User Interactive PostProcessing of Association Rules and Correlation based Redundancy Removal

C. Sweetlin
PG Student
Department of Computer Science &Engineering (PG)
National Engineering College

V. Kalaivani
Associate Professor
Department of Computer Science &Engineering (PG)
National Engineering College

ABSTRACT

Traditional association rule mining generates a large number of rules. This leads to a difficulty in finding the interested and significant rules. An efficient interactive post-processing task which includes ontology and rule schema is used to obtain user interesting rules. Correlation analysis finds significant association rules by analyzing the dependency between the antecedent and consequent parts of the rule. In this paper, correlation analysis is integrated with the interactive post-processing to obtain significant user interesting rules. A redundancy removal follows this framework to weed out the extra rules and also to reduce the ruleset further. The proposed methodology provides a significant set of non-redundant user interesting rules leading to an efficient analysis.

Keywords

Postprocessing, User Knowledge, Ontology, Rule schema, Correlation, Redundant rules.

1. INTRODUCTION

Association rule mining finds frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories. Using association rule learning, a decision maker can determine what products are frequently bought together and use this information for marketing purposes. This is referred to as market basket analysis. In this, rules indicate customer buying patterns.

An association rule is defined as the implication X → Y, described by two interestingness measures: support and confidence, where X and Y are the sets of items and $X \cap Y = \Phi$. Apriori is the first algorithm in the association rule mining field. Starting from a database, it proposes to extract all association rules satisfying minimum thresholds of support and confidence. It is very well known that mining algorithms can discover a prohibitive amount of association rules; for instance, thousands of rules are extracted from a database of several dozens of attributes and several hundreds of transactions. To obtain user interesting rules, post-processing methods are used. Effective use of post-processing methods yield exact user interesting rules. However, most of the existing post-processing methods are generally based on statistical information in the database. Since rule interestingness strongly depends on user knowledge and goals, these methods do not guarantee that interesting rules will be extracted. A pattern is interesting, if it is easily understood by humans, valid or potentially useful.

So the rule post-processing methods should be imperatively based on a strong interactivity with the user integrating user interestingness. The representation of user knowledge is an important issue. The more the knowledge represented in a flexible, expressive, and accurate formalism, the more the rule selection is efficient. Ontology is considered as the most appropriate representation to express the complexity of the user knowledge. To represent user expectations in terms of discovered rules, three levels of specification: General impressions, Reasonably Precise Concepts—representing user vague feelings, and finally, his/her Precise Knowledge can be used. But all the interesting rules are not significant. So it is essential to help the decision maker to find the truly interesting rules. Correlation analysis is useful in finding the significant rules. It analyses the dependencies between items present in the rule, whether the relations occur by chance or true dependencies exist between them. Significant user interesting rules becomes more effective if there is no redundancy in the ruleset, since redundant rules do not provide any useful information. If all the rules in the ruleset are positively correlated, then redundancies can easily be eliminated.

In this paper, the interactive post-mining framework is followed by correlation analysis which in turn is followed by redundancy removal. The proposed methodology provides significant user interesting rules without redundancy leading to a reduced ruleset.

2. RELATED WORK

In Data Mining, the usefulness of association rules is strongly limited by the huge amount of delivered rules, insignificance and the presence of redundant rules. To overcome this drawback, several methods were proposed in the literature. Interestingness measures were proposed in order to discover only those association rules that are interesting according to these measures. They have been divided into objective and subjective measures. Objective measures depend only on data structure. Many survey papers summarize and compare the objective measure definitions and properties [1],[2]. Unfortunately, being restricted to data evaluation, the objective measures are not sufficient to reduce the number of extracted rules and to capture the interesting ones. Several approaches integrating user knowledge have been proposed. In addition, subjective measures were proposed to integrate explicitly the decision-maker knowledge and to offer a better selection of interesting association rules. Silberschatz and Tuzilin [3] proposed a classification of subjective measures in unexpectedness—a pattern is interesting if it is surprising to the user—and actionability—a pattern is interesting if it can help the user take some actions.

