http://infor.seaninstitute.org/index.php/infokum/index
JURNAL INFOKUM, Volume 7, No.1, Desember 2018 ISSN 2302-9706

THE DATA MINING OF CELL PHONE MOST INTERESTED USING APRIORIAL ALGORITHM

[1]Penda Sudarto Hasugian, [2]Suprianto Panjaitan
[1,2]Program Studi Teknik Informatika
STMIK Pelita Nusantara Jl Iskandar Muda No 01 Medan, Sumatera Utara, Indonesia, 20154
sudarto89@gmail.com

Abstract

Data mining is a term used to describe knowledge discovery in a database or often called Knowledge Discovery in Database (KDD). With the development of information at this time, the need for accurate information is needed in daily life, so that information will become an important element in the development of society today and the future. However, high information needs are sometimes not balanced by the presentation of adequate information, often the information must still be extracted from very large amounts of data. The ability of information technology to collect and store various types of data far leaves the ability to analyze, summarize and extract knowledge from data. Decision-makers try to utilize data warehouse that has been owned to dig up information that is useful to help make decisions, this encourages the emergence of new branches of science to overcome the problem of extracting information or patterns that are important or interesting from large amounts of data, which is called data mining. The use of data mining techniques is expected to provide knowledge previously hidden in the data warehouse so that it becomes valuable and useful information.

Keywords: data mining, Sales, a-priori

1. Introduction

Data mining is a term used to describe knowledge discovery in a database or often called Knowledge Discovery in Database (KDD). With the development of information at this time, the need for accurate information is needed in daily life, so that information will become an important element in the development of society today and the future. However, high information needs are sometimes not balanced by the presentation of adequate information, often the information must still be extracted from very large amounts of data. The ability of information technology to collect and store various types of data far leaves the ability to analyze, summarize and extract knowledge from data. The traditional method for analyzing existing data, cannot handle large amounts of data. Utilizing existing data in information systems to support decision-making activities, it is not enough to rely solely on operational data, it is necessary to analyze data to explore the potential of existing information. Decision-makers try to utilize data warehouse that has been owned to dig up information that is useful to help make decisions, this encourages the emergence of new branches of science to overcome the problem of extracting information or patterns that are important or interesting from large amounts of data, called data mining. The use of data mining techniques is expected to provide knowledge previously hidden in the data warehouse so that it becomes valuable and useful information.

Generally, companies are currently required to have a competitive advantage by utilizing all available resources. In addition to facility, infrastructure, and human resources, information systems are one of the resources that can be used to increase competitive advantage. Information systems can be used to obtain, process, and disseminate information to support daily operational activities while supporting strategic decision-making activities.

2. Literature Review

2.1 Data Mining

Data Mining is a process that employs one or more computer learning techniques to analyze and extract knowledge automatically or a series of processes to explore the added value of a data set in the form of knowledge that has not been known manually. It is worth remembering that the word mining itself means an attempt to get a few valuable items from a large amount of basic material. Therefore Data Mining has long roots in the fields of science such as artificial intelligence, machine learning, statistics, and databases. Data mining is the process of applying this method to data to uncover hidden patterns. In other words Data mining is a process for extracting patterns from data. Data mining is becoming an increasingly important tool for turning that data into information. This is often used in various profile practices, such as marketing, surveillance, fraud detection, and scientific discovery. It has been used for years by businesses, scientists, and governments to filter out volumes of data such as flight passenger travel records, census data, and supermarket scanner data to produce a market research report.[1]–[4]

The main reason for using data mining is to assist in the analysis of a collection of observational behaviors. The data is susceptible to collinearity because a link is known. The inevitable fact of data mining is that the subsets of data sets analyzed may not represent the entire domain, and therefore may not contain examples of certain critical relationships and behaviors that exist in other parts of the domain. To overcome this kind of problem, analysis can be augmented using trial-based and other approaches, such as Choice Modeling for human-generated data. In this situation, inherent can be controlled correlation to, or removed altogether, during design construction.

