Association Rules Selection Approach Based on Interesting Measures

Pakorn Leesuthipornchai
Department of Computer Science, Faculty of Science and Technology,
Thammasat University, Thailand
pakornl@cs.tu.ac.th

ABSTRACT
In recent years, the size of data collection becomes growing significantly. The knowledge results in term of association rules obtained from the set of data are numerous and difficult to select. This paper proposed the approach to select the interesting subsets of association rules from the big association results. The selection criterion is based on well-known interesting measures that are confidence, coverage, leverage and lift. The original association rules for Thai stock market in the period of April 10, 2013 to September 5, 2014 are investigated and applied to the selection approach. The selection approach reduces 246,084 association rules into 10 rules.

KEYWORDS
Association Rule, Interesting Measure, Selection Approach, Association Mining, Stock Index Data

1 ASSOCIATION MINING
Association mining is introduced by Rakesh Agrawal et al. [1] for discovering regularities between the set of products from transactional data records in supermarkets. The example of association rules is \{onions, potatoes\} → \{burger\} that means if a customer buys onions and potatoes together, he/she is likely to buy a burger. The obtained association rule will be used for making decision about marketing activities such as product placement.

This paper considers in the selection problem when the set of association rules is large and the candidate choices are numerous. All of them are significant. No one rule is better than the other. The selection approach is based on interesting measures that are confidence, coverage, leverage and lift.

In this paper, the data [2] from the Thai stock market in the period of April 10, 2013 to September 5, 2014 are investigated. The knowledge from the data set in term of association rules are discovered by using the well-known algorithm (i.e., Apriori). Then the association rules are applied to the selection approach.

The remainder of this paper is organized as follows. Section 2 discusses association mining research works. Section 3 provides interesting measures formulation. Section 4 describes about association rules obtained from stock index data. Section 5 shows the selection approach. Section 6 discusses the obtained association results. Lastly, section 7 concludes the research work and describes the contribution.

2 ASSOCIATION MINING RESEARCH
Association mining is applied in many applications as followings.

Rakesh Agrawal et al. [1] proposed mining association rules between sets of items in large databases. The paper introduces the algorithm to generate association rules from buying items in the database.
Sun Jing et al. [3] proposed the study of association rule mining on technical action of ball games. The paper applies the association technique (i.e., Apriori algorithm) to find the relationship among technical actions of ball games.

Jiang Hong-bo and Yang De-li [4] proposed the application research on fast discovery of association rules based on air transportation. The research applies the association technique to discover the relationship of the large complex data in air transportation.

Caner Erden and Fatih Tüysüz [5] proposed the research that applied association rules to discover the relationship in cause of traffic accidents among the collection of data set about people who involved in car accidents with fatalities.

Sung Hong Na and So Yung Sohn [6] proposed the forecasting changes in Korea composite stock price index (KOSPI) using association rules. The research applies the association technique to find the relation of price direction (up or down) of stock market indices in Korea, USA, Europe, and Asia. From their association rules [6], the KOSPI index tends to have the same price direction as stock indices in USA and Europe, whereas, the KOSPI tends to have the price direction opposite to stock indices in East Asian countries such as Hong Kong and Japan indices.

3 INTERESTING MEASURES

The interesting measures are described in detail in [7]. This paper focuses on the basic measures that are confidence, support, coverage, leverage and lift. The descriptions of these measures are showed as followings.

Let $T=\{t_1, t_2, \ldots, t_n\}$ be the set of transaction records. $X\rightarrow Y$ is the association rule. $X$ and $Y$ are the item sets where $X \cap Y = \emptyset$. $X$ is antecedent (left hand side, LHS) of the rule and $Y$ is consequent (right hand side, RHS) of the rule. For example, the association rule is \{onions, potatoes\} $\rightarrow$ \{burger\}. The item set $X$ is \{onions, potatoes\} and the item set $Y$ is \{burger\}.