Klemettinen et al. [4] proposed templates to describe the form of interesting rules (inclusive templates) and not interesting rules (restrictive templates). The idea of using templates f or association rule extraction was reused in [5]. Other approaches proposed to use a rule-like formalism to express user expectations [3], [6], [7], and the discovered association rules are pruned/summarized by comparing them to user expectations.

Another related approach was proposed by An et al. in [8] where the authors introduced domain knowledge in order to prune and summarize discovered rules. The first algorithm uses data taxonomy, defined by user, in order to describe the semantic distance between rules, and in order to group the rules. The second algorithm allows grouping the discovered rules that share at least one item in the antecedent and the consequent.

Ontologies, introduced in data mining for the first time in early 2000, can be used in several ways [9]: Domain and Background Knowledge Ontologies, Ontologies for Data Mining Process, or Metadata Ontologies. Background Knowledge Ontologies organize domain knowledge and play important roles at several levels of the knowledge discovery process. Ontologies for Data Mining Process codify mining process description and choose the most appropriate task according to the given problem; while Metadata Ontologies describe the construction process of items. The first idea of using Domain Ontologies was introduced by Srikant and Agrawal with the concept of Generalized Association Rules (GAR) [10]. The authors proposed taxonomies of mined data (an is-a hierarchy) in order to generalize/specify rules.

To identify truly interesting rules, statistical correlation is used in [22] as the basis for finding rules that represent the fundamental relations of the domain. The technique here, prunes the discovered associations to remove the insignificant associations which are not useful.

[20] Proposed a rule pruning technique using *minimum improvement*, which is the difference between the confidence of a rule and the confidence of any proper subrule with the same consequent. Those rules that do not meet this minimum improvement in confidence are pruned. Toivonen et al. proposed in [11] a novel technique for redundancy reduction based on rule covers. The notion of rule cover is defined as the subset of a rule set describing the same database transaction set as the rule set. Aggarwal et al. [12] classify the redundant rule in two groups, such as: *simple redundant* and *strict redundant*. . They proposed that a rule bears *simple redundancy* in the presence of other rules if and only if those rules are generated from same frequent itemset and the support values for the those rules are the same but the confidence value for one of them is higher than the others. The authors considered rules as *strict redundancies* that are generated from two different frequent itemsets but one is the subset of another. In this paper, the proposed methodology yields significant rules from correlation analysis. So redundancy can be directly eliminated from the significant user interesting rules, by verifying each rule with the set of rules.

3. USER INTERACTIVE POSTMINING

Post-mining of huge number of association rules generated by a traditional mining algorithm is done using ontologies and rule schemas. Using item taxonomies has many advantages: the representation of user expectations is more general, but the taxonomy of items might not be enough. The user might want to use concepts that are more expressive and accurate than generalized concepts and that result from relationships other than the "is-a relation". This is why it is considered that the use of ontologies would be more appropriate. The steps of user interactive postmining are discussed below.

3.1 Construction of ontology

Since early 2000s, in the Semantic Web context, the number of available ontologies has been increasing covering a wide domain of applications. This could be a great advantage in an ontology-based user knowledge representation. This paper contributes on several levels at reducing the number of association rules. One of our most important contributions relies on using ontologies as user background knowledge representation. Thus, we extend the specification languages like General Impressions (GI), Reasonably Precise Concepts (RPC), and Precise Knowledge (PK)—by the use of ontology concepts. Domain ontology is used in this paper. This is used to strengthen the integration of user knowledge in the post processing task. This involves defining three types of concepts. They are:

1) Leaf concepts

A leaf concept is defined such that, each leaf concept is associated to one item in the database.

2) Generalized concepts

Generalized concepts are defined such that the concepts subsume other concepts in the ontology. A generalized concept is connected to the database through its subsumed concepts.

