5

On the Selection of Meaningful Association Rules

Rangsipan Marukatat
Mahidol University
Thailand

1. Introduction

In recent years, data mining has been recognized as a powerful technique to extract hidden patterns from an enormous volume of data. These patterns can be expressed in many forms. One of them is as a set of association rules. From a transactional data set, an association rule "$x \rightarrow y$" (support = s%, confidence = c%) indicates the co-occurrence of items x and y in the same transaction, with certain levels of support and confidence. Association rule mining is useful in many applications. For example, in retailing, highly associated products can be identified and sold together as a special offer package (Svetina & Zupancic, 2005). Ma et al. (2003) extracted association rules from microbiology transactions. They detected outbreak of nosocomial infection, or infection acquired by patients during their hospital stay, from low-support and low-confidence rules.

A primeval but elegant association rule mining method, *Apriori* (Agrawal & Srikant, 1994), first discovers itemsets (or sets of data items) that satisfy a minimum support criterion. It then uses these itemsets to generate rules that satisfy a minimum confidence criterion. After Apriori, a number of advanced algorithms have been developed. The list includes Brin et al. (1997), Zaki et al. (1997), Liu et al. (1999b), Han et al. (2000), Yun et al. (2003), and Ko and Rountree (2005). Nevertheless, it is the original Apriori that is still the most popular one and becomes a standard function in many data mining software.

In real practices, data is usually rudimentary and the probability that each item occurs in a data set may be very low. Association rule mining might be performed several times, each with different sets of parameters, so that plenty of rules are generated and some useful, non-trivial ones can be spotted among them. This is where Apriori's simplicity is traded off. It straightforwardly delivers rules that pass the thresholds given by users, but lacks effective pruning mechanisms. It is up to the users to handle the usually overwhelming amount of rules afterwards.

There have been techniques for post-pruning or post-selecting the association rules. For example, Ableson and Glasgow (2003) proposed statistic-based pruning. Others attempted to identify general and specific rules. Once identified, specific rules could be pruned or kept separately for further analysis (Berrado & Runger, 2007; Liu et al., 1999a; Toivonen et al., 1995). Techniques that exploit semantics conveyed in the rules include Klemettinen et al. (1994), Ma et al. (2003), and Silberschatz and Tuzhilin (1996). According to Li and Sweeney (2005), rules were selected and combined to form a new rule that expressed the knowledge more thoroughly. But the selection was performed as part of the rule generation.

Source: Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca,
ISBN 978-3-902613-53-0, pp. 438, February 2009, I-Tech, Vienna, Austria

1.1 Chapter contribution

This chapter presents an alternative approach to the selection of association rules. It suggests using an already available method, Apriori, to generate association rules. Then, the rules are selected or pruned based on their degrees of semantic redundancy and patterns. The rest of the chapter is organized as follows.

- Section 2 introduces basics of association rule mining, the Apriori algorithm, and the post-mining of association rules. It reviews how rules are summarized, interpreted, and selected or pruned in other works.
- Section 3 describes semantic analysis and pattern analysis. The former classifies rules into four groups: strongly meaningless, weakly meaningless, partially meaningful, and meaningful. The latter prunes repetitive patterns and retains ones that convey the most information.
- Section 4 demonstrates how the semantic analysis and pattern analysis were applied to a real-world application, an analysis of traffic accidents in Nakorn Pathom, Thailand.
- Section 5 discusses the proposed techniques and identifies their drawbacks.
- Section 6 concludes the chapter.

2. Association rule mining

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of distinct items, and $D = \{T_1, T_2, \ldots, T_n\}$ be a transactional database. Each transaction T contains a subset of I. Table 1 shows an example data set and its binary representation.

Transaction ID	Items	Binary Representation				
		A	B	C	D	E
1	A, B, C	1	1	1	0	0
2	B, D	0	1	0	1	0
3	A, C, E	1	0	1	0	1
4	A, B, D, E	1	1	0	1	1
5	C, E	0	0	1	0	1

Table 1. An example data set

"$A \rightarrow B$" is an association rule, given the following conditions: $A \subset I$, $B \subset I$, and $A \cap B = \varnothing$. There are three common measures for an association rule. They are *support*, *confidence*, and *lift* (or *interest*). Support is the probability that both A and B occur in a transaction, i.e. $P(A \cap B)$. Confidence is the probability that B occurs in a transaction that A has occurred, i.e. $P(B \mid A)$ or $P(A \cap B)/P(A)$. Lift normalizes the confidence with the probability of B, i.e. $P(A \cap B)/(P(A) \times P(B))$. The lift equalling one implies that A and B are independent of one another.