2.2 Apriori Algorithm

A priori algorithm is a basic algorithm proposed by Agrawal & Srikant in 1994 to determine Frequent item sets for Boolean association rules. A priori algorithm including the type of Association Rules in data mining. Rules that state the association between several attributes are often called affinity analysis or market basket analysis. One of the stages of association analysis that attracts many researchers to produce efficient algorithms is the analysis of high-frequency patterns (frequent pattern mining). The importance of an association can be known by two benchmarks, namely: support and confidence. Support (supporting value) is the percentage of combinations of items in the database, while confidence (certainty value) is the strength of the relationship between items in association rules. [5]–[8]

1. Formation of item set candidates.
   The k-item set candidate is formed from a combination (k-1) –item set obtained from the previous iteration. One way of a priori algorithm is trimming k-item set candidates whose subset contains k-1 items not included in a high-frequency pattern with a length of k-1.
1. Calculation of support for each k-item set candidate.
   Support from each k-item set candidate is obtained by scanning the database to count the number of transactions containing all items in the k-item Set candidate. This is 2 .Calculation of support for each k-item set candidate. Support from each k-item set candidate is obtained by scanning the database to count the number of transactions containing all items in the k-item set candidate. analysis of high-frequency patterns (frequent pattern mining).

2.3 Steps of Rule Association Rules Process

Association analysis is also known as one of the data mining techniques that is the basis of various other data mining techniques. Especially one of the stages of association analysis called frequent pattern mining (analysis of frequent pattern mining) attracts the attention of many researchers to produce efficient algorithms. [9]–[12]

The basic methodology of association analysis is divided into two stages:

1. Analysis of high-frequency patterns This stage looks for combinations of items that meet the minimum requirements of the support value in the database. An item's support value is obtained using the following formula:

Support (A) = Transactions Containing A and B …… .. (1) Total Transactions Containing A

While the value of the two-item support is obtained from the following formula:

Support (A, B) = P (A B) = Number of Transactions Containing ADB ……. (2) Total Transactions n

2. Formation of Association Rules

After all high-frequency patterns have been found, then the associative rules are sought that meet the minimum requirements for confidence by calculating the associative rule confidence "if A then B". The confidence value of the "if A then B" rule is obtained from the following formula:

Confidence = P (B | A) = Number of Transactions Containing AdBB …… (3) Number of Transactions Containing A

3.Results and Discussion

The process of forming a combination of items and creating rules starts from data analysis. The data used is the previous Mobile sales transaction data, then continued with the formation of a combination of items and from the pattern of interesting combination items formed association rules

The following are sales data that will be sampled for analysis and for testing

Table 1 List of Mobile Stage I sales

No	Item	Total
1	Samsung	5
2	Sony	10
3	Panasonic	13
4	LG	8
5	Oppo	17
6	Vivo	11
7	Evercross	6
8	Advan	2
9	Polytron	12
10	HTC	5
11	Digitec	4
12	Huawe	4
13	Nokia	5
14	Matrix	0

Table 2 List of Mobile Stage 2 sales

No	Item	Total
1	Samsung	17
2	Sony	5
3	Panasonic	8
4	LG	6
5	Oppo	9
6	Vivo	0
7	Evercross	11
8	Advan	0
9	Polytron	7
10	HTC	7
11	Digitec	5
12	Huawe	2
13	Nokia	13
14	Matrix	0

Table 3 List of Mobile Stage 3 sales

No	Item	TOTAL
1	Samsung	11
2	Sony	6
3	Panasonic	17
4	LG	7
5	Oppo	22
6	Vivo	4
7	Evercross	0
8	Advan	14
9	Polytron	9
10	HTC	6
11	Digitec	1
12	Huawe	14
13	Nokia	1
14	Matrix	2

Table 4 List of Mobile Stage 4 sales

No	Item	code
1	Samsung	A
2	Sony	B
3	Panasonic	C
4	LG	D
5	Oppo	E
6	Vivo	F
7	Evercross	G
8	Advan	H

9	Polytron	I
10	HTC	J
11	Digitec	K
12	Huawe	L
13	Nokia	M
14	Matrix	N

3.1 Discussion

The process of forming a combination of items and creating rules starts from data analysis. The data used is the previous Mobile sales transaction data, then continued with the formation of a combination of items and from the pattern of interesting combination items formed association rules.