3) Restriction concepts

Restriction concepts are described using logical expressions defined over items and depend on the user individually. Considering a supermarket database consisting of the items namely cookies, candy, olive oil, Ricebran oil etc, ontology can be constructed by the user as follows:

Leaf concepts: {cookies, candy, Ricebran oil, coconut oil….}
Generalized concepts: {Snacks, Oil, Food items…}
Restriction concepts: {Healthy, Fatty…}

Two data properties are also integrated in order to define whether a product is healthy or fatty. For example, the restriction concept "Healthy" is described using description logics language by:

Healthy ≡ Fooditems ∩ ∃ isHealthy.TRUE, this defines all food items that have the Boolean property isHealthy on TRUE. For our example, isHealthy is instantiated as follows:

isHealthy: {(cookies, TRUE), (Ricebran oil, TRUE)} Now, we are able to connect the ontology and the database, for example, the concept cookies is connected to the same item f(cookies) = cookies. On the contrary, the generalized concept Snacks is connected through its two subsumed concepts:

f (Snacks) ={cookies, candy}Similarly, we can describe the connection for other concepts. More interesting, the restriction concept "Healthy" will be connected through those concepts satisfying the restrictions in the definition of the concept. Thus, Healthy is connected through the concepts cookies and Ricebran oil:f (Healthy) = {cookies, Ricebran oil}

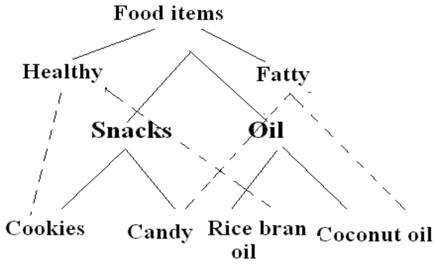

Figure: 1 Visualisation of ontology

3.2 Defining Rule schemas

To improve association rule selection, we propose a new rule filtering model, called Rule Schemas (RS). A rule schema describes, in a rule-like formalism, the user expectations in terms of interesting/obvious rules. As a result, Rule Schemas act as a rule grouping, defining rule families.

A Rule Schema expresses the fact that the user expects certain elements to be associated in the extracted association rules. It brings the expressiveness of ontologies in the post processing task of association rules combining not only item constraints, but also ontology concept constraints. This can be expressed as RS (<X1; . . .;Xn (→) Y1; . . . ; Ym>) For example: if the user is interested in buying pattern of Snacks and Oil, the rule schema can be defined by RS (Snacks → Oil)which involves the concepts of ontology. Based on the ontology database mapping, rules conforming to the rule schema are filtered as the user interesting rules.

4. CORRELATION ANALYSIS

The resulting interesting rules consist of insignificant or rules occurring by chance. Their existence may simply be due to chance rather than true correlation. In business analysis, a decision maker should find the items that can lift the sales of other items using the rules that are positively correlated. So, the decision maker is helped with a correlation analysis to find the important rules in his/her interesting ruleset. The true dependencies of items in the rule can be found by a correlation measure. The correlation measure used in this paper is "lift". Lift is calculated for each of the rules in the interesting ruleset. Lift measures how much more frequently the left-hand item of the rule is found with the right-hand than without the right-hand item.

Lift or correlation is given by

$$lift(A \rightarrow B) = \frac{Support(A \cup B)}{Support(A).Support(B)}$$

Correlation (A, B) >1means that A and B are positively correlated i.e. the occurrence of one implies the occurrence of the other and correlation (A, B) < 1means that the occurrence of A is negatively correlated with (or discourages) the occurrence of B and correlation (A, B) =1means that A and B are independent. Rules that are positively correlated are filtered as significant user interesting rules. Considering a super market database, a decision maker can find the items that can really lift the sales of other items using significant rules.

