Analysis of Frequent Itemsets Mining Algorithm Againts Models of Different Datasets

R Muliono¹, Muhathir², N Khairina³, M K Harahap⁴

¹,²,³Department of Informatics Engineering, Universitas Medan Area, Medan, Indonesia
⁴Department of Information Technology, Politeknik Ganesha, Medan, Indonesia

*Email: ¹rizkimuliono@gmail.com, ²muhathirbangdes@gmail.com, ³nurulkhairina27@gmail.com, ⁴choir.harahap@yahoo.com

Abstract. Data mining is a study that uses statistical knowledge, mathematical calculations, artificial intelligence methods, machine learning by extracting and identifying useful information and related knowledge from various large databases. One of them is looking for itemsets combination from the data stack, the search process can be done using the Apriori Association Rules algorithm, the FP-Growth Association Rule and Closed Association rule. The three algorithms are some of the implementations of frequent itemsets search methods. Two datasets will be tested, namely retail datasets and accidents datasets derived from fimi datasets, with the aim of knowing the behavior of the three algorithms against the dataset model we tested. Each algorithm will test both datasets, with minimum support values for retail datasets ranging from 0.01-0.05 and min-conf 0.01-0.05. Likewise on the accident min-sup and min-conf 0.6-1 datasets. For retail algorithm data with the fastest processing time, the FP-Growth AR algorithm is more than 85% compared to Apriori-AR, followed by AR-Apriori and Apriori-Close-AR. For detailed memory usage results the Apriori-AR algorithm is lighter and outperforms. Accident data the experimental results show that FP-Growth-AR is a little more memory efficient and the fastest process is Apriori-Close, while the Apriori-AR with the longest processing time.

1. Introduction
Frequent itemsets one of the results that will be searched in the data mining process against large data stacks, are often used to find information on association rules and were first introduced for market basketball analysis [7] [8]. Frequent itemset in the form of a combination of a combination of all items in the transaction data that are mutually interconnected and the combination can be found using several data mining algorithm techniques, one of which is a priori algorithm [1].

2. Related Research
A priori algorithm has become a well-known method in data mining techniques, especially in the search for frequent items and association rules. In previous studies [2] an analysis of this a priori algorithm has been attempted, namely analysis of a priori-close improvement with FP-Growth where apriori-close is faster 30% for the processing time of FP-Growth for frequent itemsets search process, but almost 60 % FP-Growth saves more memory in minutes 0.10 - 0.20 and then changes in process
behavior, which is almost parallel to Apriori-close until minutes 0.50 and the data used is kosarak datasets and mushrooms.

In this study will analyze the process behavior of apriori algorithms and its improvement especially the Association Rules for different dataset models and patterns, namely retail and accents taken from FIMI Repository Dataset published by IBM Almaden Quest Research Group sourced from http://fimi.ua.ac.be/data/\[19\] including the Apriori Association Rules algorithm, the FPGrowth Association Rule and Closed Association rule.

3. Data Mining
Data mining is a complex process that uses statistical techniques, mathematical calculations, artificial intelligence techniques, pattern recognition models, algorithms, system databases, machine learning to extract data to identify useful information and related knowledge from various large databases \[14\]. And data mining itself is defined as a data pattern with the potential to make non-trivial projects about unknown patterns in the dataset. Development of data mining disciplines \[15\].

(see figure 1)

![Figure 1. The development of multi-disciplines in the field of datamining today.](image)

A frequent itemset is one of the important information about the process results from the data mining process \[7\] and one of the important elements in KDD (Knowledge Database Discovery) \[8\]. In this industry era 4.0, the concept and implementation of data mining has become an inseparable need to support the completeness of the process components from large databases to be extracted in stages according to the models and needs that are thoroughly integrated into all network applications or systems such as manufacturing with IoT and Solid and complex information transfer needs.

The term data mining has the essence as a scientific discipline whose main purpose is to find, explore, or mine knowledge from the data or information we have. Data mining, currently more commonly known as Knowledge Discovery in Database (KDD). KDD is a series of activities which include the stages of collecting, using data and historically to find forms of order, patterns or relationships in large datasets \[1\] \[4\] \[6\] \[12\].

3.1. Training Method
Training methods used in data mining techniques are divided into two approaches \[12\].
1. Unsupervised learning, this method is applied without training (training) and without a teacher (teacher). The teacher here is a label of data.
2. Supervised learning is a method of learning with training and training. In this approach, to find decision functions, separator functions or regression functions, several examples of data that have output or labels are used during the training process.

