Association rules with apriori algorithm and hash-based algorithm

R D Yulanda1*, S Wahyuningsih1 and F D T Amijaya2

1Departement of Statistics, Faculty of Mathematics and Natural Sciences, Mulawarman University, Samarinda, Indonesia
2Department of Mathematics, Faculty of Mathematics and Natural Sciences, Mulawarman University, Samarinda, Indonesia

*Corresponding Author: rskylnda@yahoo.com

Abstract. The purpose of this research is to determine the pattern of hijacking cases in Samarinda by using association rules and comparing apriori and hash-based. The data used for this research are secondary data from POLRESTA Samarinda which is processed using association rules algorithm with apriori and hash-based with minimum support of 10% and minimum confidence of 70%. The result of this research, in apriori, the pattern is often formed as many as 4 5-itemset combination while the hash-based obtained pattern that often formed by 3 3-itemset combination. In 2-itemset that is if the victim is female and young, then have a possibility of 46.67% chance of experiencing minor injuries with confidence of 90.9%, in 3-itemset if the victim is in location 1 and young, then have a possibility 53.33% chance of experiencing minor injuries with confidence of 88.89%, in 4-itemset if the victim in location 2, female and young then have a possibility 40% chance of experiencing minor injuries with confidence of 100%, in 5-itemset if the victim in location two, quite, female and young then have a possibility 26.67% chance of experiencing minor injuries with confidence of 100%.

1. Introduction

Association rules is a data mining method used to find patterns that illustrate the strength of relationships in data. Association rules have two basic methods, there are frequent itemset generation to find all itemset meet the threshold minute up and rule generation to extract high confidence rules called a strong rule [1].

Association rules have many algorithms. The algorithm of association rules, among others, apriori, hash-based, generalized sequential pattern (GSP) and FP-Growth. From several algorithms, the well-known are apriori and hash-based. So, in this study, the algorithms used are apriori and hash-based.

Association rules technique is often referred to as market basket analysis because generally association rules is used to find out patterns in spending at market or supermarket. However, the application turns out that association rules have also been carried out in various fields, among others, research on determining the pattern of traffic accident relationship uses association rules method with apriori algorithm, and implementation of hash-based algorithm on association rules to determine frequent itemset case studies of seafood restaurants.
Hijacking is one of the cruel crimes because repression is a crime that not only involved the loss of the victim’s valuables but also about the victim’s life-threatening. Hijacking is an act of seizing victim’s goods in the middle of the road by stopping their victims accompanied by violence. Based on the description above, the authors are interested in reviewing the analysis by taking case research of Hijacking in Samarinda by using Association rules method with apriori algorithm and hash-based algorithm.

2. Literature citation

2.1. Association rules
Association rules is a data mining technique to find associative rules between the combination of items. Association rules became famous because its application to analyzing the contents of shopping carts at the supermarket, association rules is also often referred to as market basket analysis.[2] The basic methodology of association rules is divided into 2 stages:

1. Analyze high-frequency patterns
This stage uses apriori algorithm by looking for item combinations that meet the minimum requirements of the support value in the database. Support value of an item:

\[ \text{Support (A)} = \frac{\sum \text{transaction contains A}}{\sum \text{transaction}} \]  

(1)

The 2 items support value is obtained from the following formula:

\[ \text{Support (A, B)} = \frac{\sum \text{transaction contains A and B}}{\sum \text{transaction}} \]  

(2)

2. Formation of association rules
After all high-frequency patterns are found, then look for association rules that meet the minimum requirements for confidence by calculating association confidence rules.[5]

\[ \text{Confidence (B | A)} = \frac{\sum \text{transaction contains A and B}}{\sum \text{transaction contains A}} \]  

(3)

In association rules, besides support and confidence, there are other important parameters, namely the lift ratio. Lift ratio is a value that shows the validity of the transaction process. Lift ratio can be calculated by the formula:

\[ \text{Lift (A, B)} = \frac{\text{Support (A, B)}}{\text{Support (A)} \times \text{Support (B)}} \]  

(4)

If itemset >2 then lift ratio calculated with the formula:

\[ \text{Lift (A, B, C)} = \frac{\text{Confidence (C | A B)}}{\text{Support (C)}} \]  

(5)

