

# Applying Improved Apriori Algorithm in Figuring out the Relation between Weather Factors and Rainfall

Siti Zulaikha<sup>1,\*</sup>, Nola Ritha<sup>1</sup> and Martaleli Bettiza<sup>1</sup>

<sup>1</sup> Department of Informatics, Faculty of Engineering, Universitas Maritim Raja Ali Haji  
Jl. Politeknik Senggarang, Tanjungpinang

\*Corresponding Author: sitizulaikha51@gmail.com

## Article history

Received: 7-02-2020

Revised: 25-03-2020

Accepted: 08-04-2020

DOI: 10.31629/jit.v1i1.2133

## Abstract

Data on the rainfall is compelling to study as it becomes one of the major factors affecting the weather in a certain region and various aspects of life as well. Generally, predicting rainfall is performed by analyzing data in the past in certain methods. Rainfall is prone to follow repeated pattern in sequence of time. The utilization of big data mining is expected to result in any valuable information that used to be unrevealed in the big data store. Some methods used in data mining are Apriori Algorithm and Improved Apriori Algorithm. Improved Apriori itself is to represent the database in the form of matrix to describe its relation in the database. Data used in this research is the rainfall factor in 2016 in Tanjungpinang city. Based on the test of Improved Apriori Algorithm, it was found out that the relation of the rainfall and weather factors utilizing 2 item sets, that is, if the temperature is low (24,0 - 26,0), the humidity is high (85 - 100), then the rainfall is mild. If the temperature is low (24,0 - 26,0), the light intensity is low (0 - 3), then the rainfall is heavy, and 3 item sets if the temperature is low (24,0 - 26,0), the humidity is high (85 - 100), the sun light intensity is low (0-3), then the rainfall is medium.

**Keywords:** improved Apriori Algorithm, weather factors, rainfall

## 1. Introduction

Meteorological, Climatological, and Geophysical Agency (BMKG) is an Indonesian non-department government agency for meteorology, climatology, and geophysics. Some of its duties are observing, collecting and spreading the data, data processing, analysis and forecasting, meteorological data service, in each respective region.

The discipline of meteorology dealing with weather forecast and rainfall prediction is in fact related more to the concepts, principles, and methods in handling the reasoning model of

forecasting than what it actually is [1]. Data on the rainfall is compelling to study as it becomes one of the major factors affecting the weather in a certain region and various aspects of life as well. Generally, predicting rainfall is performed by analyzing the data in the past in certain methods. The utilization of existing system data in assisting the decision making does not suffice to solely rely on operational data. It is necessary that we perform data analysis so as to result in more potential from the existing information [2].

Any decision-makers are now striving to take advantages of existing data store to result in more useful information in making decision. This

becomes impetus to the surge of new discipline of computer science in handling the problems of informatics or mining important and compelling patterns of data in an abundant number, which is called data mining. The utilization of large data mining is expected to result in new useful and valuable information which used to be unrevealed in the large data store. In recent computer and data on development science, there have been many algorithm mining of related rules. One of the common algorithms used is Apriori Algorithm to extract frequent item set from the large database [3].

Some methods used in large data mining are Apriori Algorithm and Improved Apriori Algorithm. Apriori Algorithm is the classical one frequently used. While, Improved Apriori Algorithm is the representation of database in the form of matrix to describe the relation in the database [4]. This research is Applying Improved Apriori Algorithm in Figuring out the Relation between Weather Factors and Rainfall, a case study in Tanjungpinang city.

## 2. Methodology

### 2.1 Data Mining

Data mining is the process of extracting the pattern from the data (an important tool in transforming data into the form of information). Data mining commonly uses large data such as marketing data, surveillance, fraud detection and scientific enquiry, or any smart and automatic method that is capable of transforming data (geographical data, in time and space dimension, and etc.) in the form of information, and synthesizing knowledge [4].

### 2.2 Association Rules

Association Rules is one of the techniques in applying data mining. It is to find the pattern in large-scale database. It is commonly used in market transaction and analysis, graph mining, as application to determine chemical compounds, text analysis, as pattern for web browsing, and many more [4]. The basic methodology of association analysis is divided into two stages [5].

