Mining association rules using formal concept analysis

Nicolas Pasquier

Mining Association Rules using Formal Concept Analysis

Nicolas Pasquier

LIMOS, Université Blaise Pascal - Clermont-Ferrand II,
24 Avenue des Landais, F–63177 Aubière, France
pasquier@libd2.univ-bpclermont.fr

Abstract. In this paper, we give an overview of the use of Formal Concept Analysis in the framework of association rule extraction. Using frequent closed itemsets and their generators, that are defined using the Galois closure operator, we address two major problems: response times of association rule extraction and the relevance and usefulness of discovered association rules. We quickly review the Close and the A-Close algorithms for extracting frequent closed itemsets using their generators that reduce response times of the extraction, specially in the case of correlated data. We also present definitions of the generic and informative bases for association rules which generation improves the relevance and usefulness of discovered association rules.

1 Introduction

Data mining has been extensively addressed for the last years as the computational part of Knowledge Discovery in Databases (KDD), specially the problem of discovering association rules. Its aim is to exhibit relationships between itemsets (sets of binary attributes) in large databases. An example of association rules, fitting in the context of market basket data analysis, is "cereal $\wedge$ milk $\rightarrow$ sugar (support 10%, confidence 60%)" stating that 60% of customers who buy cereals and sugar also buy milk and that 10% of all customers buy all three items. When an association rule has support and confidence exceeding some user-defined minimum support and minimum confidence thresholds, the rule is considered as relevant for supporting decision making [AIS93]. Association rules have been successfully applied in a wide range of domains, among which marketing decision support, diagnosis and medical research support, telecommunication process improvement, web site management and access, the analysis of multimedia, spatial, geographical and statistical data, etc.

The first phase of the association rule extraction consists in selecting useful data from the database and transforming it in a data mining context. This context is a triplet $\mathcal{D} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$, where $\mathcal{O}$ and $\mathcal{I}$ are finite sets of objects and items respectively, and $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{I}$ is a binary relation. Each couple $(o, i) \in \mathcal{R}$ denotes the fact that the object $o \in \mathcal{O}$ is related to the item $i \in \mathcal{I}$. Two major problems for the association rule extraction give place to interesting research topics: the problem of response times of the extraction and the problem of the relevance and the usefulness of extracted association rules.

2 Response times of the extraction

Existing approaches for mining association rules are based on the following decomposition of the problem: the extraction of frequent itemsets[1] and their supports from the context and then the generation of all valid association rules[2]. The first phase is the most computationally intensive part of the process, since the number of potential frequent itemsets is exponential in the size of the set of items and several database passes are required. Two approaches have been proposed: levelwise algorithms for extracting frequent itemsets and algorithms for extracting maximal frequent itemsets. These algorithms give acceptable response times when mining association rules from weakly correlated data, such as market basket data, but their performances drastically decrease when they are applied to correlated data, such as statistical or medical data for instance. We recall these two approaches and present then our approach which is based on Formal Concept Analysis [GW99].

2.1 Levelwise algorithms for extracting frequent itemsets

These algorithms consider during each iteration a set of itemsets of a given size, i.e., a set of itemsets of a "level" of the itemset lattice. These algorithms are based on the following properties in order to limit the number of candidate itemsets considered: all the supersets of an infrequent itemset are infrequent and all the subsets of a frequent itemset are frequent [AS94,MTV94]. Using this property, the candidate k-itemsets[3] of the k^{th} iteration are generated by joining two frequent (k-1)-itemsets discovered during the preceding iteration. The Apriori [AS94] and OCD [MTV94] algorithms carry out a number of scans of the context equal to the size of the largest frequent itemsets. The Partition [SON95] algorithm allows the parallelization of the process of extraction and the algorithm DIC [BMUT97] reduces the number of context scans by considering itemsets of different sizes during each iteration. The Partition and DIC algorithms involve additional costs in CPU time compared to the Apriori and OCD algorithms due to the increase in the number of candidate itemsets tested.

2.2 Algorithms for extracting maximal frequent itemsets

These algorithms are based on the property that the maximal frequent itemsets, i.e., the frequent itemsets of which all the supersets are infrequent, form a border under which all itemsets are frequent. The extraction of the maximal frequent itemsets is carried out by an iterative browsing of the itemset lattice that "advances" by one level from the bottom upwards and by one or more levels from the top downwards during each iteration. Using the maximal frequent

[1] An itemset is frequent if its support is greater or equal to the minimal support threshold.

[2] An association rule is valid if its support and confidence are at least equal to the minimal support and the minimal confidence thresholds.

[3] An itemset of size k is called a k-itemset.

itemsets, all the frequent itemsets are derived and their supports are determined by performing one final scan of the context. Four algorithms based on this approach were proposed; they are the Pincer-Search [LK98], MaxClique and Max-Eclat [ZPOL97], and Max-Miner [Bay98] algorithms. These algorithms reduce the number of iterations, and thus decrease the number of context scans and the number of CPU operations carried out, compared to levelwise algorithms for extracting frequent itemsets.