As mentioned in the introduction, Apriori is a classic association rule mining algorithm that is incorporated in many data mining software. It performs two tasks: (1) generating itemsets that pass a minimum support threshold; and (2) generating rules that pass a minimum confidence threshold. For many users, finding the right thresholds is not easy because low support leads to abundant rules but high support may cause important rules to be missed. Moreover, rules with high support usually have low confidence, and vice versa. Algorithm 1 is one possible implementation that wraps the main tasks in a loop (University of Waikato, n.d.). The support threshold is gradually adjusted at the end of each loop iteration. Either confidence or lift can be used as a criterion for the rule generation.

Algorithm 1. Apriori

```
1.  // Parameters given by users are UpperMinSupport, LowerMinSupport, Delta,
2.  // Criterion, MinScore, and NumRules
3.  Set of association rules = ∅
4.  N = 0
5.  do {
6.     // Task 1: generate frequent itemsets that satisfy minimum support criterion
7.     for k = 1 to NumItems {
8.        Find frequent k-itemset, Sk, that satisfies the condition:
9.        LowerMinSupport ≤ support(Sk) ≤ UpperMinSupport
10.    }
11.    // Task 2: generate rules that satisfy minimum confidence (or lift) criterion
12.    for each frequent itemset S {
13.       for each subset SS of S {
14.          Rule R = "SS → (S – SS)"
15.          Compute confidence(R) and lift(R)
16.          if (Criterion == "lift") then score = lift(R)
17.          else score = confidence(R)
18.          if (score ≥ MinScore) then {
19.             Add R to set of association rules
20.             N = N+1
21.          }
22.       }
23.    }
24.    UpperMinSupport = UpperMinSupport – Delta
25. } until (UpperMinSupport ≤ LowerMinSupport) or (N == NumRules)
26. Sort set of association rules by Criterion
```

During the itemset generation, if an item is treated as an asymmetric attribute, it will be counted only when its value is not zero. This approach disallows negative association rules, or rules consisting of absent items such as *pasta = 0 → noodle = 1*. By ignoring these rules, one could miss valuable information about conflict or competition between items (Antonie & Zaiane, 2004; Yuan et al., 2002). In contrast, if an item is treated as a categorical attribute, each of its distinct values will be counted as a category. This allows negative association rules to be generated, but plenty of them could be redundant ones.

2.1 Post-mining of discovered rules

Learning what is conveyed in association rules usually begins with getting an overview of the findings. Toivonen et al. (1995) searched for subsets of rules that cover or summarize the whole set. Among those having the same consequences, the most general rules, or the rules sharing common antecedents with the majority, are selected. For example, given four rules *{a} → {z}*, *{b} → {z}*, *{a, b} → {z}*, and *{a, b, c} → {z}* :

- *{a} → {z}* covers itself, the third and the fourth rules.
- *{b} → {z}* covers itself, the third and the fourth rules.
- *{a, b} → {z}* covers itself and the fourth rule.
- *{a, b, c} → {z}* covers itself only.

Hence, the first and the second rules are selected as summary of this group. They are also called direction setting (DS) rules because, guided by them, one could focus on further detail by going through their related non-DS rules (Liu et al., 1999a). To ensure that a DS rule offers useful information about items' relationship, it is picked only if the chi-square correlation between its antecedents and consequences is positive.

Another rule summarization method constructs meta rules which express relationship between the discovered association rules (Berrado & Runger, 2007).

Seeing an overall picture about the domain, one can proceed to in-depth analysis by looking at specific or non-DS rules. However, going through a huge amount of them would be too onerous. Statistical pruning employs rule's common measures and their derivations. A basic idea is that if adding items to a general rule, to make it more specific, does not improve the rule's measures, then the supposedly specific rule is too trivial and is consequently pruned out (Webb & Zhang, 2002). Recall that association rules were generated on the grounds that their support and confidence had passed the criteria set by users. One may argue that they should not be pruned only because their measures are too low. Moreover, if the goal is to conduct detailed analysis, any extra information might be worth retaining.

Taking semantics or information carried by the rules into account, Klemettinen et al. (1994) and Ma et al. (2003) allowed users to construct a set of templates for rule selection. They used two types of templates, inclusive and exclusive ones. A rule is selected if it fits at least one inclusive template and does not fit any of the exclusive templates.

Silberschatz and Tuzhilin (1996) suggested that a rule is interesting if (1) it is unexpected or surprises the users; or (2) it is actionable or allows the users to use it to their advantage. The former criterion requires semantic perception. A user's belief system is first defined. Rules are compared against this system. The ones that affect or do not follow the user's belief will be considered unexpected and thus interesting. General impression can be used instead of the complex belief (Liu et al., 1997). The general impression is defined loosely as the users may have only vague feelings about the domain. Three types of rules are identified against the general impression: conforming, unexpected conclusion, and unexpected condition.