5. REDUNDANCY REMOVAL

Even if the rules obtained after the interactive post-processing and correlation analysis is user interesting and significant, there remains extra or redundant rules. So there is a need to remove redundancy to reduce the positively correlated interesting ruleset further. A rule *r* in R is said to be *redundant* if and only if a rule or a set of rules *S* where *S*∈*R* possess same intrinsic meaning of *r*. For example, consider a rule set R has three rules such as *milk→ tea*, *sugar→ tea*, and *milk, sugar→tea*. If we know the first two rules i.e. *milk→tea* and *sugar→tea*, then the third rule *milk*, *sugar→tea* becomes redundant, because it is a simple combination of the first two rules and as a result it does not covey any extra information especially when the first two rules are present and moreover, all the three rules are positively correlated. In many cases enormous redundant rules often fades away the intention of association rule mining.

One can classify association rules in two different types based on the number of items in the consequence: rules having single items in the consequence and rules having multiple items in the consequence. Depending on the application requirements, association rule mining algorithms produce ruleset, which may contains rules of both types. However, it is worth to mention that redundant rules exist in both types. Since the interesting ruleset consists of only positively correlated rules after correlation analysis, it is easy to eliminate redundant rules of these two types using the two methods: removing redundant rules with fixed *antecedent* rules, and with fixed *consequent* rules.

5.1. Finding Redundant Rules with Fixed Antecedent Rules

The algorithm for this finds those redundant rules that have multiple items in the consequence but have the same antecedent itemset in the antecedent, from the positively correlated ruleset. It first iterates through the whole rule set and finds those rule that have multiple itemset in the consequence. Once it comes across such a rule, checking is carried out to see whether *n* numbers of (*n-1*)-itemset of the consequence are in the rule set with the same antecedent. If it finds *n* number of rules in the rule set and since they are positively correlated, we delete that rule from the rule set otherwise that rule remains in the rule set. The pseudo code for this is given in Figure 2.

```
For all rules $r \in R$
  r = U {A, C}
  n = Length(C)
  if (n>1)
    for all (n-1) – subsets $e \in C$
        if( $r_i$ = U{A,e})
         e.i ++
    end for
     if (i ==n)
       R=R-r
  end if
End for
```

Figure 2: Pseudo code for *Finding Redundant Rules With Fixed Antecedent Rules.*

5.2. Finding Redundant Rules with Fixed Consequent Rules

The algorithm for this finds those redundant rules that have multiple items in the antecedent but have the same antecedent itemset in the consequence, from the positively correlated ruleset. First, it finds the antecedents that have multiple itemset. Once it comes across such rule a check is made to see whether *n* numbers of (n-1) itemset of the antecedent are in the rule set with the same consequence. If it finds *n* number of rules in the rule set and since they are positively correlated, we delete that rule from the rule set otherwise that rule remains in the rule set. The pseudo code for this is given in Figure 3.

The proposed methods for redundancy are not based on any bias assumptions. In addition it verifies each rule with set of rules in order to find redundant rule. Hence it eliminates redundant rules without losing any important knowledge from the resultant rule set.

```
For all rules r∈R
    r = U {A, C}
    n = Length (A)
    if (n>1)
     for all (n-1) – subsets e ∈A
          if( $r_i$ = U{C,e})
           e.i ++;
      end for
       if (i ==n)
          R=R-r
    end if
 End for
```

Figure 3: Pseudo code for *Finding Redundant Rules With Fixed Consequent Rules*

The integration of the three steps discussed above is shown in figure 4.

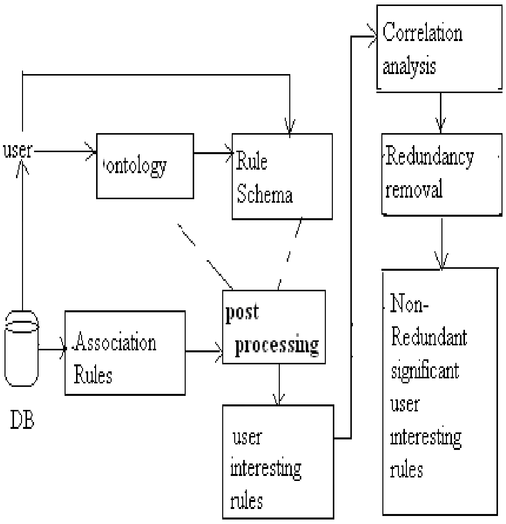

Figure: 4 User interactive postmining of association rules and correlation based redundancy removal (UPACBR) framework

