

IMPROVING POST-MINING OF ASSOCIATION RULES WITH ONTOLOGIES

Claudia Marinica¹ and Fabrice Guillet¹

¹LINA – Polytechnique School of University of Nantes,
Rue Christian Pauc, BP 50609, 44306, Nantes, France
E-mail: {Claudia.Marinica, Fabrice.Guillet}@univ-nantes.fr

Abstract: In data mining, the discovery of association rules is strongly limited by the huge amount of delivered rules. In this paper, we propose to improve post-processing of association rules by a better integration of user (decision maker) goals and knowledge. On one hand, we integrate a domain ontology associated to data which extends Generalized Association Rules. On the other hand, we generalize General Impressions with Rule Schemas. Ontologies will offer a powerful representation of user knowledge, and rule schemas a more expressive representation of user expectations in term of rules.

Keywords: Data Mining, Knowledge Management, Association Rules, Post-processing, Domain Knowledge, Ontologies.

1. Introduction

Data mining is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns from data (Fayyad *et al.*, 1996). One important topic in data mining is concerned with the discovery of interesting association rules (Agrawal *et al.*, 1993). Association rules mining allows non-supervised discovery of implicative and interesting tendencies in databases. An association rule $a \rightarrow b$ implies the presence of the itemset b when an itemset a occurs in a database transaction.

As a huge number of rules is generated, analyzing discovered rules is a hard task for the decision maker. To fulfil this hole, is crucial to help the decision maker with an efficient post-processing step adapted to association rules. A post-processing step refers to reducing the rule number using different methods like interesting measures (filtering and ranking), redundancy reduction or visualisation. In this paper, we are interested in post-processing methods base on interesting measures integrating decision maker knowledge and beliefs. Recently, Liu *et al.* (1997) have introduced the notion of General Impressions (GI) to improve the quality of discovered knowledge. General impressions represent expert’s vague feelings about the database domain.

Srikant and Agrawal (1995) worked on a problematic of finding Generalized Association Rules using a taxonomy of mined data (an *is-a* hierarchy). But these representations are limited to an *is-a* relation. To fulfil this hole, ontologies are proposed as knowledge representation languages.

In this paper, we propose to improve post-processing of association rules by a better integration of user (decision maker) goals and knowledge. Our basic idea relies on the fact that the better the user knowledge is given, the better the algorithm can focus on useful patterns. More precisely, we extend Generalized Association Rules (Srikant and Agrawal, 1995) and General Impressions (Liu *et al.*, 1997). On one hand, we integrate a domain ontology associated to data which extends Generalized Association Rules. On the other hand, we generalize General Impressions with Rule Schemas. Ontologies will offer a powerful representation of user knowledge, and Rule Schemas a more expressive representation of user expectations in term of rules.

The paper is structured as follows. Section 2 describes the research domain and reviews main related works. Section 3 explains the main principles of proposed framework. Section 4 is devoted each to one element of the framework: the mining process, the ontologies and the rule schemas. Finally, section 5 presents the conclusion and shows directions for future research.

2. Related work

Subjective measures of interestingness to integrate user knowledge were highlighted as early as 1994 in the KEFIR system (Piatetsky-Shapiro and Matheus, 1994) which discovers and explains *key findings*. In

1994, Klemettinen *et al.* employ the *templates* notion to describe the form of interesting rules, and also to specify which rules are not interesting. Silbershatz and Tuzilin (1996) suggest a classification of user beliefs in hard and soft beliefs and a classification of interestingness measures is actionability and unexpectedness.

In 1997, Liu *et al.* propose two user knowledge representation types: user's vague feelings (General Impressions - GI) and user's prior knowledge (Reasonably Precise Knowledge - RPK). Padmanabhan and Tuzhuilin (1997) introduce user beliefs represented as a set of rules over attributes, and propose to discover surprising rules using an interestingness measure based on logical contradiction.

Ontologies can be used in data mining as domain ontologies, ontologies for data mining process or metadata ontologies (Nigro *et al.*, 2007). Domain ontologies express application domain knowledge. Ontologies for data mining process codify all knowledge about the process, for example it can select the best algorithms according the problem. Metadata ontologies represent the description of variables process construction.