Li and Sweeney (2005) constructed robust rules from sets of rules that carry the same pieces of information. In each set, the most general expression is assigned as the robust rule's antecedent, and the most specific expression as the robust rule's consequence. A rationale is that the general expression states a broad hypothesis whereas the specific expression gives exact description about the knowledge. This technique requires concept hierarchy and the rule selection is performed during, not after, the rule generation.

3. Semantic classification and pattern analysis

Given positive and negative association rules, one approach to pruning or selecting them is based on how their meanings can be formulated. A general idea is that rules are useful if their meanings offer insight about the domain. Naturally, a domain is composed of several perspectives. If a rule reveals only one of them and contains all negative terms (nonexistent items), then it is not useful from a decision making point of view. An example is a rule describing vehicles involved in a traffic accident $\{bicycle = 0, sedan = 0\} \rightarrow \{truck = 0\}$. On the other hand, a rule that reveals more than one perspectives and contains only a few negative terms offers useful, albeit incomplete, insight about those perspectives.

Suppose that a data set corresponds to a domain. Variables are grouped into *subjects* which correspond to the domain's perspectives. For example, in a traffic accident data set, binary

variables are grouped into three subjects: vehicles involved, causes of accidents, and human losses. The details are as follows.

1. Vehicles involved. Binary variables and their item representation are
   - V0 represents bicycle or tricycle
   - V1 motorcycle
   - V2 sedan
   - V3 van or bus
   - V4 pick-up
   - V5 truck or trailer
   - V6 pedestrian
2. Causes of accidents. Binary variables and their item representation are
   - C0 represents vehicle overloaded or malfunctioned
   - C1 speeding
   - C2 violating traffic signs
   - C3 illegal blocking
   - C4 illegal overtaking
   - C5 swerving in close distance
   - C6 driving in the wrong lane or direction
   - C7 failing to signal
   - C8 careless driving
   - C9 following in close distance
3. Human losses. Binary variables and their item representation are
   - H1 represents dead
   - H2 seriously injured
   - H3 slightly injured

3.1 Formal definitions

Let $\{S_1, S_2, \ldots, S_n\}$ be subjects; $\{V_{a1}, V_{a2}, \ldots V_{ap}\}$ be antecedent variables in a rule; and $\{V_{c1}, V_{c2}, \ldots, V_{cq}\}$ be consequence variables in a rule. Let *nil* refer to "absent" or "unknown" category, or categories outside the scope of interest. This definition enables the semantic analysis to cover not only binary variables but also categorical ones. As a refinement on previous work (Marukatat, 2007), criteria to determine whether a rule is semantically useful are as follows.

1. A rule is classified as strongly meaningless if it has the following form:

$$\{V_{ai} = nil \mid V_{ai} \in S, i = 1 \text{ to } p\} \rightarrow \{V_{ck} = nil \mid V_{ck} \in S, k = 1 \text{ to } q\} . \quad (1)$$

   That is, all the variables have absent or unknown values, and they are members of the same subject. For instance, *{V0 = nil, V2 = nil}* → *{V5 = nil}* implies that an accident *not* involving bicycle and sedan tends to *not* involve truck. In other words, these vehicles are all absent from the accident. Since there are many types of vehicles in the domain, it is impossible to infer which and how the remaining ones would fit into this accident. This type of rules does not make an individual aspect of the domain (vehicles involved, in this case) any clearer and, therefore, can be removed from the analysis.
2. A rule is classified as weakly meaningless if it has the following form:

$$\{V_{ai} = nil \mid V_{ai} \in S_t, i = 1 \text{ to } p\} \rightarrow \{V_{ck} = nil \mid V_{ck} \in S_t, k = 1 \text{ to } q\} \text{ where } t = 1 \text{ to } n . \quad (2)$$

That is, all the variables have absent or unknown values, and they are members of more that one subjects. For instance, *{V6 = nil, H1 = nil}* → *{C1 = nil}* implies that an accident *not* involving pedestrian and *not* resulting in human death tends to *no*t being caused by speeding. Although this rule does not offer insight about individual subjects (vehicles involved, human losses, and causes of accidents, in this case), it reveals some vague interaction between them. Nevertheless, it may not be worth squeezing information out of these rules if there are plenty other rules available.

3. A rule is classified as partially meaningful if it has the following form:

$$\{V_{ah} \neq nil,\ V_{ai} = nil \mid h,\ i = 1 \text{ to } p;\ h \neq i\} \rightarrow \{V_{cj} \neq nil,\ V_{ck} = nil \mid j,\ k = 1 \text{ to } q;\ j \neq k\}\ . \qquad (3)$$

Some variables have absent or unknown values. For instance, *{V3 = 1, C1 = nil}* → *{V1 = 1}* implies that an accident involving bus and *not* being caused by speeding tends to involve motorcycles. These rules are complementary to meaningful ones as they help understand negative association between variables.