3.2. Datamining Grouping
There are several techniques that data mining has based on tasks that can be done, namely \[13\]:

---

**Figure 1.** The development of multi-disciplines in the field of datamining today.

---

A frequent itemset is one of the important information about the process results from the data mining process \[7\] and one of the important elements in KDD (Knowledge Database Discovery) \[8\]. In this industry era 4.0, the concept and implementation of data mining has become an inseparable need to support the completeness of the process components from large databases to be extracted in stages according to the models and needs that are thoroughly integrated into all network applications or systems such as manufacturing with IoT and Solid and complex information transfer needs.

The term data mining has the essence as a scientific discipline whose main purpose is to find, explore, or mine knowledge from the data or information we have. Data mining, currently more commonly known as Knowledge Discovery in Database (KDD). KDD is a series of activities which include the stages of collecting, using data and historically to find forms of order, patterns or relationships in large datasets \[1\] \[4\] \[6\] \[12\].

3.1. Training Method
Training methods used in data mining techniques are divided into two approaches \[12\].
1. Unsupervised learning, this method is applied without training (training) and without a teacher (teacher). The teacher here is a label of data.
2. Supervised learning is a method of learning with training and training. In this approach, to find decision functions, separator functions or regression functions, several examples of data that have output or labels are used during the training process.

3.2. Datamining Grouping
There are several techniques that data mining has based on tasks that can be done, namely \[13\]:

---
1. Description, Researchers usually try to find ways to describe patterns that are hidden in the data according to the growing trend of their time.

2. Estimation, Estimation is similar to classification, except that the destination variable is more numerical than the category.

3. Predictions, Predictions are similar to estimates and classifications. It’s just that the prediction of the results shows something that hasn’t happened (it might happen in the future).

4. Classification, in variable classification, objectives are categorical. For example, we will classify income in three classes, namely high income, moderate income, and low income.

5. Clustering, more towards grouping records, observations, or cases in a class that has similarities.

6. Association, identifies the relationship between various events that occur at one time.

3.3. Data Mining Phases
As a series of processes, data mining can have several process steps illustrated in Figure 2. These stages are interactive, the user is directly involved or through the mediation of a knowledge base [16] [17]. (see figure 2).

![Figure 2. Main phases of a data mining process.](image)

From the picture above can be explained that the process in data mining is:

1. Data cleaning (data cleaning), is the process of dissipating noise and inconsistent data or irrelevant data.

2. Data integration (data integration), is a combination of data from various databases into a new database.

3. Data selection (data selection), the data in the database is often not all used, therefore only the data that is suitable for analysis will be taken from the database.

4. Data transformation (data transformation), data is changed or merged into a format suitable for processing in data mining.

5. Mining process, is a main process when the method is applied to find valuable and hidden knowledge from the data. Some methods can be used based on data mining grouping.

6. Evaluation of patterns (pattern evaluation) To identify interesting patterns into the knowledge based.

7. Knowledge presentation (knowledge presentation) Is a visualization and presentation of knowledge about the method used to obtain knowledge gained by users.

4. Frequent Pattern Mining Algorithms
Apriori Algorithm is one of the data mining algorithms that is used to find frequent itemsets from a database or also known as Frequent pattern mining [8]. Itemsets search on a priori must have parameters: minimum support and minimum confidence [2] [3], in a priori process has a search pattern for frequent itemsets combinations in this a priori using the search method for the breadth-first-search algorithm [7] [11].

The purpose of a priori algorithm is to find all frequent itemset by repeating the above process by searching repeatedly for all the items in the dataset and ending when the itemsets candidate is generated Ck-1 [4]. The looping process that occurs in the a priori algorithm when searching and selecting combinations in the lattice tree is to do the trimming process to determine the appropriate
Combination between the one in the lattice tree traversal and the one in the dataset, this process is called the join and pruning lattice traversal itself is the development of capabilities from the selection of combinations of a priori algorithms which are also referred to as Eclat a priori [6].

This a priori algorithm has undergone many changes with a better level of optimization in terms of memory usage and processing time for larger and more complex transaction datasets. The most important thing first sought in frequent pattern mining is association rule mining [7] [10]. There are two stages in the association rule:

4.1. Frequent Itemset Generation
Frequent 1-itemset is found by reading the database to collect counts for each item, and collecting items that meet the minimum support. The resulting set is denoted by $L_1$. Next, $L_1$ is used to find $L_2$, frequent itemset 2-itemset, which is used to find $L_3$, and so on, until no more frequent items are found. Every $L_k$ search is carried out by re-reading the database until there is no more suitable combination found [1].

