Mining Concise Representations of Frequent Patterns through Conjunctive and Disjunctive Search Spaces

Tarek HAMROUNI
Computer Science Department, Faculty of Sciences of Tunis, Tunis El Manar University, Tunisia
tarek.hamrouni@fst.rnu.tn
Computer Science Research Center of Lens, Artois University, France
hamrouni@cril.univ-artois.fr

The last years witnessed an explosive progress in networking, storage, and processing technologies resulting in an unprecedented amount of digitalization of data. There is hence a considerable need for tools or techniques to delve and efficiently discover valuable, non-obvious information from large databases. In this situation, Knowledge Discovery in Databases offers a complete process for the non-trivial extraction of implicit, previously unknown, and potentially useful knowledge from data. Amongst its steps, data mining offers tools and techniques for such an extraction. Much research in data mining from large databases has focused on the discovery of association rules which are used to identify relationships between sets of items in a database. The discovered association rules can be used in various tasks, such as depicting purchase dependencies, classification, medical data analysis, etc. In practice however, the number of frequently occurring itemsets, used as a basis for rule derivation, is very large, hampering their effective exploitation by the end-users. In this situation, a determined effort focused on defining manageably-sized sets of patterns, called concise representations, from which redundant patterns can be regenerated. The purpose of such representations is to reduce the number of mined patterns to make them manageable by the end-users while preserving as much as possible the hidden and interesting information about data.

Within the traditional association analysis, the conjunction connector – linking items – got the monopoly. This was motivated by the original application pertaining to market basket analysis. In this respect, a growing number of approaches explored the conjunctive search space where items are characterized by the frequency of their *simultaneous occurrence* (or *co-occurrence*). The aim of such an exploration is to get out a lossless nucleus of itemsets, from which the remaining ones can be exactly derived. Many exact concise representations of frequent patterns were thus proposed in the literature. Among the numerous concise representations, the ones based on closed itemsets and minimal generators (*aka* minimal keys, free itemsets, key itemsets, and intent reducts) got a large interest since their respective proposals [1]. The representation based on closed itemsets heavily relies on an operator which makes it possible mapping an important number of elements – from the frequent itemset search space – into a single element within that of frequent closed itemsets. On its side, the minimal generator-based representation takes advantage from its efficient computation thanks to the interesting structural properties of the minimal generator set. In fact, these itemsets are closely linked. Indeed, once applied, the aforementioned operator partitions the set of frequent itemsets into equivalence classes. Each class contains itemsets characterizing the same set of objects. These itemsets hence share the same closure obtained by intersecting the associated objects. The closed itemset is then the *unique maximal* set of items characterizing a set of objects. While, *often several* minimal generators constitute the *minimal* elements of each class. The aforementioned link between closed patterns and minimal generators explains why they are often simultaneously used for concisely representing pattern classes, like frequent itemsets, (classification) association rules, sequential patterns, etc. In this respect, the interesting structural properties of the minimal generator set made it a key step for mining important pattern classes as well as for knowledge interpretation. Indeed, according to the minimum description length (MDL) principle, minimal generators are the preferred representation of an equivalence class in applications like model selection, classification, etc. Unfortunately, the number of minimal generators is usually larger than that of closed itemsets. This is explained by the fact that several minimal generators cohabit in the same equivalence class and, hence, convey the same information. An intra-class combinatorial redundancy would then logically results from the inherent absence of a unique minimal generator associated to a given closed itemset. Thus, a same piece of knowledge is redundantly conveyed by distinct minimal generators.

This motivated us to carry out an in-depth study aiming at only retaining irreducible minimal generators in each equivalence class, and pruning the remaining ones without information loss. Our aim is thus to approach as much as possible the "ideal" case which consists in only retaining a unique irreducible minimal generator per equivalence class. In this respect, we propose lossless reductions of the minimal generator set thanks to a new substitution-based process that avoids the internal interchangeability of the minimal generators [2; 5]. This consists in localizing minimal generators mutually reachable by permutation of their respective subsets. This allows to group them in *finer* equivalence classes, induced by a *dedicated substitution operator*. Then, only a representative in each class is retained, while the remaining ones are omitted since redundant. We then carry out a thorough study of the associated properties of the obtained families [2; 5]. Our theoretical results can be useful for the different extensions and applications of minimal generators or their similar constructs. In this thesis, they are applied