4. A rule is classified as meaningful if it does not fall into any of the above categories.

3.2 Further rule selection

Among rules classified as meaningful and partially meaningful, there are still redundant or repetitive patterns. This is because Apriori generated rules by permuting items in frequent itemsets and choosing ones that passed the user's criteria without pruning. In addition, the algorithm may have been executed multiple times, with multiple sets of parameters, leading to even more redundancy.

Let S_1 and S_2 be sets of items in rules R_1 and R_2, respectively. When the rules are compared, their relationship is defined as follows.

1. If S_1 equals S_2, then R_1 is equivalent to R_2.
2. If S_1 includes all items in S_2 and at least one item in S_1 does not exist in S_2 ($S_2 \subset S_1$ and $|S_1| > |S_2|$), then R_1 covers R_2. In other words, R_2 is covered by R_1.

The comparison takes every item into account irrespective of whether it is antecedent or consequence. The effect of it being one or the other is captured by the rule's measures, as illustrated by the following example. Given rules R_1 and R_2:

1. R_1: *{V2 = 1, C1 = nil}* → *{V1 = 1}* ;
   R_2: *{V2 = 1}* → *{C1 = nil, V1 = 1}* ;
   R_1 and R_2 are equivalent since $S_1 = S_2 =$ *{V2=1, C1=nil, V1=1}* .
2. Confidence(R_1) = P(*V2=1* ∩ *C1=nil* ∩ *V1=1*) / P(*V2=1* ∩ *C1=nil*) .
3. Confidence(R_2) = P(*V2=1* ∩ *C1=nil* ∩ *V1=1*) / P(*V2=1*) .
4. Lift(R_1) = P(*V2=1* ∩ *C1=nil* ∩ *V1=1*) / (P(*V2=1* ∩ *C1=nil*) × P(*V1=1*)) .
5. Lift(R_2) = P(*V2=1* ∩ *C1=nil* ∩ *V1=1*) / (P(*V2=1*) × P(*C1=nil* ∩ *V1=1*)) .

From a set of equivalent rules, the most significant one is selected. The rules' significance are determined according to user's criteria such as lift and confidence. Rules covering the others are selected while the ones being covered are pruned out. The definitions of "cover" and "being covered" are in the opposite direction of those mentioned in Section 2.1. There, the aim was to summarize the entire domain by using general rules. Here, the aim is to dig out as much information as possible from specific rules.

4. Case study: An analysis of traffic accident

This section demonstrates an application of semantic analysis and pattern analysis to real practice. Nakorn Pathom is a province located near Bangkok, the capital of Thailand. Over the past years, economic and human losses due to traffic accidents in Nakorn Pathom have been ranked among the highest of the country. Traffic accident data, dated from January 1st, 2003 to March 31st, 2006, were collected from its local police stations. The data set contains more records and more variables than the one used in previous publication (Marukatat, 2007). There are 1103 records, 20 binary variables, and 8 categorical variables in total. Subjects of binary variables were described in Section 3. Categorical variables were also grouped into the following subjects.

1. District. This subject includes only one categorical variable.
   D1 represents Nakorn Pathom's district area
2. Time. This subject includes three categorical variables.
   T1 represents quarter of day
   T2 day of week
   T3 quarter of year
3. Scenes of accidents. This subject includes four categorical variables.
   S1 represents type of road (highway, local road, etc.)
   S2 road feature (straight, intersection, etc.)
   S3 road material (concrete, laterite, etc.)
   S4 traffic direction (one-way, two-way)

Weka's Apriori (University of Waikato, n.d.) was employed to extract association rules from this data set. The algorithm was described in Section 2. The data set can be transformed to enable or disable the generation of negative association rules. In case that only positive rules are allowed, the *nil* value must be replaced by "?" which represents Weka's missing value. If negative rules are also allowed, *nil* is a user-defined value that represents a *nil* category. This case study used the latter setting. The other parameters were set as follows:

- NumRules = 500
- LowerMinSupport = 0.1
- UpperMinSupport = 0.4, 0.5, …, 0.9
- Delta = 0.05
- Criterion = lift
- MinScore = 1.5, 2, 3, 4

The algorithm was executed 24 times by using different combinations of UpperMinSupport and MinScore. Only 12 combinations produced the results. There are 4368 rules in total, as summarized in Table 2.

The rules' contents varied from 2 to 9 items. There was slim chance that every item in the rule was *nil*. Consequently, only 2.2% of the discovered rules were classified as meaningless (see Table 3). About 32% were classified as meaningful, and 65.8% as partially meaningful. Next, pattern analysis was performed to remove redundant patterns and find the most specific ones. As a result, 5.3% and 5.4% of the meaningful and partially meaningful rules were retained, respectively. Their confidence and lift measures are displayed in Fig. 1 and Fig. 2.