4.2. Rule Generation
There are two attributes of the Rule Mining Association (ARM), namely Support and Confidence. Frequent itemset results that are obtained using a priori algorithm, then is to get a rule that meets confidence. Because the resulting rule comes from frequent itemset, in other words, in calculating the rule using confidence, there is no need to calculate the support because all the resulting rules have fulfilled the minutes specified [18].

This calculation also does not need to repeat scanning on the database to calculate confidence, just by taking itemset from the support results. Support ($s$) is the visitor's value or the percentage of a combination of an item in the database and confidence ($c$) is defined as the proportion of the number of transactions including the anterior and after all records containing $D$ [6]. For example, ABCD and AB are frequent, so the ABCD rule is obtained if the ratio of support (ABCD) to support (AB) is at least equal to the minimum confidence. This rule has a minimum support because ABCD is frequent.

\[
Support (A) = \left( \frac{\text{The number of transactions containing } A}{\text{Total transactions}} \right) \times 100\% \quad (1)
\]

\[
Confidence P(B|A) = \frac{\text{The number of transactions containing } A \text{ and } B}{\text{Total transactions } A} \times 100\% \quad (2)
\]

4.3. Apriori Association Rule
Frequent itemset are searched and iteratively calculated, in ascending order according to the size of the data. The process requires as many as $k$-iterations, where $k$ is the size of the largest Itemset frequent for each iteration of $k$, the database is read once and the number of frequent items is available. The first iteration calculates the $L_1$ set from frequent 1-itemsets. Then the iteration consists of two phases. First, a set of $C_i$ from candidate $i$-itemset is made by the frequent merge ($i1$) - it seems in $L_1$. Found in the previous iteration. This phase is realized by the Apriori-Gen function that generates combinations. Furthermore, the database is scanned to determine candidate support in $C_i$ and frequent $i$-itemset is extracted from candidates. This process is repeated until no more candidates can be produced [2]. (see figure 3).
Figure 3. Pseudocode Apriori Algorithm.

```
L_1 := { large 1-itemsets };
k := 2;  // k represents the pass number
while (L_{k-1} ≠ ∅) do
begin
    C_k := New candidates of size k generated from L_{k-1}; (apriori-gen)
    forall transactions t ∈ D do
        Increment the count of all candidates in C_k that are contained in t;
    L_k := All candidates in C_k with minimum support;
    k := k + 1;
end
Answer := U_k L_k;
```

From the pseudocode figure above the a priori algorithm can be explained more clearly the procedure line.

1) Produces frequent itemset along k (initially k = 1).
2) Repeat the process until there are no frequently identified item sets from (3) to (6).
3) Generate length (k + 1) itemset candidate of length k frequent itemset (Generation and Trimming Candidate Items).
4) Length of prune length (k + 1) itemset candidates that contain subset of length k which are rare (Generation and Trimming Candidate Items).
5) Calculate support for each candidate (Counting Support).
6) Elimination of length (k + 1) candidates are rare.

4.4. FP-Growth Association Rule

FP-growth is one of the fastest algorithms for finding frequent item sets within datamining [2] [11] [12]. This algorithm looks for itemset without doing candidate generation itemset and using tree data structure or (FP-Tree). FP-tree is a compressed data storage structure. FP-tree is built by mapping each transaction data into each particular path in the FP-tree. Because in each transaction that is mapped, there may be transactions that have the same item, then the path allows to overwrite each other. The more transaction data that has the same item, the compression process with the FP-tree data structure is more effective.

The FP-Growth method can be divided into 3 main stages, namely as:
1. Stage of generating conditional pattern base,
2. Stage of generation of FP-Tree conditionals, and
3. Frequent itemset search phase.
4. These three stages are the steps that will be taken to get frequent itemset.

4.5. Closed Association Rule

Apriori-Close pruning candidates who often appear are carried out with one step, in which the subset of candidates who are frequently checked. By default all subsets are marked as having proximity, which is changed if the subset support value is the same as the candidate that was actually checked during comparison. As a result, in Apriori-Close all subsets of candidate frequent are generated, where many processes are in the tree scheme. The useless comparison process will be avoided if the closest itemset filtering is carried out in the phase when the candidate generation and pruning are carried out using the applied approach method. With this method, a subset has been determined, so checking the equality of support values does not require extra comparison [2].