6. RESULTS

The rule filtering framework discussed in this paper results in a reduced ruleset. This reduced ruleset consist of significant user interesting rules without any redundancy. A supermarket database of 4627 transactions which include 39 items is taken for experiment. The rules are generated using apriori algorithm with 60% of confidence threshold. The resulting ruleset consists of 2755 rules.

Using the concepts of ontology in a rule schema, user interesting rules are filtered. By applying various levels of filters discussed above, an analyzable set of 399 association rules are obtained and is shown in Figure 5.

Compared to taxonomies used in the specification language proposed in [6], ontologies offer a more complex knowledge representation model by extending the only is-a relation presented in taxonomy with the set R of relations.

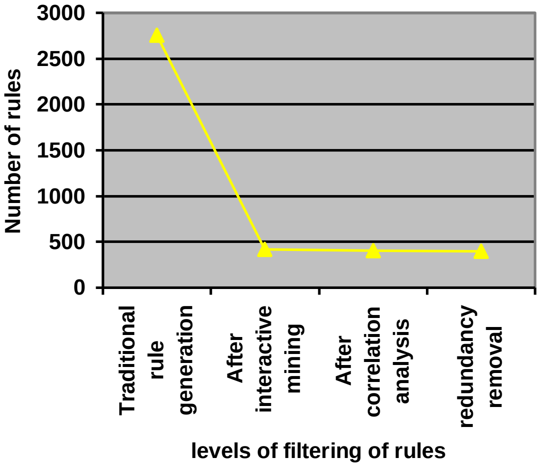

Figure: 5 Rules obtained in successive levels of filtering

The proposed methodology uses ontology to express the user interestingness more accurately. Accordingly, the rules are filtered accurately.

The methodology in [21] ARIPSO framework, the user is not helped with correlation analysis to filter only the important rules in his/her's interesting ruleset. The proposed methodology here (UPACBR), uses correlation analysis using lift to filter the significant rules. [21] Uses a redundancy removal technique using "minimum improvement", which is the difference between the confidence of a rule and the confidence of any proper subrule with the same consequent. Since confidence cannot determine a rule's true significancy, here correlation analysis is followed and redundancy can directly be removed from the positively correlated ruleset by finding rules with complete set of its subrules.

Table1. Comparison of levels of filtering in existing and proposed methodologies

	User interesting rules	Significant rules	Non redundant rules
ARIPSO	Interactive post-processing	-	Using minimum improvement filter(confidence based)
UPACBR	Interactive post-processing	Correlation analysis	Correlation based

Since the proposed methodology uses correlation analysis and redundancy removal successively after the interactive post-processing, a significant set of non redundant user interesting rules is obtained.

6. CONCLUSION

In association rule mining, large number of rules consisting of insignificant and redundant rules does not make any sense. Methods that include user interestingness yield user interesting rules. But they do not help the decision maker to find the important or significant rules. This paper discusses a new approach which integrates user interactive post-processing with correlation analysis, followed by a correlation based redundancy removal. Thus the proposed methodology yields a reduced, non-redundant significant user interesting ruleset.

7. REFERENCES

[1] F. Guillet and H. Hamilton, Quality Measures in Data Mining. Springer, 2007.

[2] P.-N. Tan, V. Kumar, and J. Srivastava, “Selecting the Right Objective Measure for Association Analysis,” Information Systems, vol. 29, pp. 293-313, 2004.

[3] Silberschatz and A. Tuzhilin, “What Makes Patterns Interesting in Knowledge Discovery Systems,” IEEE Trans. Knowledge and Data Eng. vol. 8, no. 6, pp. 970-974, Dec. 1996.