In domain ontology filed, Cespivova *et al.* (2004) suggest that an ontology of background knowledge can bring benefits in all phases of a KDD cycle described in CRISP-DM. From business understanding to deployment, the authors deliver a complete example of using ontologies in a cardiovascular risk domain. Starting from the same idea, Euler and Scholz (2004) use ontologies in the pre-processing phase of CRISP-DM methodology.

Chen *et al.* (2004) use ontology to improve rule support in data. They raise items to a lower level in ontology and then, they apply the Apriori algorithm. The difference with generalized association rules is that here they use a level raising and mining.

3. The Framework

In our approach we propose to define a new formal environment for integrating knowledge into this specific mining process. It is composed of two main parts (as shown in the Figure 1). On the one hand, we have the basic rule mining process with its main elements like database and association rules sets. On the other hand, we have two distinct knowledge representations.

[1] The dataset consists in n transactions described using p attributes. Let $I = \{I_1, I_2, \dots, I_p\}$ be a set of attributes. Also, let $T = \{t_1, t_2, \dots, t_n\}$ be the transaction set. Each transaction t_i is a set of items, called itemset, such as $t_i \subset I$. Let X be an itemset; we say that t_i contains X if $X \subset t_i$. An itemset selects a set of transactions as following: $T(X) = \{t \in T | X \subseteq t\}$.

[2] An *association rule* is an implication $X \rightarrow Y$, where X and Y are two itemsets and $X \cap Y = \emptyset$. This rule holds on T with the *confidence* c and the *support* s .

[1]-[2] Since its early definition, association rules are generated using *Apriori* algorithm proposed for the first time in Agrawal *et al.*, 1993.

[3] **Definition 1.** Formally, an ontology is a 4-tuple $O = \{C, R, H, A\}$. $C = \{C_1, C_2, \dots, C_o\}$ is a set of concepts and $R = \{R_1, R_2, \dots, R_r\}$ is a set of relations defined over concepts. H is a directed acyclic graph (DAG) over concepts defined by the subsumption relation (is-a relation, \leq) between concepts. We say that C_2 "is-a" C_1 , $C_2 \leq C_1$, if the concept C_1 subsumes the C_2 concept.

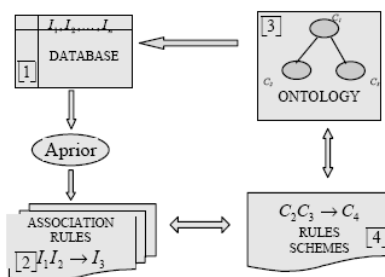


Fig. 1. Graphical representation of the proposed approach

[1]-[3] The ontology integrating our approach allows mining items representation in a hierarchical structure. In this scenario, it is fundamental to connect the ontology to the database. Each concept is instantiated in several transactions in database. For example, the easiest way to connect concepts to database is to

associate each concept from the ontology to an itemset in the database and to select the transactions containing the itemset: $f_o : C \rightarrow 2^I$ and $c \rightarrow f_o(c) = X$ where c is a concept and X is an itemset. More generally, to connect concepts to database we use logical expression defined over items: $f : C \rightarrow 2^I$ and $c \rightarrow f(c) = \text{logical expression defined over } I$.

[4] To improve association rule selection, we propose a filtering box based on a new type of rules that we call Rules Schemas. A rule schema describes the user knowledge and, based on ontology concepts, it is defined in a rule-like formalism comparable to association rule one. A semantically extension of General Impressions notion, rules schemas bring the complexity of ontologies in rule mining combining not only item constraints, but also ontology concept constraints.

Definition 2. A rule schema is a rule $X_1, X_2, \dots, X_{s1} \rightarrow Y_1, Y_2, \dots, Y_{s2}$ where X_i and Y_j are defined like different constraints over concepts. For example, a rule schema $C_2, \overline{C_3} \rightarrow C_4$ corresponds to “all association rules whom condition verifies C_2 and doesn't verify the concept C_3 , and whom conclusion verifies C_4 ”.