The Apriori-Close algorithm is basically different from the existing algorithm, because it is based on Pruning in the closest itemset to find frequent itemset. Itemset close is a maximum set of items that often appear. For example, in database D, itemset {B, C, E} are the closest itemset because this is a series of the most common items known to objects {2,3,5}. {B, C, E} is called frequent closed itemset for minimum support = 2 as support ({B, C, E}) = || {2,3,5} || = 3 ≥ min-support. In the database, this means that 60% of customers (3 customers out of a total of
5) buy the most items B, C, E. Itemset {B, C} not the closest itemset, since not the maximum item group that is common for some objects: all customers buy goods B and C also buy item E.

Apriori-Close has a process:
1. The process is to find all the nearest itemset that often appear in D, which is the closest itemet and has a min_support value greater or equal to min_support.
2. Lower all frequent itemset from frequent itemsset which is often found in phase 1. This phase will generate all subset of the closest itemset that is maximal and decrease it based on the nearest itemset min_sup which is frequent.
3. Every frequent itemset found in phase 2, make all association rules that can be derived that have a confidence value greater than or equal to the confidence.

4.6. Dataset
The dataset that will be used in this research is retail.txt of 4.2MB and accidents.dat of 35MB taken from the FIM Dataset Repository published by IBM Almaden Quest Research Group sourced from http://fimi.ua.ac.be/data/ [19].

4.7. Tools
In this study, we use the SPMF application which is an open-source mining data library application that is built with the java programming language, which specifically searches for data patterns in the implementation of data mining algorithms [20]. And the computer device used is the 8GB macbook pro ram, Intel Core i7 Mid 2012.

4.8. Research framework
In this study, we intend to analyze the trend patterns of the Apriori Association Rule, FP-Growth Association Rule and Closed Association rule. Conducting frequent itemset search and value experiments found the association rules with minimum-support values on retail and accident datasets ranging from 0.01 - 0.05 and minimum-confidence 0.01 - 0.05. The experiment will document resource use (mb) and processing time (ms). From the description it can look like figure 5.

5. Discussion
This study conducted experiments on retail datasets for min-sup 0.01 - 0.05 and min-conf 0.01 - 0.05 with the results shown in table 1.
Table 1. Result of Retail.txt Ar (Association Rule), L (Frequent itemsets, mb (memory), ms (millisecons))

| Min-Supp | Min-Conf | Apriori AR | FP-Growth AR | Apriori-Close AR |
|----------|----------|------------|--------------|-----------------|
|          |          | AR (L) (ms) | L (mb) (ms) | AR (L) (ms) | L (mb) (ms) | AR (L) (ms) | L (mb) (ms) |
| 0.05     | 0.05     | 32 (16)    | 93 (204)     | 32 (16)    | 269 (215)  | 32 (16) | 739 (1788) |
| 0.04     | 0.04     | 34 (18)    | 194 (266)    | 34 (18)    | 266 (21011) | 34 (18) | 740 (1525) |
| 0.03     | 0.03     | 56 (32)    | 296 (201)    | 56 (32)    | 374 (230)  | 56 (32) | 743 (1337) |
| 0.02     | 0.02     | 130 (55)   | 337 (318)    | 130 (55)   | 372 (241)  | 130 (55) | 742 (1452) |
| 0.01     | 0.01     | 250 (159)  | 439 (2155)   | 250 (159)  | 347 (282)  | 230 (350) | 744 (1991) |

Next is an experiment on the accidents.dat dataset for min-sup 0.6 - 1 and min-conf 0.6 - 1 with the results shown in table 2.

Table 2. Mining results accidents.dat AR (Association Rule), L (Frequent itemsets, mb (memory), ms (millisecons)).

| Min-Supp | Min-Conf | Apriori AR | FP-Growth AR | Apriori-Close AR |
|----------|----------|------------|--------------|-----------------|
|          |          | AR (L) (ms) | L (mb) (ms) | AR (L) (ms) | L (mb) (ms) | AR (L) (ms) | L (mb) (ms) |
| 1        | 1        | 0 (0)      | 886 (914)    | 0 (0) | 855 (1713) | 0 (0) | 809 (903) |
| 0.9      | 0.9      | 180 (31)   | 849 (3223)   | 180 (31) | 788 (1671) | 180 (31) | 1029 (921) |
| 0.8      | 0.8      | 1432 (149) | 1015 (3592)  | 1432 (149) | 757 (1800) | 1432 (149) | 840 (1049) |
| 0.7      | 0.7      | 5226 (529) | 1165 (11798) | 5226 (529) | 951 (2015) | 5226 (529) | 842 (1471) |
| 0.6      | 0.6      | 54170 (2674) | 850 (39171) | 54170 (2074) | 1014 (2361) | 54170 (2074) | 714 (2950) |