2.3 Algorithms for extracting frequent closed itemsets

In contrast to the two previous approaches, our approach [PBTL99a] is based on Formal Concept Analysis. The closure operator γ of the Galois connection [GW99] is the composition of the application ϕ, that associates with $O \subseteq \mathcal{O}$ the items common to all objects $o \in O$, and the application ψ, that associates with an itemset $I \subseteq \mathcal{I}$ the objects related to all items $i \in I$ (the objects "containing" I). The closure operator $\gamma = \phi \circ \psi$ associates with an itemset I the maximal set of items common to all the objects containing I, i.e., the intersection of these objects. Using this closure operator, the *frequent closed itemsets* are defined.

Definition 1 (Frequent closed itemsets). *A frequent itemset $I \subseteq \mathcal{I}$ is a frequent closed itemset iff $\gamma(I) = I$.*

The frequent closed itemsets constitute, together with their supports, a generating set for all frequent itemsets and their supports and thus for all association rules, their supports and their confidences [PBTL99a]. This property relies on the properties that the support of a frequent itemset is equal to the support of its closure and that the maximal frequent itemsets are maximal frequent closed itemsets. Two efficient levelwise algorithms, called Close [PBTL99a] and A-Close [PBTL99b], for extracting frequent closed itemsets from large databases were proposed. In order to improve the efficiency of the extraction, the Close and the A-Close algorithms consider the *generator itemsets* of the frequent closed itemsets.

Definition 2 (Generator itemsets). *An itemset $G \subseteq \mathcal{I}$ is a generator of a closed itemset I iff $\gamma(G) = I$ and $\nexists G' \subseteq \mathcal{I}$ with $G' \subset G$ such that $\gamma(G') = I$.*

Close and A-Close perform a breadth-first search for the (frequent) generators of the frequent closed itemsets in a levelwise manner. During an iteration k, the Close algorithm considers a set of candidate generators of size k, it determines their supports and their closures, and then deletes all infrequent generators. The supports and the closures of the candidate k-generators are computed by performing one database pass and, for each generator G, intersecting all the objects containing G (their number gives the support of G). During the $(k+1)^{th}$ iteration, the candidate $(k+1)$-generators are constructed by joining two frequent k-generators if their $k-1$ first items are identical, and the candidate $(k+1)$-generators obtained are pruned if they are known to be infrequent or their closure

is already computed. In the A-Close algorithm, the generator itemsets are identified according to their supports only, since the support of a generator itemset is different from the supports of all its subsets, and one more database pass is performed at the end of the algorithm for computing the closures of all frequent generators discovered. Both algorithms initialize at the begining the set of candidate 1-generators with the list of all itemsets of size 1. Experimental results show that these algorithms are particularly efficient for mining association rules from dense or correlated data that represent an important part of real life databases. On such data, Close outperforms A-Close, and they both clearly outperform algorithms for extracting frequent itemsets, whereas for weakly correlated data, A-Close outperforms Close and is in the range of algorithms described in section 2.1.

3 Relevance of extracted association rules

The problem of the usefulness and the relevance of discovered association rules is related to the huge number of rules extracted and the presence of many redundancies among them for many datasets, especially for correlated data. Several approaches for solving this problem have been proposed. We first quickly review these approaches and present then the approach we propose that consists in generating non-redundant association rules with minimal antecedents and maximal consequents using Formal Concept Analysis.

3.1 Previous work

The use of statistic measures other than confidence, such as conviction, Pearson's correlation or χ^2 test, to compute the precision of rules is proposed in [BMS97,SBM98]. Generalized association rules, that are rules between itemsets that belong to different levels of a taxonomy of the items, are defined in [HF95,SA95]. In [Hec96,ST96], deviation measures, i.e., measures of distance between association rules used for pruning similar ones, are defined using support and confidence. Item constraints [BAG99,NLHP98] are boolean expressions that allow the user to specify the form of association rules that will be selected. In [BG99], A-maximal rules, that are rules for which the population of objects concerned is reduced when an item is added to the antecedent, are defined. In [PBTL99c], the Duquenne-Guigues basis for global implications [DG86,GW99] and the Luxenburger basis for partial implications [Lux91] are adapted to the association rules framework. These bases are minimal with respect to the number of rules extracted, but they are not made up of the most informative association rules that are non-redundant rules with minimal antecedents and maximal consequents, called *minimal non-redundant association rules*. We believe that these rules are the most relevant and useful from the point of view of the user, considering the fact that in practice the user cannot infer all other valid rules from the rules extracted while visualizing them. None of the approaches proposed in previous work allows to generate only these rules.