3.1 Support
Support is the proportion of transactions which contains the item set $X$ as shown in Eq. (1).

$$\text{sup}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$ (1)

3.2 Confidence
Confidence is the probability of the support of item set that contains antecedent $X$ and consequent $Y$ over the support of item set $X$ as shown in Eq. (2).

$$\text{con}(X \rightarrow Y) = \frac{\text{sup}(X \rightarrow Y)}{\text{sup}(X)} = \frac{\text{sup}(X \cup Y)}{\text{sup}(X)}$$ (2)

3.3 Coverage
Coverage is the antecedent ($X$) support. The coverage indicates that how often does the association rule $X \rightarrow Y$ is applied. The coverage formulation is showed in Eq. (3).

$$\text{cov}(X \rightarrow Y) = \text{sup}(X)$$ (3)

3.4 Leverage
Leverage is the difference between the support that $X$ and $Y$ appear together and the support that $X$ and $Y$ appear independently. Leverage indicates that how many items $X$ and $Y$ are sold together more than the independent sells. The leverage formulation is showed in Eq. (4).

$$\text{lev}(X \rightarrow Y) = \text{sup}(X \rightarrow Y) - \left(\text{sup}(X) \cdot \text{sup}(Y)\right)$$
$$= \text{sup}(X \cup Y) - \left(\text{sup}(X) \cdot \text{sup}(Y)\right)$$ (4)
3.5 Lift
Lift is the proportion of the support that X and Y appear together over the support that X and Y appear independently. Lift measures that how many times items X and Y are sold together over than the independent sells. The lift formulation is showed in Eq. (5).

\[
lift(X \rightarrow Y) = \frac{\text{con}(X \rightarrow Y)}{\text{sup}(Y)} = \frac{\text{sup}(X \cup Y)}{\text{sup}(X)\text{sup}(Y)}
\]  

(5)

4 STOCK INDEX ASSOCIATION RULES
In this paper, the data set was gathered from the Stock Exchange of Thailand (SET). The transactional data of SET50 index were captured every half an hour in the period of April 10, 2013 to September 5, 2014. The data set consists of 3,892 records. Each record consists of trading date, trading time, opening, high, low and closing price. The currency of the price is Thai Baht.

The transactional data were preprocessed before applied to the Apriori algorithm. For parameter setting, the paper considered that the minimum support count is 1 from 3,892 (i.e., the percentage of minimum support is 0.0257% that means a large number of frequent item set is considered). The minimum confidence is 100%. The number of association rules obtained from Apriori algorithm with 100% confidence is 246,084. The number of rules that the RHS (i.e., PCH) is in range A, B, C, D are 2088, 27116, 207960 and 8920 respectively. The summation of the number of rules is showed in Table 3.

From Table 1, the rule number 1 shows the condition that is RSI=7=D MACD=C MACD-fst=B CHK=C 577 ===>$ PCH=C 577 \Rightarrow$ conf:(1). The meaning is when RSI indicator is more than 75, MACD indicator is between -0.56 to 10.45, MACD-fst indicator is between -7.74 and 1.33, and CHK indicator is between -94,763.73 and 85,248.43. The percentage of price change is oscillated between -0.60% and 1.96%. The number 577 in the LHS is the support of item set \{RSI=D, MACD=C, MACD-fst=B, CHK=C\} and the number 577 in the RHS is the support value of item set \{RSI=D, MACD=C, MACD-fst=B, CHK=C, PCH=C\}. The confidence is 557 (from RHS) / 557 (from LHS) = 1.

Table 1. Stock Index Association Rules

<table>
<thead>
<tr>
<th>Rule Number</th>
<th>Condition</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RSI7=D MACD=C MACD-fst=B CHK=C 577 ===&gt;$ PCH=C 577 \Rightarrow$</td>
<td>conf:(1)</td>
</tr>
<tr>
<td>2</td>
<td>RSI7=D MACD=C MACD-fst=B CHK=C Vol=A 566 ===&gt;$ PCH=C 566 \Rightarrow$</td>
<td>conf:(1)</td>
</tr>
<tr>
<td>3</td>
<td>RSI7=D PROC7=C MACD-fst=B Vol=A 448 ===&gt;$ PCH=C 448 \Rightarrow$</td>
<td>conf:(1)</td>
</tr>
<tr>
<td>4</td>
<td>RSI7=D PROC7=C MACD=C MACD-fst=B Vol=A 421 ===&gt;$ PCH=C 421 \Rightarrow$</td>
<td>conf:(1)</td>
</tr>
<tr>
<td>...</td>
<td>246084. TIME=C SMA7=C SMA14=B RSI7=D PROC7=C MACD=D MACD-fst=D STH7=C STH-fst=D CHK=C Vol=A 2 ===&gt;$ PCH=D 2 \Rightarrow$</td>
<td>conf:(1)</td>
</tr>
</tbody>
</table>