2.2 Apriori Algorithm
Apriori algorithm is a very basic algorithm in finding frequent itemset from large database transactions. Apriori algorithm name is taken from the fact that this algorithm uses prior knowledge from frequent itemset for the next iteration process. Apriori algorithm steps are as follows:
1. Set $k = 1$ (points to 1-itemset)
2. Calculate all $k$-itemset (itemset that has $k$-items)
3. Calculate the support of all itemset candidates. Select the itemset based on the minimum support calculation
4. Combine all items to produce candidate itemset $k + 1$
5. Set $k = k + 1$
6. Repeat steps 3-5 until there are no longer itemset that can be formed
7. Make the final set of itemset by creating a union of all $k$-itemset. [3]

2.3 Hash-Based Algorithm

Hash-based algorithm can be a solution to determine the frequent of candidate itemset optimally. This algorithm reduces the number of $k$-itemset candidate at the beginning. The number of itemset generated can be smaller so that database scanning that is performed to determine the next itemsets becomes more efficient.

1-itemset hashing process using the formula:

$$h(x) = \left(\text{order of item } x\right) \mod n$$

for 2-itemset using the formula:

$$h(x) = \left(\left(\text{order of item } x\right) \times 10 + \text{order of item } y\right) \mod n$$

for 3-itemset using the formula:

$$h(x) = \left(\left(\text{order of item } x\right) \times 100 + \left(\text{order of item } y\right) \times 10 + \text{order of item } z\right) \mod n$$

It can be seen that the formula pattern is:

$$h(x) = \left(\sum_{a=1}^{k} \text{order of item } x_a + i^{a-1}\right) \mod n$$

with

$$n = 2m + 1$$

The steps of the hash-based algorithm are as follows:
1. Create an order of item table
The order of item table is a table that is used to code by numbering sequentially on an item. The order of item table can be seen in Table 1.

| Item | Order |
|------|-------|
| $I_1$ | 1     |
| $I_2$ | 2     |
| $I_3$ | 3     |
| ...  | ...   |
| $I_j$ | $j$   |

2. Set $k = 1$ (points to 1-itemset)
3. Calculate values
4. Create a hash table

The hash table is used to filter itemsets that will be used in the next iteration. The hash table consists of addresses, itemset, links, support values, confidence values, and bit vector values. The hash table can be seen in Table 2.
5. Make a frequency table
The frequency table is a table that is used to view the frequency of an itemset in a transaction. The frequency table can be seen in Table 3.

| Itemset | Support |
|---------|---------|
| 10      | S10     |
| 3       | S3      |
| ...     | ...     |
| 2       | S2      |

6. Determine the combination
7. Repeat steps 2 through 5 until all bit vectors are zero. [4]

3. Methods
The data source used in this research is the data cases of hijacking in Samarinda as many as 15 cases. The variables used in this research are placed the incident, condition, age, gender, and hijacking. Hijacking that used in this research are death, severe injuries, and minor injuries. The method used is Descriptive Statistics, Association Rules with Apriori Algorithm and Hash-Based Algorithm.

4. Results and discussion

4.1. Descriptive statistics

| No. | Characteristics | Percentage |
|-----|-----------------|------------|
| 1   | Location 1      | 20%        |
|     | Location 2      | 80%        |
| 2   | Crowded         | 20%        |
|     | Quite           | 80%        |
| 3   | Male            | 53.33%     |
|     | Female          | 46.67%     |
|     | Child           |            |
| 4   | Young           | 73.33%     |
|     | Adult           | 20%        |
|     | Elderly         | 6.67%      |
|     | Death           | -          |
| 5   | Severe injuries | 20%        |
|     | Minor injuries  | 80%        |

Based on Table 4 obtained the percentage of the hijacking at Samarinda. Most victims of the hijacking were driving at location 2 as many as 80%, who drive when the situation was quite is 80%, was female gender is 53.33%, victims in young age were 73.33%. The rates of abuse experienced by victims mostly suffered minor injuries as much as 80%, severe injuries as much as 20% and no victims died. This shows that the case of rape in Samarinda in 2017 tends to cause minor injuries to the victim not to make the victim die.
4.2. Apriori Algorithm

Association rules in the case of hijacking in Samarinda in 2017 use apriori algorithm with the help of R 3.5.0 software. The itemset selection criteria used by researchers are based on research conducted by Hakim et al. [2]. Can be seen in Table 5.

| Table 5. Selection Criteria for Itemset in the Process of Apriori Algorithms |
|-----------------------------|-----------------|-----------------|-----------------|
| Criteria | Support | Confidence | Lift | Decision (Y) |
| Value | ≥ 0,1 | ≥ 0.7 | ≥ 1 | Level of hijacking |

Rules analysis is done by looking at whether the support, confidence, lift, and decision (Y) values meet the criteria or not.