Parameters		Results		
MinScore (Lift)	UpperMinSupport	Max Lift	Max Confidence	Number of Rules
4.0	0.8	4.43	0.77	24
3.0	0.8	4.43	0.95	210
2.0	0.8	2.99	0.91	500
2.0	0.7	4.37	0.91	500
2.0	0.6	4.37	0.95	500
2.0	0.5	4.37	0.95	500
2.0	0.4	3.42	0.95	38
1.5	0.8	1.65	0.90	500
1.5	0.7	2.15	0.97	500
1.5	0.6	2.32	0.91	500
1.5	0.5	3.73	0.94	500
1.5	0.4	3.42	0.95	96
Total				4368

Table 2. Summary of rules discovered by Apriori

Semantic Classification	Before		After	
	No. of Rules (1)	% of Total	No. of Rules	% of (1)
Meaningful	1398	32	74	5.3
Partially meaningful	2874	65.8	155	5.4
Meaningless	96	2.2	-	-
Total	4368	100	229	5.2

Table 3. Semantic classification, before and after pattern analysis

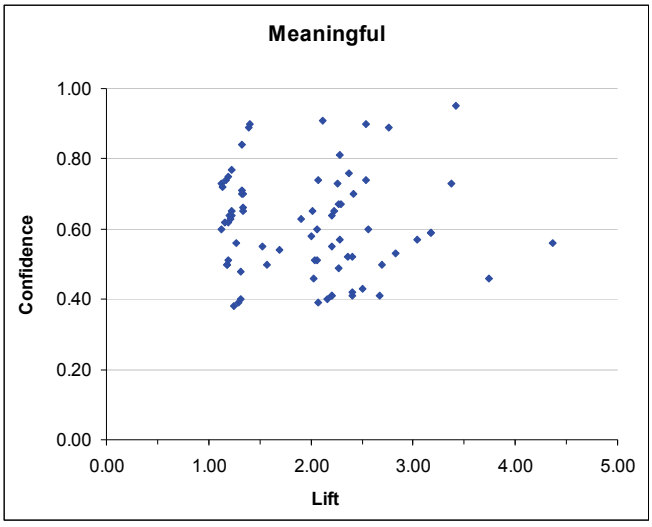

Fig. 1. Distribution of meaningful rules (after pattern analysis)

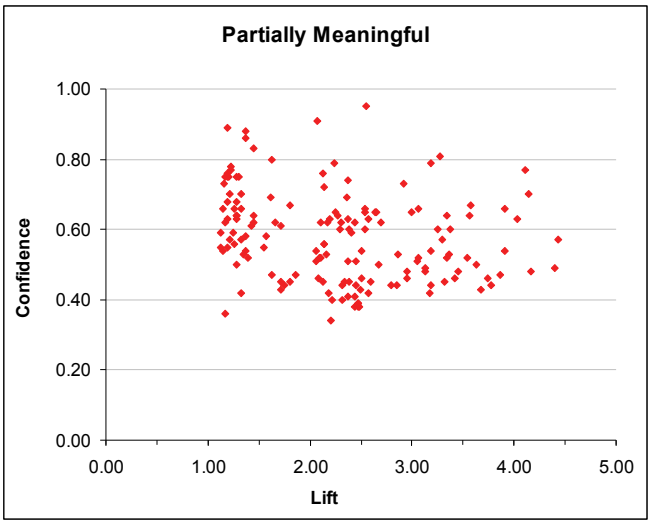

Fig. 2. Distribution of partially meaningful rules (after pattern analysis)

The following are some of the retained rules. To make them more understandable, items were substituted by categories' description, e.g. *S1 = 1* was substituted by *highway*.

Meaningful	
1.	12.01-18.00, local road, intersection → illegal blocking
2.	12.01-18.00, straight, dead → truck
3.	00.01-06.00, truck → dead
4.	00.01-06.00, highway, straight → speeding
5.	18.01-24.00, pedestrian → speeding, dead
6.	Swerving in close distance → bicycle
7.	Bus, swerving in close distance → intersection
8.	Local road, illegal overtaking → curve
9.	Violating traffic signs → pedestrian
10.	Violating traffic signs → 06.01-12.00, community area
Partially meaningful	
11.	Truck → highway, *no* sedan, *not* speeding, *not* swerving in close distance
12.	Local road, slightly injured → two ways, asphalt surface, *no* sedan, *not* speeding
13.	Bicycle, truck → straight, *no* pick-up, *no* slightly injured
14.	*No* motorcycle, dead → highway, speeding
15.	18.01-24.00, pedestrian → *no* motorcycle, dead

Nakorn Pathom is known as a gateway to the western and the southern regions of Thailand. Heavy vehicles usually travel through the province at night and in the early morning, rather than in the afternoon. Thus, rules 3 and 4 look like typical accident patterns in Nakorn Pathom while rule 2 is slightly unexpected.