1. Transaction Data

Transaction data is a data representation obtained from daily sales. The following data is the Mobile sales transaction

Table 5 types of Mobile types

No	Type of Brand Mobile (items) Sales Transaction
1	Samsung, LG, Evercross, Polytron, Puji Elektrik, Nokia
2	Samsung, Panasonic, Oppo, Evercross, HTC, Nokia
3	Panasonic, Oppo, Advan, HTC,Nokia, Matrix
4	Sony, LG, Vivo, Advan, Digitex, Huawe
5	Samsung, Sony, LG, Oppo, Vivo, Polytron
6	Samsung,Oppo, Polytron, HTC, Huawe, Matrix

Table 6 Format of Tabular Data Transaction Data items

No	Items	Transaction					
		1	2	3	4	5	6
1	Samsung	1	1	0	0	1	1
2	Sony	0	0	0	1	1	0
3	Panasonic	0	1	1	0	0	0
4	LG	1	0	0	1	1	0
5	Oppo	0	1	1	0	1	1
6	Vivo	0	0	0	1	1	0
7	Evercross	1	1	0	0	0	0
8	Advan	0	0	1	1	0	0
9	Polytron	1	0	0	0	1	1
10	HTC	0	1	1	0	0	1
11	Digitec	0	0	0	1	0	0
12	Huawe	1	0	0	1	0	1
13	Nokia	1	1	1	0	0	0
14	Matrix	0	0	1	0	0	1
	Jumlah	6	6	6	6	6	6

Table 7 Candidate Table First Item

No	Item
1	Samsung
2	Sony
3	Panasonic
4	LG
5	Oppo
6	Vivo
7	Evercross
8	Advan
9	Polytron
10	HTC
11	Digitec
12	Huawe
13	Nokia
14	Matrix

support from each k-itemset candidate is obtained by scanning a database to count the number of transactions containing all items in the k-itemset candidate. This is also a feature of the a priori algorithm, which is the calculation required by the entire database of the longest k-itemset. To determine the value of support can be known by using the following formula.

SupportA = ∑ Transactions containing A * 100
∑ Transactions = ∑ Transactions containing Samsung * 100%
∑ Transactions = 4/6*100% =66,66%

Table 8 List of candidate items for initial candidates

No	Items	Support
1	Samsung	66,66%
2	Sony	33,33%
3	Panasonic	33,33%
4	LG	50,00%
5	Oppo	66,66%
6	Vivo	33,33%
7	Evercross	33,33%
8	Advan	33,33%
9	Polytron	50,00%
10	HTC	50,00%
11	Digitec	16,66%
12	Huawe	50,00%
13	Nokia	50,00%
14	Matrix	33,33%

Support = 35%
And then compare with minimal support.

Table 9 List that meets the minimum support

No	Items/kode	Support	Support(%)
1	A	3	66,66%
2	D	4	50,00%
3	E	3	66,66%
4	I	4	50,00%
5	J	4	50,00%
6	M	3	50,00%

Table 10 List of Candidate items two combinations of 2 itemsset

No	Items
1	AD
2	AE
3	AI
4	AJ
5	AM
6	DE
7	DI
8	DJ
9	DM
10	EI
11	EJ
12	EM
13	IJ
14	IM
15	JM

Support from each k-itemset candidate is obtained by using a database scan to count the number of transactions containing all items in the k-itemset candidate. This is a typical feature of the a priori algorithm, which is the calculation required by scanning the entire database of the longest k-itemset. To determine the value of support can be known by using the following formula:

Table 11 List of candidate items support items 3 itemsset

Items	Support
EJI	33,33

Support (E, I, J) = ∑ Transactions containing E, I and J * 100%
∑ Transactions Containing A

= 2/6 * 100% = 33.33%
The process of calculating 3 itemset combination support, then no 35% minimum support is found, then the item calculation process stops.