[2]-[4] The goal of our approach is to select interesting rules for decider maker post-processing discovered rules. In this purpose Let $R: X \rightarrow Y$ be an association rule and $T(X)$ and $T(Y)$ be the transactions selected by itemsets X , and, respectively, Y . Let $S: C_1 \rightarrow C_2$ be a rule schema with $C_1 \rightarrow f(C_1)$ and $C_2 \rightarrow f(C_2)$, and $T(f(C_1))$ and $T(f(C_2))$ the transactions selected by C_1 , respectively C_2 . We say that R is conforming to S if: $T(X) \subseteq T(f(C_1))$ and $T(Y) \subseteq T(f(C_2))$.

4. Case of study

We illustrate our approach on a simple example based on a database describing pizza composition with the advantage of being easy to interpret. Each item of the 99 ones corresponds to a pizza ingredient and each transaction of the 35 ones existing in the database corresponds to a pizza.

Different implementations of the *Apriori* algorithm have been proposed in the literature (Bodon, 2006, Borgelt et Kruse, 2002). We use *ARMiner*¹ software to extract association rules, and we fixe a minimum support of 8% and a minimum confidence of 70%. The algorithm extracts 253 association rules.

4.1. Ontology conceptual structure and ontology-database connection

We use the ontology *pizza.owl* proposed in the example set by Protégé software. The advantage of *pizza*-ontology is that it has generally the same granularity with our database. The ontology contains three main hierarchies based on three root concepts: *Pizza*, *PizzaTopping* and *Spiciness*. The first hierarchy, *Pizza*, describes sets of pizzas (for example: *CheeseyPizza*, *FishyPizza*, *SweetyPizza*).

Definition 3. Let C_d be a concept belonging to *Pizza* hierarchy. C_d is defined by a logical expression over a set of concepts of two other hierarchies. Let us consider the concept *CheezyPizza*; in natural language, a pizza belongs to this type if it contains at least one cheese ingredient. *CheezyPizza* concept represents all pizzas which contain at least one concept subsumed by the concepts *CheezyTopping*.

The second one, *PizzaTopping* is composed by all pizza ingredients. For example, *ChoppedCheese* and *RicottaCheese* concepts are generalized by *GoatCheese* concept, and *GoatCheese* is generalized by *CheeseTopping* concept. The third one, *Spiciness*, contains three specific concepts describing the spiciness of ingredients: *Hot*, *Medium* and *Mild*. Moreover, the subsumption relation is completed by two other relations between concepts: *hasTopping* property between *Pizza* concept and *PizzaTopping* concept, and *hasSpiciness* property between *PizzaTopping* concept and *Spiciness* concept.

Concerning the ontology-database connection, several connection types can be conceived. The simplest ontology-database connection is the direct one. It connects one concept of the *PizzaTopping* hierarchy to an item or an itemset (semantically, the nearest one) and associate the concept to the corresponding transaction set. Let *Mushrooms* be a concept of the ontology defined by: "*Mushrooms*" $\rightarrow f(\text{"Mushrooms"}) = \text{"Champignons"}$. It will select all the transactions containing the item *Champignons*.

A second type of connection implies connecting concepts of *Pizza* hierarchy to database. We know that C_d concept of *Pizza* hierarchy is defined by a logical expression over a set of concepts. Considering

¹ <http://www.cs.umb.edu/~laur/ARMiner/>

that each concept C_1, \dots, C_k in C_d description are direct connected to an item. Thus, C_d is defined by a logical expression over a set of items: $f : C \rightarrow 2^I$ and $c_d \rightarrow f(c_d) = \text{logical expression defined over } I$

Definition 4. Considering the concept *CheezyPizza*. It is defined over the set of items I by the following logical expression: $C_{d1} = \text{"CheezyPizza"}$ and $C_{d1} \rightarrow f(C_{d1}) = \text{"CheeseTopping"}$. The concept *CheezyPizza* selects all transactions satisfying the logical expression.