3.2 Minimal non-redundant association rules

From the point of view of the user, an association rule is redundant if it conveys the same information – or less general information – than the information conveyed by another rule of the same range (support) and the same precision (confidence). In previous work for reducing redundant implication rules (functional dependancies), the notion of non-redundancy considered is related to the inference system using Armstrong axioms [Arm74]. This notion is not to be confused with the notion of non-redundancy we consider here. To our knowledge, such an inference system for association rules, i.e., taking into account supports and confidences of the rules, does not exist. An association rule $r \in E$ is non-redundant and minimal if there is no other association rule $r' \in E$ with same support and confidence and, which antecedent is a subset of the antecedent of r and which consequent is a superset of the consequent of r.

Definition 3 (Minimal non-redundant association rules). *An association rule $r : I_1 \rightarrow I_2$ is a minimal non-redundant association rule iff not exists an association rule $r' : I'_1 \rightarrow I'_2$ such that* support(r) = support(r'), confidence(r) = confidence(r'), *$I'_1 \subseteq I_1$ and $I_2 \subseteq I'_2$.*

Given this characterization, we define the generic basis for exact association rules (100% confidence rules) and the informative basis for approximate association rules. These bases are constituted of the minimal non-redundant exact and approximate association rules respectively. Let $\mathcal{FC}$ be the set of frequent closed itemsets and let $\mathcal{FG}$ be the set of their (minimal) generators.

Definition 4 (Generic basis). *The generic basis contains all rules with the form $r : G \rightarrow (F \setminus G)$ between a generator itemset $G \in \mathcal{FG}$ and its closure $\gamma(G) \in \mathcal{FC}$ such that $G \neq \gamma(G)$.*

Definition 5 (Informative basis). *The informative basis contains all rules with the form $r : G \rightarrow (F \setminus G)$ between a generator itemset $G \in \mathcal{FG}$ and a frequent closed itemset $F \in \mathcal{FC}$ that is a superset of its closure: $\gamma(G) \subset F$. The transitive reduction of this basis, i.e., for $\nexists F' \in \mathcal{FC}$ such that $\gamma(G) \subset F' \subset F$, is also a basis for all approximate association rules.*

All valid association rules, their supports and their confidences can be deduced from the union of the generic basis and the informative basis or its transitive reduction. Results of experimentations conducted on real-life databases show that their generation is efficient and useful in practice, particularly when mining association rules from correlated data.

References

[AIS93] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. *Proc. SIGMOD conf.*, 207–216, May 1993.

[AS94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. *Proc. VLDB conf.*, 478–499, September 1994.

[Arm74] W. W. Armstrong. Dependency structures of data base relationships. *Proc. IFIP congress*, pp 580–583, August 1974.

[Bay98] R. J. Bayardo. Efficiently mining long patterns from databases. *Proc. SIGMOD conf.*, 85–93, June 1998.

[BAG99] R. J. Bayardo, R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense databases. *Proc. ICDE conf.*, 188–197, March 1999.

[BG99] R. J. Bayardo, and R. Agrawal. Mining the most interesting rules. *Proc. KDD Conference*, 145–154, August 1999.

[BMUT97] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. *Proc. SIGMOD conf.*, 255–264, May 1997.

[BMS97] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlation. *Proc. SIGMOD conf.*, 265–276, May 1997.

[DG86] V. Duquenne and J.-L. Guigues. Famille minimale d'implications informatives résultant d'un tableau de données binaires. *Mathématiques et Sciences Humaines*, 24(95):5–18, 1986.

[GW99] B. Ganter and R. Wille. *Formal Concept Analysis : Mathematical foundations.* Springer, 1999.

[HF95] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. *Proc. VLDB conf.*, 420–431, September 1995.

[Hec96] D. Heckerman. Bayesian networks for knowledge discovery. *Advances in Knowledge Discovery and Data Mining*, 273–305, 1996.

[LK98] D. Lin and Z. M. Kedem. Pincer-Search : A new algorithm for discovering the maximum frequent set. *Proc. EBDT conf.*, 105–119, March 1998.

[Lux91] M. Luxenburger. Implications partielles dans un contexte. *Mathématiques, Informatique et Sciences Humaines*, 29(113):35–55, 1991.

[MTV94] H. Mannila, H. Toivonen, and A. I. Verkamo. Efficient algorithms for discovering association rules. *AAAI KDD workshop*, 181–192, July 1994.

[NLHP98] R. T. Ng, V. S. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning optimizations of constrained association rules. *Proc. SIGMOD conf.*, 13–24, June 1998.

[PBTL99a] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1):25–46, 1999.

[PBTL99b] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. *Proc. ICDT conf.*, 398–416, January 1999.

[PBTL99c] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Closed set based discovery of small covers for association rules. *Proc. BDA conf.*, 361–381, October 1999.

[SON95] A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. *Proc. VLDB conf.*, 432-444, September 1995.

[ST96] A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974, December 1996.

[SBM98] C. Silverstein, S. Brin, and R. Motwani. Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2(1):39–68, January 1998.

[SA95] R. Srikant and R. Agrawal. Mining generalized association rules. *Proc. VLDB conf.*, 407–419, September 1995.

[ZPOL97] M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. *Proc. KDD conf.*, 283–286, August 1997.