Table 2. Ranges of Attribute Values

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Range</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>10:00-11:00</td>
<td>11:01-12:30</td>
<td>14:30-15:30</td>
<td>15:31-16:30</td>
<td></td>
</tr>
<tr>
<td>SMA7</td>
<td>&lt; 890.40</td>
<td>890.40 to 957.76</td>
<td>957.77 to 1025.12</td>
<td>&gt; 1025.12</td>
<td></td>
</tr>
<tr>
<td>SMA14</td>
<td>&lt; 892.94</td>
<td>892.94 to 959.28</td>
<td>959.29 to 1025.61</td>
<td>&gt; 1025.61</td>
<td></td>
</tr>
<tr>
<td>RSI7</td>
<td>&lt; 25</td>
<td>25 to 50</td>
<td>50.01 to 75</td>
<td>&gt; 75</td>
<td></td>
</tr>
<tr>
<td>PROC</td>
<td>&lt; -2.94</td>
<td>-2.94 to -0.99</td>
<td>0.10 to 3.11</td>
<td>&gt; 3.11</td>
<td></td>
</tr>
<tr>
<td>MACD</td>
<td>&lt; -11.55</td>
<td>-11.55 to -0.55</td>
<td>-0.56 to 10.45</td>
<td>&gt; 10.45</td>
<td></td>
</tr>
<tr>
<td>MACD-fst</td>
<td>&lt; -7.74</td>
<td>-7.74 to 1.33</td>
<td>1.34 to 10.80</td>
<td>&gt; 10.80</td>
<td></td>
</tr>
<tr>
<td>STH7</td>
<td>&lt; 25</td>
<td>25 to 50</td>
<td>50.01 to 75</td>
<td>&gt; 75</td>
<td></td>
</tr>
<tr>
<td>STH-fst</td>
<td>&lt; 25.68</td>
<td>25.68 to 49.55</td>
<td>49.56 to 73.41</td>
<td>&gt; 73.41</td>
<td></td>
</tr>
<tr>
<td>CHK</td>
<td>&lt; -274,482</td>
<td>-274,482 to -94,763.72</td>
<td>-94,763.73 to 85,248.43</td>
<td>&gt; 85,248.43</td>
<td></td>
</tr>
<tr>
<td>Vol</td>
<td>&lt; 248,482</td>
<td>248,482 to 496,943</td>
<td>496,943 to 745,401</td>
<td>&gt; 745,401</td>
<td></td>
</tr>
<tr>
<td>PCH</td>
<td>&lt; -3.14</td>
<td>-3.14 to -0.59</td>
<td>-0.60 to 1.96</td>
<td>&gt; 1.96</td>
<td></td>
</tr>
</tbody>
</table>

The number of association rules obtained from Apriori algorithm with 100% confidence is 246,084. The number of rules that the RHS (i.e., PCH) is in range A, B, C, D are 2088, 27116, 207960 and 8920 respectively. The summation of the number of rules is showed in Table 3.
The large number of original association rules will be applied to the selection approach that will be described in the next section.

5 ASSOCIATION RULES SELECTION APPROACH

The association rules selection approach is calculated based on two options. The first assumption considers the rule with rare item has high priority. Rare item is the situation that the item set may appear infrequent but every time the cause appears, the result always happens. The second assumption considers the rule with popular item has high priority.

For the rare item situation, the rule with high confidence value will be reserved first then the lift, leverage and coverage values are considered respectively. For doing this, the interesting measures will be normalized to the probability values with 2 decimal digits (i.e., 0.00 to 1.00). The set of interesting measures (i.e., \{con, lift, lev, cov\}) will be multiplied with the place values that are 10^{-11}, 10^{-8}, 10^{-5}, and 10^{-2} respectively. The multiplication result is assigned to the selection score as shown in Table 5.

From Table 5, the top-\(k\) association rules with highest score will be selected. Suppose \(k=4\), rules no.1, no.2, no. 7 and no. 8 are selected because they are the top four highest selection score.

For the popular item situation, the rule with high confidence value will be reserved first then the leverage, coverage and lift values are considered respectively. The set of interesting measures (i.e., \{con, lev, cov, lift\}) will be multiplied with the place values that are 10^{-11}, 10^{-8}, 10^{-5}, and 10^{-2} respectively. The multiplication result is assigned to the selection score as shown in Table 6.