4.2. Rules Schemas and Results

A rule schema allows user knowledge representation and it permits to user to supervise association rule mining. Let us consider the rule schema *SweetyPizza* \rightarrow *CheezyPizza* (RS_1). The *CheezyPizza* concept is described in the section §4.1 and the *SweetyPizza* concept is described by the pizzas containing at least one sweet ingredient. Considering the following association rule extracted by the *Apriori* algorithm:

$$\text{Sugar} \rightarrow \text{Cream} \quad \text{Support} = 0.08571429 \quad \text{Confidence} = 0.75$$

This association rule is conforming to rule scheme (RS_1) since the itemset in the condition of the rule selects a set of transactions included in the set of transactions selected by the concept *SweetyPizza*:

$$\begin{aligned} T(\text{"SweetyPizza"}) &= \{1, 16, 23, 24, 33, 34, 35\} \\ T(\text{"Sugar"}) &= \{1, 33, 34, 35\} \end{aligned} \Rightarrow T(\text{"Sugar"}) \subset T(\text{"SweetyPizza"})$$

In the same way, the itemset in the conclusion of the rule selects a set of transactions included in the set of transactions selected by the concept *CheezyPizza*. Thus, one association rules is selected:

$$\text{Sugar} \rightarrow \text{Cream} \quad \text{Support} = 0.08571429 \quad \text{Confidence} = 0.75$$

Selecting rules conforming to a rule schema is interesting for the decision maker since he wants to know if the implication between the concepts of the rule schema exists. But, if the decision maker is convinced by the existence of this implication, is more interesting to show him what he doesn't know.

Therefore, starting from RS_1 , we can select other types of rules. For example, we can filter exception rules selecting those one conforming to a new rule schema created by modifying RS_1 as it follows:

$$\text{"SweetyPizza"} \wedge ? \rightarrow \text{"CheezyPizza"}.$$

5. Conclusion

This paper discusses the problem of helping the decision maker in the post-processing step of association rule mining. We propose to integrate user knowledge and beliefs to reduce the number of rules extracted by *Apriori* algorithm. User knowledge are modeled in an ontology connected to data. Rule schemas allow user beliefs representation, and, combined with ontologies, they improve the selection of interesting rules.

We intend to pursue this approach improving it by two directions: developing the rule schema formalism and integrating the approach is the discovering algorithm.

References

- Agrawal, R.; Imielinski, T. and Swami, A. 1993. *Mining Association Rules between Sets of Items in Large Databases*, 12th ACM SIGMOD International Conference on Management of Data, 207–216.
- Cespivova, H.; Rauch, J.; Svatek, V.; Kejkula, M. and Tomeckova, M. 2004. *Roles of Medical Ontology in Association Mining CRISP-DM Cycle*, Workshop Knowledge Discovery and Ontologies in ECML/PKDD.
- Chen, X.; Zhou, X.; Scher, R. and Geller, J. 2003. *Using an interest Ontology for Improved Support in Rule Mining*, 5th International Conference of Data Warehousing and Knowledge Discovery, 320–329.
- Euler, T. and Scholz, M. 2004. *Using Ontologies in a KDD Workbench*. Workshop on Knowledge Discovery and Ontologies at ECML/PKDD.
- Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P. and Uthurusamy, R. 1996. *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press.
- Klemettinen, M.; Mannila, H.; Ronkainen, P.; Toivonen, H. and Verkamo, A. I. 1994. *Finding Interesting Rules from Large Sets of Discovered Association Rules*. Int. Conference on Information and Knowledge Management (CIKM), 401–407.
- Liu, B.; Hsu, W. and Chen, S. 1997. *Using General Impressions to Analyze Discovered Classification Rules*. Knowledge Discovery and Data Mining, 31–36.

- Nigro, H. O.; Gonzalez Cisaro, S. E. and Xodo, D. H. 2007. *Data Mining With Ontologies: Implementations, Findings and Frameworks*, Idea Group Reference.
- Padmanabhan, B. and Tuzhilin, A. 1997. *Unexpectedness as a Measure of Interestingness in Knowledge Discovery*. Workshop on Information Technology and Systems (WITS), 81–90.
- Piatetsky-Shapiro, G. and Matheus, C. J. 1994. *The Interestingness of Deviations*. Knowledge Discovery in Databases, AAAI Workshop, 25–36.
- Silberschatz, A. and Tuzhilin, A. 1996. *What Makes Patterns Interesting in Knowledge Discovery Systems*, IEEE Transactions on Knowledge and Data Engineering, 8: 970–974.
- Srikant, R. and Agrawal, R. 1995 *Mining Generalized Association Rules*. 21st International Conference on Very Large Databases, 407–419.