#### 2.2.1 Analysis of high frequency pattern

This phase is to find out the combination of item that meets the minimum requirement of the

supporting value in the database. The supporting value of an item is obtained from the following formula:

$$Support(A) = \frac{\text{Number of transactions with } A}{\text{Total number of transactions}} \quad (1)$$

Where:

- Support A is the supporting value percentage of combination of item A in the database.
- The number of transactions containing in A is the occurrence of item A in the whole transaction.
- The total transaction is the number of transactions in the database

While, the supporting value from 2 items is obtained from the following formula:

$$Support(A,B) = P(A \cap B) \quad (2)$$

$$Support(A,B) = \frac{\text{Number of transactions with } A \text{ and } B}{\text{Total number of transactions}} \quad (3)$$

Where:

A is the supporting value percentage of combination of item A in the database. The number of transaction A is the occurrence of item A in the whole transaction. The total transaction is the number of transactions in the database

#### 2.2.2 Forming Association Rules

After all high frequent patterns are set, Association Rules is determined as it has quite strong dependence on each item in antecedent and consequent and that has met minimum requirement for confidence by computing its Association Rules Confidence  $A \rightarrow B$ . The formula to determine the value of confidence from the association rules of  $A \rightarrow B$  is as follows:

$$Confidence = P(B|A) = \frac{\text{Number of transactions with both } A \text{ and } B}{\text{Total number of transactions with } A} \quad (4)$$

Where:

Confidence  $P(A|B)$  is the value of certainty factor between item A and item B in the Association Rules or some frequent simultaneous occurrences of item A and item B

### 2.3 Improved Apriori Algorithm.

Improved Apriori Algorithm based on Matrix is proposed by some researchers though in different techniques in determining its frequent itemset [5-6]. Improved Apriori Algorithm represents the database in the form of matrix to describe its relation in database. Then, matrix is computed to determine the support value of the frequent itemset candidate that meets the criteria to result in frequent itemset without re-scanning of the database using “AND” operation on the matrix lines matching with the item in the frequent itemset candidate and in adding the result of AND operation with the result is support [4,7].

The Steps of this algorithm such as [7]:

1. Converting database into matrix.
  - a. Conversion of database containing  $I_n$  item and transaction  $T_m$  into matrix form. Line from the matrix representing transactions and columns representing the item. If on transactions there are items, the value is 1 and 0 if otherwise.
  - b. Sum of the values from the column is the support count value and sum of values from rows is number of items in a transaction
2. Check the number of columns and the number of rows.
  - a. Delete columns for the number of columns is less than the minimum support value
  - b. Delete rows where there are less rows of equal to the value of k (k-frequent itemset).
3. Combine each column using a cross products to find combinations often 2-itemet and use the AND Operation for get its value.
4. Check the number of columns and the number of rows
  - a. Delete columns for the number of columns is less than minimum support value.
  - b. Delete rows where there are less rows of equal to the value of k (k-frequent itemset).
5. And to search for  $K_{th}$  -frequent itemset. Combine each column and delete columns that are less than the minimum support and delete rows where the number of rows is less than the same as k

### 3. Result and Discussion

The following is 10 sample data to process applying Improved Apriori Algorithm as it

complies with the attributes that have been determined in figuring out the relation between the data of weather factors and rainfall.

**Table 1.** Sample Data

NO	RR (Rainfall)	T (Temperature)	H (Humidity)	M (Intensity)
1	46,2	25,9	93	0,4
2	21,8	25,6	93	0,0
3	5,3	26,8	89	1,0
4	53,4	26,3	93	2,3
5	52,8	28,1	86	2,1
6	50,1	26,9	89	5,5
7	85,5	26,1	91	1,2
8	148,2	24,8	91	1,5
9	19,6	28,3	79	8,9
10	0	27,1	80	7,9

Next, these sample data are converted into the selected class. RR represents the Rainfall in millimeter (mm), T represents Temperature in Celsius (0C), H represents Humidity in Relative Humidity (RH) and M represents Light Intensity in Candela (cd).

**Table 2.** Data Conversion in Class

No	RR (mm)	T (°C)	K (RH)	M (cd)
1	RR1	T1	K3	M1
2	RR1	T1	K3	M1
3	RR1	T2	K3	M1
4	RR2	T2	K3	M1
5	RR2	T2	K3	M1
6	RR2	T2	K3	M2
7	RR2	T2	K3	M1
8	RR3	T1	K3	M1
9	RR1	T2	K2	M3
10	RR1	T2	K2	M3

In computing Improved Apriori applying Association Rules method, some steps to do are:

- a. Assuming the value of minimum support to use is 2
- b. Assuming the value of minimum confidence to use is 60%
- c. Performing transformation of data into matrix,

in which the attributes of data on rainfall is represented by line and data on weather factors are represented by column

**Table 3.** Transformation of Data into Matrix

	T1	T2	T3	K1	K2	K3	M1	M2	M3
RR1	1	1	0	0	1	1	1	0	1
RR2	0	1	0	0	0	1	1	1	0
RR3	1	0	0	0	0	0	1	0	0

Then, to determine frequent 1-itemset, eliminate the number of columns that have value less than minimum support and eliminate the number lines that have value less than or the same as minimum support.