6. Conclusions
The final result of this research produces itemsets association value Rule (AR), itemsets Frequent value (L), memory (mb) and milliseconds (ms) with the same number of all algorithms, except that there is a difference between mb and ms. Then the algorithmic process pattern for the dataset used has a significant difference. For retail algorithm data with the fastest processing time, the FP-Growth-AR algorithm is more than 85% compared to Apriori-AR, followed by Apriori-AR and Apriori-Close-AR. For detailed memory usage results the Apriori-AR algorithm is lighter and outperforms. In the accident data the experimental results show that FP-Growth-AR is a little more memory efficient and the fastest process is Apriori-Close, while the Apriori-AR with the longest processing time. The accuracy of the results of memory usage and processing time must still be tested on other data patterns that are more complex so that there will be shortcomings from previous studies.

References
[1] Agrawal, R., & Srikan, R 1994 Fast algorithms for mining association rules in large databases. VLDB ’94: Proceedings of the 20th International Conference on Very Large Data Bases (pp. 487-499). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
[2] Muliono, R. 2017. Analisis Efisiensi Algoritma Data Mining. Seminar Nasional Teknologi Informatika (Semantika), J. pp. 117-123.
[3] Bhandari, A., A. Gupta, and D. Das 2015 Improvised apriori algorithm using frequent pattern tree for real time applications in data mining, in Procedia Computer Science. p. 644-651.
[4] Agrawal, R., Imielinski, T. & Swami, A 1993 Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD Conference, pp. 207-216.
[5] Darshan. 2014. Improved Apriori Algorithm for mining Association Rules. International Journal of Information Technology and Computer Science (IJITCS'07) pp.15-23

[6] Sinthuja, M., Aruna, P., & Puvirasan, N 2017 Experimental Evaluation of Apriori and Equivalence Class Clustering and Bottom up Lattice Traversal (Eclat) Algorithms.

[7] R. Agrawal and R. Srikant, 1994 Fast algorithms for mining association rules, Proceedings of the 20th Very Large Databases Conference (VLDB’94), Santiago de Chile, Chile.

[8] Vanitha, K., & Santhi, R. 2011 Evaluating the Performance of Association Rule Mining Algorithms.

[9] Padhy, N., Mishra, P., & Panigrahi, R 2012 The Survey of Data Mining Applications And Feature Scope. CoRR, abs/1211.5723.

[10] Muliono, R. 2017, Implementasi Algoritma Apriori Pada Data Benchmark Kosarak Dan Mushrooms, Journal of Informatics and Telecommunication Engineering, 1(1) :34-41

[11] Lu, Y., Richter, F., & Seidl, T 2018 Efficient Infrequent Itemset Mining Using Depth-First and Top-Down Lattice Traversal. In International Conference on Database Systems for Advanced Applications (pp. 908-915). Springer, Cham.

[12] Gunadi, G., & Sensuse, D. I 2016 Penerapan metode data mining market basket analysis terhadap data penjualan produk buku dengan menggunakan algoritma apriori dan frequent pattern growth (fp-growth): studi kasus percetakan pt. Gramedia. Telematika MKOM, 4(1), 118-132.

[13] Larose, D.T 2005 Discovering Knowledge in Data: An Introduction to Data Mining. John Willey & Sons, Inc.

[14] Lausch, A., Schmidt, A., Tischendorf, L 2014 Data mining and linked open data - New perspectives for data analysis in environmental research Ecol. Model. p 13.

[15] I.H. Witten F. Eibe M.A. Hall 2011 Data Mining Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers.

[16] Inthasone S., Pasquier N., Tettamanzi A.G.B., da Costa Pereira C 2014 The BioKET Biodiversity Data Warehouse: Data and Knowledge Integration and Extraction. In: Blockeel H., van Leeuwen M., Vinciotti V. (eds) Advances in Intelligent Data Analysis XIII. IDA 2014. Lecture Notes in Computer Science, vol 8819. Springer, Cham.

[17] Mariscal, G., Marban, O., & Fernandez, C 2010 A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review, 25(2), 137-166.

[18] M. J. Zaki 2000 Generating non-redundant association rules. in Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. New York. p 34-43.

[19] Frequent Itemset Mining Dataset Repository by IBM Almaden Quest Research Group (Online) http://fimi.ua.ac.be/data/. (Accessed at 03 October 2018).

[20] Fournier-Viger, P., Lin, C.W., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z., Lam, H. T. (2016). The SPMF Open-Source Data Mining Library Version 2. Proc. 19th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2016) Part III, Springer LNCS 9853, pp. 36-40.