[4] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A.I. Verkamo, “Finding Interesting Rules from Large Sets of Discovered Association Rules,” Proc. Int’l Conf. Information and Knowledge Management (CIKM), pp. 401-407, 1994.

[5] Baralis and G. Psaila, “Designing Templates for Mining Association Rules,” J. Intelligent Information Systems, vol. 9, pp. 7-2, 1997.

[6] Liu, W. Hsu, K. Wang, and S. Chen, “Visually Aided Exploration of Interesting Association Rules,” Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), pp. 380-389, 1999.

[7] B. Padmanabhan and A. Tuzhuilin, “Unexpectedness as a Measure of Interestingness in Knowledge Discovery,” Proc. Workshop Information Technology and Systems (WITS), pp. 81-90, 1997

[8] An, S. Khan, and X. Huang, “Objective and Subjective Algorithms for Grouping Association Rules,” Proc. Third IEEE Int’l Conf. Data Mining (ICDM ’03), pp. 477-480, 2003.

[9] H. Nigro, S.G. Cisaro, and D. Xodo, Data Mining with Ontologies: Implementations, Findings and Frameworks. Idea Group, Inc., 2007.

[10] R.Srikant and R. Agrawal, “Mining Generalized Association Rules,” Proc. 21st Int’l Conf. Very Large Databases, pp. 407-419, 1995.

[11] H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hatonen, and H. Mannila, “Pruning and Grouping of Discovered Association Rules,” Proc. ECML-95 Workshop Knowledge Discovery in Databases, pp. 47-52, 1995.

[12] Charu C. Agrarian and Philip S. Yu, “A new Approach to Online Generation of Association Rules”. *IEEE TKDE*, Vol. 13, No. 4 pages 527- 540.

[13] Bing Liu, Wynne Hsu and Yiming Ma, “Pruning and Summarize the Discovered Associations”. *In the proc. of ACM SIGMOD* pp.125 134, San Diego, CA, August 1999.

[14] L.M. Garshol, “Metadata? Thesauri? Taxonomies? Topic Maps Making Sense of It All,” J. Information Science, vol. 30, no. 4, pp. 378-391, 2004.

[15] M.A. Domingues and S.A. Rezende, “Using Taxonomies to Facilitate the Analysis of the Association Rules,” Proc. Second Int’l Workshop Knowledge Discovery and Ontologies, held with ECML/ PKDD, pp. 59-66, 2005.

[16] Bellandi, B. Furletti, V. Grossi, and A. Romei, “Ontology- Driven Association Rule Extraction: A Case Study,” Proc. Workshop Context and Ontologies: Representation and Reasoning, pp. 1-10, 2007.

[17] R. Natarajan and B. Shekar, “A Relatedness-Based Data-Driven Approach to Determination of Interestingness of Association Rules,” Proc. 2005 ACM Symp. Applied Computing (SAC), pp. 551- 552, 2005.

[18] A.C.B. Garcia and A.S. Vivacqua, “Does Ontology Help Make Sense of a Complex World or Does It Create a Biased Interpretation?” Proc. Sense making Workshop in CHI ’08 Conf. Human Factors in Computing Systems, 2008.

[19] A.C.B. Garcia, I. Ferraz, and A.S. Vivacqua, “From Data to Knowledge Mining,” Artificial Intelligence for Eng. Design, Analysis and Manufacturing, vol. 23, pp. 427-441, 2009.

[20] Bayardo, R., Agrawal, R, and Gunopulos, D."Constraint-based rule mining in large, dense databases. "To appear in *ICDE-99*, 1999.

[21] Knowledge-Based Interactive Postmining of association rules using ontologies and rule schemas IEEE Transactions on knowledge and data engineering, VOL. 22, NO. 6, JUNE 2010.

[22] Pruning and Summarizing the Discovered Associations *ACM SIGKDD* International Conference on Knowledge Discovery & Data Mining (KDD-99), August 15-18, 1999, San Diego, CA, USA.