Rule 1 implies that an accident occurring at the intersection of local roads, in the afternoon, is likely to be caused by illegal blocking. Rule 10 implies that an accident caused by violating traffic signs is likely to occur in community area, in the morning. An investigation into the

amount of traffic during rush hours, the adequacy of traffic lights around the intersections, and the law/regulation enforcement should be made to complete the picture.
Problems with law/regulation enforcement is confirmed by rule 9, but this rule describes a rather typical accident pattern in many parts of Thailand. That is, pedestrians often cross the roads wherever they like and vehicles seldom stop for them.
The analysis of partially meaningful rules offers more insight about traffic accidents, or at least spawns a few questions for further investigation. For example, rule 11 implies that an accident which involves truck and highway does not involve sedan, and is not caused by speeding or swerving in close distance. One might suspect if there is any type of vehicles other than truck potentially fitting the pattern described by rule 4.
Rule 15 can serve as a complement to rule 5. From rule 5, one learns an accident pattern that happens in the evening until midnight and involves pedestrian. It is likely to be caused by speeding and result in human death. From rule 15, one also learns that this accident is likely to not involve motorcycle.

5. Discussion

5.1 Semantic analysis

Association rules can be classified by their meanings or semantics. A rule is meaningful if its whole content describes something which exists in the domain. In contrast, it is meaningless if its whole content describes something which does not exist. A partially meaningful rule is somewhere in the middle. It is comparable to a negative association rule in "A → ~B" or "~A → B" forms. As mentioned in Section 2, this type of association is useful in marketing research. It helps identify conflicting items that should not be promoted together, or a replacement item in case that the other is in short supply. Wu et al. (2004) gave another example. Suppose that A normally triggers an alert of event B. Rule "A → ~B" suggests that the alert can be postponed because B has not yet happened. This chapter has also demonstrated how such rules were used to gain a better understanding about the domain. However, not all of them are useful for the analysis. Consider the following:

1. *{V3 = 1, C5 = 1} → {S2 = 2, V1 = 1, V2 = nil, V4 = nil, V5 = nil, V6 = 1}*
2. *{V3 = 1, C5 = 1} → {S2 = 2, C1 = nil, C3 = nil, C4 = nil, C8 = nil}*

Both rules say that accidents involving bus and being caused by swerving in close distance is likely to occur at the intersection. They give further detail about vehicles involved and causes of accidents, respectively. The extra detail regarding vehicles involved is interesting because it is normal that an accident involves more than one vehicles. On the other hand, it is unlikely (but sometimes possible) that an accident is caused by so many reasons. Hence, adding that there is no other cause of accident is unnecessary.
The above example shows different natures of the subjects. In some subjects, multiple or all the items can exist at the same time. But in the others, one or only a couple of items can co-exist. Further refinement should be made to the semantic analysis to handle this.
A few works have mentioned negative association rules in "~A → ~B" form (Antonie & Zaiane, 2004; Wu et al., 2004; Yuan et al., 2002). None explained how to exploit such rules. Only Yuan et al. (2002) suggested that "~A → ~B" is equivalent to "B → A", but did not elaborate any further. To make sense of this, "~A → ~B" is interpreted as that the absence of A causes the absence of B. Therefore, the presence of B would imply the presence of A. It is probable if the association (→) is perceived as cause-and-effect relationship.

The cause-and-effect perception is weak when every item belongs to the same subject, as in a strongly meaningless rule *{V0 = nil, V2 = nil} → {V5 = nil}*. It is more natural when items belong to different subjects, as in a weakly meaningless rule *{V6 = nil, H1 = nil} → {C1 = nil}*. However, it is imprudent to infer that any form of the inverse, e.g. *{V6 = 1, H1 = 1} → {C1 = 1}* or *{C1 = 1} → {V6 = 1, H1 = 1}*, is true.
Although weakly meaningless rules give a little insight about the domain, it is still unclear how to exploit them effectively. At the moment, they serve only as a confirmation of what has been learned from meaningful rules.