- 2-itemset

Based on the resulting output from software R, 26 combinations with 2-itemset are obtained, there are only 4 itemset combinations that meet the criteria. The result can be seen in Table 6.

| Table 6. Output Association Rules with 2-itemset Apriori Algorithm |
|-----------------------------|-----------------|-----------------|-----------------|-----------------|
| No. | Condition (X) | Decision (Y) | Support | Confidence | Lift | Rules |
| 1 | location 1 | minor injuries | 0.2000 | 1.0000 | 1,2500 | yes |
| 2 | crowded | minor injuries | 0.2000 | 1.0000 | 1,2500 | yes |
| 3 | female | minor injuries | 0.4667 | 1.0000 | 1,2500 | yes |
| 4 | young | minor injuries | 0.6667 | 0.9091 | 1,1364 | yes |

Based on Table 5, there are 4 rules of 2-itemset association are obtained as follows:

1. "If the incident happened at location 1 then the victim suffered minor injuries"
2. "If the condition is crowded then the victim suffered minor injuries"
3. "If the gender is female then the victim suffered minor injuries"
4. "If the age is young then the victim suffered minor injuries"

- 3-itemset

Based on the resulting output from software R, 59 combinations with 3-itemset are obtained, there are only 11 itemset combinations that meet the criteria. The result can be seen in Table 7.

| Table 7. Output Association Rules with 3-itemset Apriori Algorithm |
|-----------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| No. | Condition (X) | Decision (Y) | Support | Confidence | Lift | Rules |
| 1 | location 1, male | minor injuries | 0.1333 | 1.0000 | 1,2500 | yes |
| 2 | location 1, young | minor injuries | 0.1333 | 1.0000 | 1,2500 | yes |
| 3 | location 1, quiet | minor injuries | 0.1333 | 1.0000 | 1,2500 | yes |
| 4 | crowded, female | minor injuries | 0.1333 | 1.0000 | 1,2500 | yes |
| 5 | crowded, young | minor injuries | 0.1333 | 1.0000 | 1,2500 | yes |
| 6 | location2, crowded | minor injuries | 0.1333 | 1.0000 | 1,2500 | yes |
| 7 | female, young | minor injuries | 0.4667 | 1.0000 | 1,2500 | yes |
| 8 | location 2, female | minor injuries | 0.4000 | 1.0000 | 1,2500 | yes |
| 9 | quiet, female | minor injuries | 0.3333 | 1.0000 | 1,2500 | yes |
| 10 | location 2, young | minor injuries | 0.5333 | 0.8889 | 1,1111 | yes |
| 11 | quiet, young | minor injuries | 0.5333 | 0.8889 | 1,1111 | yes |

Based on Table 7, there are 11 rules of 3-itemset association are obtained as follows:

1. "If the incident happened at location 1 and the gender is male then the victim suffered minor injuries"
2. "If the incident happened at location 1 and the age is young then the victim suffered minor injuries"
3. "If the incident happened at location 1 and the condition is quiet then the victim suffered minor injuries"
4. "If the condition is crowded and the gender is female then the victim suffered minor injuries"
5. "If the condition is crowded and the age is young then the victim suffered minor injuries"
6. "If the incident happened at location 2 and the condition is crowded then the victim suffered minor injuries"
7. "If the gender is female and the age is young then the victim suffered minor injuries"
8. "If the incident happened at location 2 and the gender is female then the victim suffered minor injuries"
9. "If the gender is female and the condition is crowded then the victim suffered minor injuries"
10. "If the incident happened at location 2 and the age is young then the victim suffered minor injuries"
11. "If the condition is quite and the age is young then the victim suffered minor injuries"

- 4-itemset

Based on the resulting output from software R, 49 combinations with 4-itemset are obtained, there are only 8 itemset combinations that meet the criteria. The result can be seen in Table 8.