From Table 6, the top-\(k\) association rules with highest score will be selected. Suppose \(k=4\), rules no.5, no.6, no. 1 and no. 2 are selected because they are the top four highest selection score.
Table 6. Association Rules with Selection Score for Popular Item Consideration

<table>
<thead>
<tr>
<th>Association Rules</th>
<th>Interesting Measures</th>
<th>Selection Score (MSD to LSD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>{TIME=D, SMA7=C, MACD=A} → {PCH=A}</td>
<td>[1.00 0.00 0.00 1.00]</td>
<td>100 000 000 100</td>
</tr>
<tr>
<td>{TIME=D, SMA14=C, MACD=A} → {PCH=A}</td>
<td>[1.00 0.00 0.00 1.00]</td>
<td>100 000 000 100</td>
</tr>
<tr>
<td>{TIME=D, MACD=B, MACD-fst=B, CHK=B} → {PCH=B}</td>
<td>[1.00 0.00 0.00 0.03]</td>
<td>100 000 000 003</td>
</tr>
<tr>
<td>{TIME=D, PROC=B, MACD=B, MACD-fst=B, CHK=B} → {PCH=B}</td>
<td>[1.00 0.00 0.00 0.03]</td>
<td>100 000 000 003</td>
</tr>
<tr>
<td>{RSI7=D, MACD=C, MACD-fst=B, CHK=C} → {PCH=C}</td>
<td>[1.00 0.04 0.15 0.01]</td>
<td>100 000 000 015</td>
</tr>
<tr>
<td>{RSI7=D, MACD=C, MACD-fst=B, CHK=C, Vol=A} → {PCH=C}</td>
<td>[1.00 0.03 0.15 0.01]</td>
<td>100 000 000 015</td>
</tr>
<tr>
<td>{RSI7=D, MACD=C, MACD-fst=B, CHK=C} → {PCH=D}</td>
<td>[1.00 0.00 0.00 0.29]</td>
<td>100 000 000 029</td>
</tr>
<tr>
<td>{RSI7=D, MACD=C, MACD-fst=B, CHK=C, Vol=A} → {PCH=D}</td>
<td>[1.00 0.00 0.00 0.29]</td>
<td>100 000 000 029</td>
</tr>
</tbody>
</table>

For the selection criterion, the rule r will be selected to the top-k association rules even if 1) the selection score of rule r is greater than or equal to the lowest score of top-k rules and 2) the LHS of the rule r must not be the superset of the LHS of the rule k when the RHS of rule r is same set of the RHS of rule k.

For example, from Table 5, the top-4 rules are no.1, no.2, no.7 and no.8. The LHS of rule no.7 is {SMA7=C, MACD=D} and the LHS of rule no.8 is {SMA7=C, RSI7=D, MACD=D}. If rule no.7 is in the top-k rules, the rule no.8 is not selected because the LHS of rule no.8 is the superset of rule no.7 when the RHS of rules no.7 and no.8 are the same (i.e., PCH=D). The condition of rule no.7 is more general than the condition of rule no.8. The association rule with small set of condition (LHS) is preferred because it is easy to determine and likely to be applied.

In the other hand, if rule no.8 is in the top-k rules, the rule no.7 is selected because the LHS of rule no.7 is not the superset of rule no.8 when the RHS of rules no.7 and no.8 are the same (i.e., PCH=D). In this case, the rule no.8 will be replaced by the rule no.7. The overview of the selection approach is showed as follows.

For the rare item situation, the association rules from Table 5 are applied with the selection approach. The final top-4 rules are rules no.1, no.2, no.7 and no.3.

For the popular item situation, the association rules from Table 6 are applied with the selection approach. The final top-4 rules are rules no.5, no.1, no.2 and no.7.

6 ASSOCIATION RESULTS AFTER THE SELECTION APPROACH

This section discusses on the association results after the selection approach. The example of original association rules with 100% confidence value are showed in Table 1 (246,084 rules). This paper applied the selection approach for selecting the interesting association rules from 246,084 rules. The top-5 interesting rules from the selection approach with rare item and popular item considerations are showed in Table 7.
7 CONCLUSIONS

This paper proposed the approach to select the interesting subsets of association rules from the big association results. The selection criterion is based on well-known interesting measures that are confidence, coverage, leverage and lift. The two selection criteria were proposed that are rare item and popular item situations. For rare item consideration, the interesting measures are set in the order as confidence, lift, leverage, and coverage. For popular item consideration, the interesting measures are set in the order as confidence, leverage, coverage, and lift. The main research contribution is to propose the selection approach for large association results by setting interesting measures order. The paper showed that the original 246,084 association rules will be selected to only 10 interesting rules.

REFERENCES