2. Formation of Association Rules

After a high-frequency pattern is found. Then the confidence is calculated for each combination of items. Iteration stops when all items have been counted until there are no more item combinations. To find confidence from the A-B rules, the following formula is obtained:

Confidence P (a / b) = ∑ number of transactions containing A and B∑ transactions containing A
Confidence = 80%

Table 11 List of Association Rules that Meet Minimum Confidence

Rules	Support	Confidence(%)
If sold l (E), Then it will be sold (J)	3/3	100%
If sold (J), Then it will be sold (E)	¾	75%
If sold (I), Then it will be sold (J)	¾	75%
If sold (J), Then it will be sold (I)	¾	75%

Association rules from the table above are rules that are formed from a combination of three-item patterns, the table above is divided into several parts. Rules are rules that result from a combination of two itemset patterns. Support is the value found from the support of two items set divided by the value of the antecedent support multiplied by one hundred percent. The formation of the next association rule is formed from the combination of the two itemset patterns shown in the following table:

Table 12 List of 2 interesting itemset rules

Rules	Support	Confidence(%)
If sold (E), Then it will be sold (J)	40%	100%
If sold (J), then it will be sold (E)	40%	75%
If sold (I), Then it will be sold (J)	40%	75%
If sold (J), Then it will be sold (I)	40%	75%

Based on the discussion above, we can know that the most interesting types of Mobile items most often purchased by customers in transaction 1 are {E, J, J, E, I, J, J, I} from the description the company can find out a good analysis process for the future in the implementation of the Mobile stock process based on the interests of consumers. That the most frequently purchased item {E, J, I} is OPPO.

4. Conclusion

In the analysis of a number of data, it was found that the smaller the minimum support and confidence determined, the more rules can be generated, with the consequence that the processing time will be longer than the greater minimum support.

Reference

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.
[2] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, 2014.
[3] S. Agarwal, "Data mining: Data mining concepts and techniques," in *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*, 2014.
[4] R. Sowmya and K. R. Suneetha, "Data Mining with Big Data," in *Proceedings of 2017 11th International Conference on Intelligent Systems and Control, ISCO 2017*, 2017.

[5] H. Toivonen, "Apriori Algorithm," in *Encyclopedia of Machine Learning and Data Mining*, 2017.
[6] D. K. Pane, "Implementasi Data Mining Pada Penjualan Produk Elektronik Dengan Algoritma Apriori (Studi Kasus : Kreditplus)," *Pelita Inform. Budi Darma*, 2013.
[7] D. S. Kusumo, M. A. Bijaksana, and D. Darmantoro, "DATA MINING DENGAN ALGORITMA APRIORI PADA RDBMS ORACLE," *TEKTRIKA - J. Penelit. dan Pengemb. Telekomun. Kendali, Komputer, Elektr. dan Elektron.*, 2016.
[8] F. Rahmawati and N. Merlina, "Metode Data Mining Terhadap Data Penjualan Sparepart Mesin Fotocopy Menggunakan Algoritma Apriori," *PIKSEL Penelit. Ilmu Komput. Sist. Embed. Log.*, 2018.
[9] F. Rumaisa, "Penentuan Association Rule Pada Pemilihan Program Studi Calon Mahasiswa Baru Menggunakan Algoritma Apriori Studi Kasus Pada Universitas Widyatama Bandung," *Semin. Nas. Apl. Teknol. Inf.*, 2012.
[10] P. M. Hasugian, "PENGUJIAN ALGORITMA APRIORI DENGAN APLIKASI WEKA DALAM PEMBENTUKAN ASOSIATION RULE," *J. Mantik Penusa*, 2017.
[11] A. P. Utomo and I. I. Sungkar, "Analisis dan Perancangan Dashboard untuk Monitoring dan Evaluasi Pasien Rawat Inap," in *Prosiding Matheematics and Sciences Forum 2014*, 2014.
[12] F. Apriana, F. Jansen, and E. M. Lintong, "Perencanaan Pengembangan Sisi Udara Bandar Udara Mutiara Sis Al-Jufri Di Kota Palu Provinsi Sulawesi Tengah," *J. Sipil Statik*, 2017.