| No. | Condition (X) | Decision (Y) | Support | Confidence | Lift | Rules |
|-----|---------------|--------------|---------|------------|------|-------|
| 1   | location 1, quiet, young | minor injuries | 0.1333  | 1.0000     | 1.2500 | yes   |
| 2   | crowded, female, young | minor injuries | 0.1333  | 1.0000     | 1.2500 | yes   |
| 3   | location 2, crowded, female | minor injuries | 0.1333  | 1.0000     | 1.2500 | yes   |
| 4   | location 2, crowded, young | minor injuries | 0.1333  | 1.0000     | 1.2500 | yes   |
| 5   | location 2, female, young | minor injuries | 0.4000  | 1.0000     | 1.2500 | yes   |
| 6   | quiet, female, young | minor injuries | 0.1333  | 1.0000     | 1.2500 | yes   |
| 7   | location 2, quiet, female | minor injuries | 0.2667  | 1.0000     | 1.2500 | yes   |
| 8   | location 2, quiet, young | minor injuries | 0.4000  | 0.8571     | 1.0714 | yes   |

Based on Table 8, there are 8 rules of 4-itemset association are obtained as follows:
1. "If the incident happened at location 1, the condition is quiet and the age is young then the victim suffered minor injuries"
2. "If the condition is crowded, the gender is female and the age is young then the victim suffered minor injuries"
3. "If the incident happened at location 2, the condition is crowded and the gender is female then the victim suffered minor injuries"
4. "If the incident happened at location 2, the condition is crowded and the age is young then the victim suffered minor injuries."
5. "If the incident happened at location 2, the gender is female and the age is young then the victim suffered minor injuries"
6. "If the condition is quiet, the gender is female and the age is young then the victim suffered minor injuries"
7. "If the incident happened at location 2, the condition is quiet and the gender is female then the victim suffered minor injuries"
8. "If the incident happened at location 2, the condition is quiet and the age is young, then the victim suffered minor injuries"

- 5-itemset

Based on the resulting output from software R, 11 combinations with 5-itemset are obtained, there are only 2 itemset combinations that meet the criteria. The result can be seen in Table 9.
Based on Table 9, there are 2 rules of 5-itemset association are obtained as follows:
1. "If the incident happened at location 2, the condition is crowded, the gender is female and the age is young then the victim suffered minor injuries"
2. "If the incident happened at location 2, the condition is quiet, the gender is female and the age is young then the victim suffered minor injuries"

### 4.3. Hash Based Algorithm

Association rules in the case of hijacking in Samarinda in 2017 use hash-based algorithm with the help of Lazarus 1.6.0 software. The selection criteria for itemset used in this study can be seen in Table 10.

| Criteria | Support | Confidence | Lift | Bit Vector | Decision (Y) |
|----------|---------|------------|------|------------|--------------|
| Value    | ≥ 0.1   | ≥ 0.7      | ≥ 1  | 1          | Level of hijacking |

Rules analysis is done by looking at whether the support, confidence, lift, and decision (Y) values meet the criteria or not.

- **2-itemset**

Based on the resulting output from software R, 45 combinations with 2-itemset are obtained, there are 4 itemset combinations that meet the criteria. The result can be seen in Table 11.

| No. | Condition (X) | Decision (Y) | Support | Confidence | Lift | Bit vector | Rules |
|-----|---------------|--------------|---------|------------|------|------------|-------|
| 1   | location 1    | minor injuries | 0.2000  | 1.0000     | 1.2500| 1          | Yes   |
| 2   | crowded       | minor injuries | 0.2000  | 1.0000     | 1.2500| 1          | Yes   |
| 3   | female        | minor injuries | 0.4667  | 1.0000     | 1.2500| 1          | Yes   |
| 4   | young         | minor injuries | 0.6667  | 0.9091     | 1.1364| 1          | Yes   |

Based on Table 11, there are 4 rules of 2-itemset association are obtained as follows:
1. "If the incident happened at location 1 then the victim suffered minor injuries"
2. "If the condition is crowded then the victim suffered minor injuries"
3. "If the gender is a female then the victim suffered minor injuries"
4. "If the age is young then the victim suffered minor injuries"

- **3-itemset**

Based on the resulting output from software R, 35 combinations with 3-itemset are obtained, there are only 11 itemset combinations that meet the criteria. The result can be seen in Table 12.

