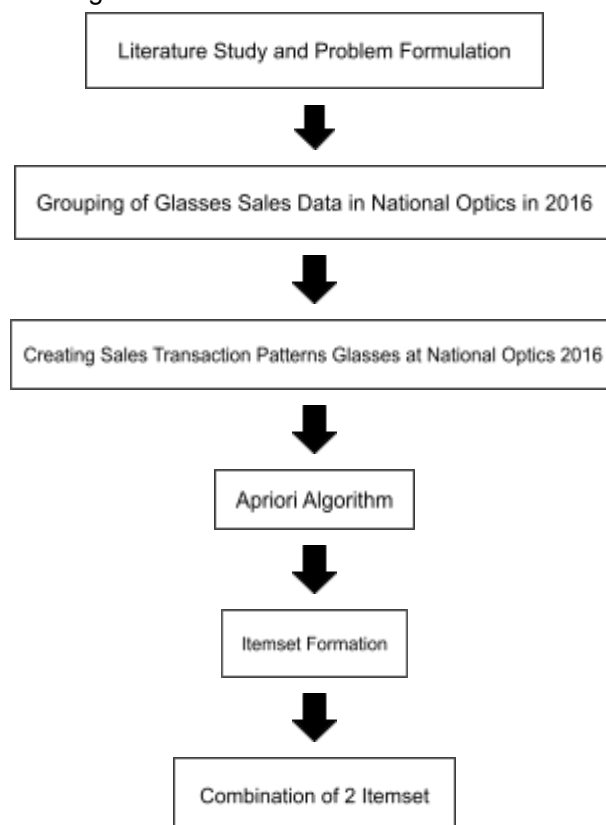Itemset Formation

Combination of 2 Itemset

Figure 1. Research Stages

3. Results and Discussion

The goal of the a priori method is to identify all a priori rules that satisfy the minimal support criteria, which are the combinations of each item in the database. And the minimal level of confidence (certainty value), which refers to the strength of the link between elements in the association rules.

These transactions might be gathered based on the transaction of selling Glasses to an Optics firm across the country. Table 3 shows the total number of sales transactions for Glasses at National Optics that were utilised as objects over the course of a year in 2016.

Table 3. Pattern of Glasses Sales Transactions in 2016

MONTH	ITEM SET
1	FERRARI,CARTIER,RAYBAN
2	CHANEL,RAYBAN,DUNHILL,CARTIER
3	DIOR,GUCCI,GUESS,LEVI'S,FERRARI
4	LEVI'S,DUNHILL,GUESS,RAYBAN
5	MERCEDES BENZ,LEVI'S,BVLGARI,BILLABONG,ADIDAS
6	EMPORIO ARMANI, GUCCI, MERCEDES BENZ, GUCCI, ADIDAS
7	MONTBLANC,NIKE,OAKLEY,GUCCI, RAYBAN, FERRARI
8	OAKLEY,LEVI'S,POLICE,PUMA,BVLGARI,D&G
9	CARRERA, CHANEL, FERRARI, GUCCI, LEVI'S, MERCEDES BENZ
10	MONTBLANC, NIKE, POLICE, RODENSTOCK, BVLGARI, CARRERA
11	DIOR, D&G, DUNHILL, NIKE, POLICE, PORSCHE DESIGN, CHANEL
12	RODENSTOCK, ADIDAS, D&G, GUCCI, MERCEDES BENZ, DUNHILL, DIOR

Source: National Optics 2016

    a. Itemset Formation

Based on the data in Table 3, the following is a case settlement. The process of producing C1, also known as 1 itemset, with a minimum of 30% support. The formula is as follows:

Support A = $\dfrac{\sum \text{Transactions Containing A}}{\sum \text{Transaction}} \times 100\%$

Table 4. Support Itemset

Itemset	Support
FERRARI	33%
RAYBAN	33%
CHANEL	30%
DUNHILL	33%
DIOR	30%
GUCCI	41%
LEVI'S	41%
MERCEDES BENZ	33%
BVLGARI	30%
ADIDAS	30%
NIKE	30%
POLICE	30%
D&G	30%

a.   **Combination of 2 Itemset**

The process of forming C2 or called 2 itemset with a minimum amount of support = 15% It can be done with the following formula :

Support (A,B) = P(A∩B)

Support A.B = $\frac{\sum \text{Transactions Containing A and B}}{\sum \text{Transaction}}$ x 100%

Table 5. Candidate 2 Itemset

Itemset	Amount	Support
FERRARI - BVLGARI	0	0%
FERRARI - ADIDAS	0	0%
FERRARI - NIKE	1	8%
FERRARI - POLICE	0	0%
FERRARI - D&G	0	0%
RAYBAN - CHANEL	1	8%
RAY BAN - DUNHILL	2	16%
RAYBAN - DIOR	0	0%
RAYBAN - GUCCI	1	8%
RAYBAN - LEVI'S	1	8%
RAYBAN - MERCEDES BENZ	0	0%
RAY BAN - BVLGARI	0	0%
RAY BAN - ADIDAS	0	0%

RAYBAN - NIKE	1	8%
RAYBAN - POLICE	0	0%
RAYBAN - D&G	0	0%
CHANEL - DUNHILL	2	16%
CHANEL - DIOR	1	8%
CHANEL - GUCCI	1	8%
CHANEL - LEVI'S	1	8%

Because the stated minimum support is 15%, any combination of two itemsets that fails to fulfil this requirement will be eliminated, as indicated in Table 6 below.:

Table 6. Minimum Support 2 Itemset 15%

Itemset	Jumlah	Support
FERRARI - RAYBAN	2	16%
FERRARI - GUCCI	3	25%
FERRARI - LEVI'S	2	16%
RAY BAN - DUNHILL	2	16%
CHANEL - DUNHILL	2	16%
DUNHIL - DIOR	2	16%
DUNHILL - D&G	2	16%

DIOR - GUCCI	2	16%
DIOR - D&G	2	16%
GUCCI - LEVI'S	2	16%
GUCCI - MERCEDES BENZ	3	25%
GUCCI - ADIDAS	2	16%
LEVI'S - BVLGARI	2	16%

b. Establishment of Association Rules

After all of the high-frequency patterns have been identified, the confidence of the associative rules A→B is calculated to find the association rules that fulfil the minimal confidence criteria. 70 percent confidence level is the minimum level of confidence. The following formula is used to get the confidence value from rule A→B :

Confidence = $\frac{\sum \text{Transactions Containing A and B}}{\sum \text{Transaction A}}$ x 100%

The magnitude of the support and confidence values of the candidate association rules may be observed in table 7 from the combination of two itemsets that were discovered.

Table 7. Association Rules

ATURAN	CONFIDENCE	
If Buy FERRARI then will buy RAYBAN	2/4	50%
If Buy FERRARI then will buy GUCCI	**3/4**	**75%**
If Buy FERRARI then will buy LEVI'S	2/4	50%
If Buy RAY BAN then will buy DUNHILL	2/4	50%

If Buy CHANEL then will buy DUNHILL	2/3	66%
If Buy DUNHIL then will buy DIOR	2/4	50%
If Buy DUNHIL then will buy D&G	2/4	50%
If Buy DIOR then will buy GUCCI	2/3	66%
If Buy DIOR then will buy D&G	2/3	66%
If Buy GUCCI then will buy LEVI'S	2/5	40%
If Buy GUCCI then will buy MERCEDES BENZ	3/5	60%
If Buy GUCCI then will buy ADIDAS	2/5	40%
If Buy LEVI'S then will buy BVLGARI	2/5	40%
If Buy MERCEDES BENZ then will buy ADIDAS	2/4	50%
If Buy BVLGARI then will buy POLICE	2/3	66%
If Buy NIKE then will buy POLICE	2/3	66%
If Buy POLICE then will buy D&G	2/3	66%

c. **Final Association Rules**

Table 8 shows how the final association rules are arranged depending on the minimal support and confidence that have been calculated.

Table 8. Final Association Rules

Aturan	Support	Confidence
If Buy FERRARI then will buy GUCCI	75%	75%

The most commonly sold eyewear brands at the National Optical firm may be observed in the graph below, which is based on the aforementioned association rules:
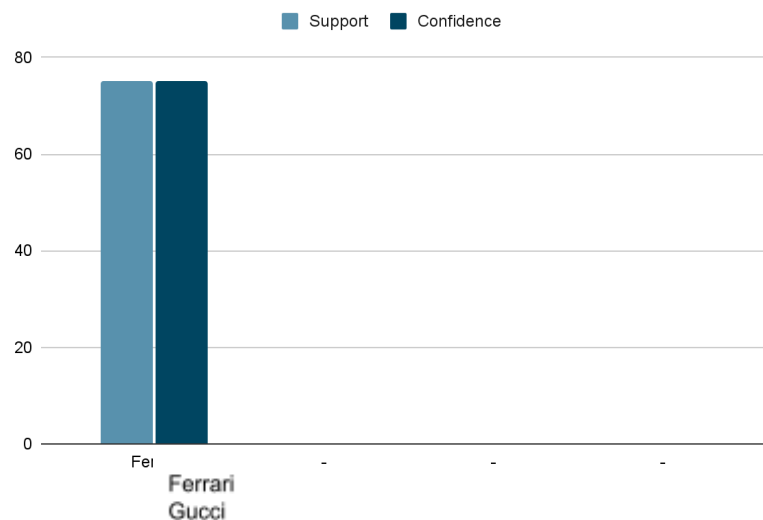

Figure 2. Most Final Sales Association Rules Formation Result Chart

4. Conclusion

The most commonly sold Glasses brands are Ferrari and Gucci, with the most widely sold items being known, according to an a priori algorithm based on the graph above. Companies may utilise the Apriori algorithm to build marketing plans to promote items with other brands by studying what the product's biggest benefits are. Some suggestions for future research include: the object of research should not only cover one year, but should cover up to five years in order for the data to be more accurate; and in future research, it is better to develop a system that is built on an application that is simple to use for data processing.

References

[1] Haryanto D, Oslan Y, Dwiyana D. "**Implementation of Shopping Cart Analysis With Association Rules Using Apriori Algorithm in Motorcycle Parts Sales**". 2014;

[2] Susanto S, Suryadi D. "**Introduction to data mining: digging knowledge from chunks of data**". Andi Publisher; 2010.

[3] Setiawati DD. "**THE USE OF APRIORI METHOD FOR MARKET CART ANALYSIS ON MINIMARKET SALES TRANSACTION DATA**". Inform Tech Study Program Thesis.

[4] DK pane." **Implementation of Data Mining on Sales of Electronic Products Using Apriori Algorithm (Case Study: Kreditplus)**". Pelita Inform Budi Darma. 2013;4(3):25–9.

[5] Anggraeni HD, Saputra R, Noranita B. "**Data Mining Application Analysis of Drug Sales Transaction Data Using Apriori Algorithm (Case Study at Pharmacy Setya Sehat Semarang)**". J Informatics Technol. 2014;2(2):24–30.