5.2 Pattern analysis

The pattern analysis helps remove repetitive or redundant patterns. It aims to retain rules which describe the most information. These rules normally have much lower support and confidence than general ones. In spite of this, the measures are supposed to be accepted by the users, according to their own (Apriori's) criteria. Otherwise, the support and confidence thresholds could have been raised to avoid generating these rules.
Based on the analysis, rules *{V2 = 1} → {C1 = nil, V1 = 1}* and *{V2 = 1, C1 = nil} → {V1 = 1}* are equivalent. The one with the higher confidence or lift will be selected. In some practices, both of them are considered important. The likes of *{V2 = 1} → {C1 = nil, V1 = 1}* are characteristic rules while the likes of *{V2 = 1, C1 = nil} → {V1 = 1}* are discriminant rules (Brijs et al., 2000; Cheung et al., 2000). The former are useful for description purpose, since they characterize single antecedent items (concepts) with multiple consequence items. The latter are useful for prediction purpose, since they discriminate consequence items (classes) by using multiple antecedent items. However, this chapter sees both of them as conveying the same piece of information, only in slightly different forms. Therefore, keeping any one of them would be sufficient.
This strategy has some drawbacks. First, consider the following cases:
1. R_0: *{V2 = 1} → {S1 = 1, H2 = 1}* and R_1: *{V2 = 1, C1 = nil} → {V1 = 1}* are selected.
2. R_0: *{V2 = 1} → {S1 = 1, H2 = 1}* and R_2: *{V2 = 1} → {C1 = nil, V1 = 1}* are selected.

The second case would make analysis job easier because the rules can be grouped easily and insight about the accidents involving sedan can be obtained quickly. But the first case may happen if R_1 has higher confidence or lift than R_2. The fact that specific rules usually have many more items than general ones makes the analysis even harder.

6. Conclusion

Association rule mining produces a large amount of rules. Many of them are redundant ones. This chapter has presented techniques to select rules that are semantically useful and carry the most information. They aim for a complete understanding, rather than an overall picture, about the domain. Prior to the analysis, variables are grouped into subjects which correspond to the domain's perspectives. These subjects are key factors to classify the rules into strongly meaningless, weakly meaningless, partially meaningful, and meaningful ones. Rules that have equivalent patterns are identified and the most significant one is selected. Furthermore, between a general and a more specific rule, the latter is selected since it offers more insight about the domain.
These rule selection strategies still have some drawbacks, as discussed in Section 5. Further refinement on the semantic analysis would help filter out even more semantically redundant

rules. Throughout this chapter, it was assumed that rules which offer as much information as possible are the ones users would like to see. But as shown in Section 4, some rules only describe what the users already knew. There are techniques, as mentioned in Section 2, that take the users' existing knowledge and the unexpectedness of the rules into account (Liu et al., 1997; Silberschatz & Tuzhilin, 1996). Their ideas could be incorporated to improve the rule selection capability. Finally, visualization systems (Blanchard et al., 2003; Bruzzese & Davino, 2005; Techapichetvanich & Datta, 2005) would make the analysis job easier.

8. References

Ableson, A. & Glasgow, J. (2003). Efficient statistical pruning of association rules, In: PKDD 2003, *Lecture Notes in Computer Science*, Vol. 2838, Lavrac, N. et al. (Eds.), pp. 23-34, Springer.

Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules, *Proceedings of the 20th International Conference on Very Large Databases (VLDB)*, pp. 487-499, Santiago, Chile, September 1994, Morgan Kaufmann.

Antonie, M.-L. & Zaiane, O. R. (2004). Mining positive and negative association rules: An approach for confined rules, In: PKDD 2004, *Lecture Notes in Computer Science*, Vol. 3202, Boulicaut, J. F. et al. (Eds.), pp. 27-38, Springer.

Berrado, A. & Runger, G. C. (2007). Using meta rules to organize and group discovered association rules, *Data Mining and Knowledge Discovery*, Vol. 14, No. 3, pp. 409-431, Springer.

Blanchard, J., Guillet, F. & Briand, H. (2003). Exploratory visualization for association rule rummaging, *The 4th International Workshop on Multimedia Data Mining (MDM/KDD)*, Washington, DC, August, 2003.

Brijs, T., Vanhoof, K. & Wets, G. (2000). Reducing redundancy in characteristic rule discovery by using integer programming techniques, *Intelligent Data Analysis*, Vol. 4, No. 3-4, pp. 229-240, IOS Press.

Brin, S., Motwani, R., Ullman, J. D. & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 255-264, Tucson, AZ, May 1997, ACM Press.

Bruzzese, D. & Davino, C. (2003). Visual post-analysis of association rules, *Journal of Visual Languages and Computing*, Vol. 14, No. 6, pp. 621-635, Elsevier.

Cheung, D. W., Hwang, H. Y., Fu, A. W. & Han, J. (2000). Efficient rule-based attribute-oriented induction for data mining, *Journal of Intelligent Information Systems*, Vol. 15, No. 2, pp. 175-200, Kluwer Academic Publishers.

Han, J., Pei, J. & Yin, Y. (2000). Mining frequent patterns without candidate generation, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 1-12, Dallas, TX, May 2000.

Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H. & Verkamo A. I. (1994). Finding interesting rules from large sets of discovered association rules, *Proceedings of the 3rd International Conference on Information and Knowledge Management (ICKM)*, pp. 401-407, Gaithersburg, MD, November 1994.