| No. | Condition (X) | Decision (Y) | Support | Confidence | Lift | Bit vector | Rules |
|-----|---------------|--------------|---------|------------|------|------------|-------|
| 1   | location 2, young | minor injuries | 0.5333  | 0.8889     | 1.1111| 1          | Yes   |
| 2   | female, young  | minor injuries | 0.4667  | 1.0000     | 1.2500| 1          | Yes   |
| 3   | quiet, young   | minor injuries | 0.5333  | 0.8889     | 1.1111| 1          | Yes   |
Based on Table 12, there are 11 rules of 3-itemset association are obtained as follows:
1. "If the incident happened at location 2 and the age is young then the victim suffered minor injuries"
2. "If the gender is female and the age is young then the victim suffered minor injuries"
3. "If the condition is quite and the age is young then the victim suffered minor injuries"

5. Conclusion
Based on the research that has been done, the results of this research can be the conclusion as follows:
1. Based on the research that has been done on the association rule method with a priori algorithm, obtained 5 itemset and the tendency of the relationship patterns that are often formed is as follows:
   a. In the 2-itemset that is "If the age is young then the victim has a 66.67% chance suffered a level of hijacking namely minor injuries with a 90.9% confidence level"
   b. In the 3-itemset that is "If the incident at location 2 and the age is young then the victim has a 53.33% chance suffered a level of hijacking namely minor injuries with an 88.89% confidence level"
   c. In the 4-itemset that is "If the incident at location 2, the gender is female and the age is young then the victim has a 40% chance suffered a level of hijacking namely minor injuries with a 100% confidence level"
   d. In the 5-itemset that is "If the incident at location 2, the condition is quiet, the gender is female and the age is young then the victim has a 26.67% chance to experience a level of hijacking namely minor injuries with a 100% confidence level"

While the hash-based algorithm obtained 3-itemset and obtained the tendency of relationship patterns that are often formed, namely:
   a. In the 2-itemset that is "If the age is young then the victim has a 66.67% chance of experience a level of hijacking namely minor injuries with a 90.9% confidence level"
   b. In the 3-itemset that is "If the incident at location 2 and the age is young then the victim has a 53.33% chance to experience a level of hijacking namely minor injuries with an 88.89% confidence level"

2. Based on the research that has been done there are differences in association rules which between apriori algorithm and hash-based algorithm. In apriori algorithm, there are 4 rules on 2 itemset, 11 rules on 3 itemset, 8 rules on 4 itemset and 2 rules on 5 itemset. In the hash based algorithm, there are 4 rules in 2 itemset and 3 rules in 3 itemset. It can be concluded that the based hash algorithm is more efficient because it reduces itemset pairs that are better than apriori algorithm.

Acknowledgments
The research would have taken far longer to complete without the encouragement of many others. It is a delight to acknowledge those who have supported me. First, I would thank Mrs. Dr. Sri Wahyuningsih, M.Si., Mr. Fidia Deny Tisna Amijaya, M.Si., Mrs. Memi Nor Hayati, M.Si., Mr. Nanda Arista Rizki, M.Si for guidance, advice, and input for the author. Then, I would thank my parents, my brothers, my sisters and also my friends for support until this research was complete.

References
[1] Prasetyo E 2012 Data Mining Konsep dan Aplikasi Menggunakan MATLAB (Yogyakarta: Penerbit Andi) pp 6.
[2] Kusrini and Luthfi E T 2009 Algoritma Data Mining (Yogyakarta: Penerbit Andi) pp 149-50.
[3] Hakim L and Fauzy A 2015 Penentuan Pola Hubungan Kecelakaan Lalu Lintas Menggunakan Metode Association Rules dengan Algoritma Apriori on Jurnal University Research Colloquium pp 73-81.
[4] Ramadhan F 2017 Implementasi Algoritma Hash Based Terhadap Aturan Asosiasi untuk Menentukan Frequent Itemset Studi Kasus Rumah Makan Seafood “Kita” on Seminar Nasional Teknologi Informasi dan Multimedia pp 97-101.
[5] Virgiawan D M and Mukhlas I 2013 Aplikasi Association Rule Mining untuk Menemukan Pola Pada Data Nilai Mahasiswa Matematika ITS on Jurnal Sains dan Seni Pomits I pp 1-6.