**Table 4.** Frequent 1-itemset

	T1	T2	K3	M1
RR1	1	1	1	1
RR2	0	1	1	1
RR3	1	0	0	1

Next, perform AND operation to determine 2-itemset candidate.

**Table 5.** Frequent 2-itemset candidate

	T1,K3	T1,M1	T2,K3	T2,M1	K3,M1
RR1	1	1	1	1	1
RR2	0	0	1	1	1
RR3	0	1	0	0	0

Data whose columns less than the value of minimum support will be eliminated and the the number of lines whose value less than or the same as 2 will also be eliminated

**Table 6.** Frequent 2-itemset

	T1,M1	T2,K3	T2,M1	K3,M1
RR1	1	1	1	1
RR2	0	1	1	1

Next, eliminate the data whose columns less

than the value of minimum support and the number of lines whose value less than or the same as 2. Then, perform AND operation to determine 3-itemset candidate

**Table 7.** Frequent 3-itemset candidate

	T1,K3,M1	T2,K3,M1
RR1	1	1
RR2	0	1

As the matrix has reached null point, the process of searching frequent itemset stops. Then, that is the result of computation that applies Association Rules

**Table 8.** The Result of Association Rules

Item set	Combination
RR1	$T1 \cap K3 \cap M1 \rightarrow RR1$
	$T2 \cap K3 \cap M1 \rightarrow RR2$
RR2	$T2 \cap K3 \cap M1 \rightarrow RR2$
RR3	$T1 \cap M1 \rightarrow RR3$

#### 4. Conclusion

The conclusion is drawn from this research that the applying Improved Apriori Algorithm in determining the relation between the rainfall and weather factors, which, among other things are the pattern in Improved Apriori with 2 item sets and 3 itemset. 2 item sets show that if the temperature is low (24,0 - 26,0), the humidity is high (85 - 100), then the rainfall will be light. If the temperature is low (24,0 - 26,0), the intensity of the sun light is low/dark (0 - 3), then the rainfall will be heavy. By utilizing 3 item sets, if the temperature is low (24,0 - 26,0), the humidity is high (85 - 100), the intensity of the sun light is low/dark (0-3), then the rainfall will be medium. In short, by applying Improved Apriori Algorithm, it shows that the computation resulted from the searching process of each pattern has relation with the data of the rainfall and weather factors.

#### Acknowledgement

The writer would like to thank to the staffs of Meteorological, Climatological, and Geophysical Agency (BMKG) of Tanjungpinang city.

## References

- [1] N. Ritha and R. Wardoyo. 2016. Implementasi Neural Fuzzy Inference System dan Algoritma Pelatihan Levenberg Marquardt untuk Prediksi Curah Hujan, *Indonesian Journal of Computing and Cybernetics Systems*, Vo. 10, No. 2
- [2] M. I. Azhar and W. F. Mahmudy. 2018. Prediksi Curah Hujan Menggunakan Metode Adaptive Neuro Fuzzy Inference System (ANFIS), *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, Vol. 2, No. 11.
- [3] K. B. Agyapong, and J. B. Hayfron-Acquah. 2015. An Improved Apriori Algorithm Established On Probability Matrix, *International Journal Of Scientific & Technology, Reserch* Volume 4, Issue 11, November 2015.
- [4] N. Krisdianto, and A. M. Arymurthy. 2012. Improve Apriori Berbasis Matrix Dengan Incremental database untuk Market Basket Analysis, *Skripsi, Univ. Sultan Ageng Tirtayasa, Univ. Indonesia*.
- [5] Han and Kamber. 2006. *Data Mining Concept and Technique*, Second Edition, The Morgan Kaufmann, San Fransisco.
- [6] Y. Septiana and D. Dharmayanti. 2016. Penerapan Improve Apriori Pada Aplikasi Data Mining Di Perusahaan Calvin Socks Production, *Jurnal Ilmiah Komputer Dan Informatika (Komputa)*, Volume 5, 01, ISSN : 2089-9033.
- [7] V. Mangla, C. Sarda, and S. Marda. 2013. Improving The Efficiency of Apriori Algorithm in Data Mining, *International Journal of Engineering and Innovative Technology (IJEIT)*, Vol. 3