Ko, Y. S. & Rountree, N. (2005). Finding sporadic rules using Apriori-Inverse, In: PAKDD 2005, *Lecture Notes in Computer Science*, Vol. 3518, Ho, T. B. et al. (Eds.), pp. 97-106, Springer.

Li, Y. & Sweeney, L. (2005). *Adding semantics and rigor to association rule learning: the GenTree approach*, Technical Report CMU ISRI 05-101, School of Computer Science, Carnegie Mellon University.

Liu, B., Hsu, W. & Chen, S. (1997). Using general impressions to analyze discovered classification rules, *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 31-36, Newport Beach, CA, August 1997, AAAI Press.

Liu, B., Hsu, W. & Ma, Y. (1999a). Pruning and summarizing the discovered association, *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 125-134, San Diego, CA, August 1999.

Liu, B., Hsu, W. & Ma, Y. (1999b). Mining association rules with multiple minimum supports, *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 337-341, San Diego, CA, August 1999.

Ma, L., Tsui, F.-C., Hogan, W. R., Wagner, M. M. & Ma, H. (2003). A framework for infection control surveillance using association rules, *AMIA Annual Symposium Proceedings*, pp. 410-414, American Medical Informatics Association.

Marukatat, R. (2007). Structure-based rule selection framework for association rule mining of traffic accident data, In: CIS 2006, *Lecture Notes in Computer Science*, Vol. 4456, Wang, Y. et al. (Eds.), pp. 231-239, Springer.

Silberschatz, A. & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 6, pp. 970-974.

Svetina, M. & Zupancic, Joze. (2005). How to increase sales in retail with market basket analysis. *Systems Integration,* pp. 418-428.

Techapichetvanich, K. & Datta, A. (2005). VisAr: A new technique for visualizing mined association rules, In: ADMA 2005, *Lecture Notes in Computer Science*, Vol. 3584, Li, X. et al. (Eds.), pp. 88-95, Springer.

Toivonen, H., Klemettinen, M., Ronkainen, P., Hatonen, K. & Mannila, H. (1995). Pruning and grouping of discovered association rules, *Workshop Notes of the ECML 95 Workshop in Statistics, Machine Learning, and Knowledge Discovery in Databases*, pp. 47-52, Greece, 1995.

University of Waikato (n.d.). Weka 3 – data mining software in Java (version 3.4.12) [software]. Available from http://www.cs.waikato.ac.nz/ml/weka/.

Webb, G. I. & Zhang, S. (2002). Removing trivial association in association rule discovery, *Proceedings of the 1st International NAISO Congress on Autonomous Intelligent Systems (ICAIS)*, Geelong, Australia, 2002, NAISO Academic Press.

Wu, X., Zhang, C. & Zhang, S. (2004). Efficient mining of both positive and negative association rules, *ACM Transactions on Information Systems*, Vol. 22, No. 3, pp. 381-405.

Yuan, X., Buckles, B. P., Yuan, Z. & Zhang, J. (2002). Mining negative association rules, *Proceedings of the 7th IEEE International Symposium on Computers and Communications (ISCC)*, pp. 623-628, Taormina, Italy, July 2002, IEEE Computer Society.

Yun, H., Ha, D., Hwang, B. & Ryu, K. H. (2003). Mining association rules on significant rare data using relative support, *Journal of Systems and Software*, Vol. 67, No. 3, pp. 181-191, Elsevier.

Zaki, M. J., Parthasarathy, S., Ogihara, M. & Li, W. (1997). New algorithms for fast discovery of association rules, *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 283-286, Newport Beach, CA, August 1997, AAAI Press.

Data Mining and Knowledge Discovery in Real Life Applications

Edited by Julio Ponce and Adem Karahoca

ISBN 978-3-902613-53-0
Hard cover, 436 pages
Publisher I-Tech Education and Publishing
Published online 01, January, 2009
Published in print edition January, 2009

This book presents four different ways of theoretical and practical advances and applications of data mining in different promising areas like Industrialist, Biological, and Social. Twenty six chapters cover different special topics with proposed novel ideas. Each chapter gives an overview of the subjects and some of the chapters have cases with offered data mining solutions. We hope that this book will be a useful aid in showing a right way for the students, researchers and practitioners in their studies.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Rangsipan Marukatat (2009). On the Selection of Meaningful Association Rules, Data Mining and Knowledge Discovery in Real Life Applications, Julio Ponce and Adem Karahoca (Ed.), ISBN: 978-3-902613-53-0, InTech, Available from: http://www.intechopen.com/books/data_mining_and_knowledge_discovery_in_real_life_applications/on_the_selection_of_meaningful_association_rules

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821