within the association rule framework in order to reduce as much as possible the number of retained rules without information loss [2]. We then give a thorough formal study of the related inference mechanism allowing to derive all redundant association rules, starting from the retained ones. A new approach is then proposed, dedicated to the extraction of a lossless subset of generic association rules based on redundancy-free minimal generators as a starting point [2]. In order to validate our approach, computing means for the new pattern families are presented together with empirical evidences about their relative sizes *w.r.t.* the entire sets of patterns.

In the literature, other item links such as their complementary/mutually occurrences – rather than item co-occurrences – were neglected and only some recent works highlight the added-value of this type of knowledge. Indeed, in practice the following situations can arise. Suppose that a market basket data is under treatment, and the manager is searching for items c_1, c_2, ..., and c_n whose selling implies that of at least one of two competitive products a and b (or probably both), *i.e.*, the items fulfilling the condition: $c_1 \vee c_2 \vee \ldots \vee c_n \Rightarrow a \vee b$ is always true. Such a rule conveys knowledge about the items sold simultaneously with a or b. Since the disjunction connector $\vee$ is inclusive, the simultaneous selling of c_i and c_j ($i \neq j$) is possible. On the other hand, in a textmining application related to text translation from a language l_1 to a language l_2, an analyst may be interested in the possible translations in the language l_2 of a given term t belonging to the language l_1. In this respect, t may have several translations tr_1, tr_2, ..., and tr_n in the language l_2 according to its usage context. Thus, a rule like $tr_1 \vee tr_2 \vee \ldots \vee tr_n \Rightarrow t$ is interesting since it summarizes the possible translation of t. In both cases, more computations may be performed to get more precise information about the effect of a given product (*resp.* terms) among c_i (*resp.* t_j) on the appearance of a and b (*resp.* t). Various other applications of disjunctive itemsets are possible in the contexts of social network analysis, bioinformatics, etc. In such situations, the disjunction connector linking items can bring key information as well as a summarizing method of the conveyed knowledge. Such knowledge may not be obtained even by a collection of conjunctive patterns.

Due to the close link between frequent itemsets and association rules, it is more advantageous to mine a concise representation of frequent itemsets that offers direct access to the disjunctive support of frequent itemsets. Such a representation can be used as a starting point for mining generalized rules (*i.e.*, also involving disjunction and negation of items) based on frequently occurring itemsets. In this respect, we initially focus on the unique representation in the literature exploring the disjunctive search space, namely that based on *essential itemsets*. Being the equivalent of minimal generators within the disjunctive search space, essential itemsets bring interesting knowledge about the *complementary occurrence* of items in a dataset. However, several essential itemsets can characterize the same set of objects. This motivated us to propose a new *disjunctive closure operator* dedicated to the disjunctive search space to avoid such a redundancy [4]. This operator offers a redundancy-free representation of the disjunctive search space since it allows to select a unique element – the maximal one – to represent itemsets covering the same set of data. It hence makes it possible the proposal of reduced exact concise representations of frequent itemsets based on *disjunctive closed itemsets* [4]. This representation allows the efficient derivation of the different types of itemset supports, *i.e.*, conjunctive, disjunctive and negative. In addition, the proposed closure operator constitutes an interesting tool for the efficient exploration of the disjunctive search space. Such an exploration can be used for example towards the derivation of *generalized association rules* [3]. These latter rules generalize classic association rules – positive rules – to also offer disjunction and negation connectors between items, in addition to the conjunctive one. The associated quality measures can thus be derived using a representation based on disjunctive itemsets [3]. In the literature, generalized association rules are useful in different applications. For example, they are used as an intermediate step for defining concise representations for frequent itemsets. They are also exploited to provide the end-users with some new forms of association rules. Dedicated tools were then designed and implemented for extracting disjunctive itemsets and generalized association rules. Our experiments showed the usefulness of our exploration and highlighted interesting compactness rates.

Ph. D. dissertation committee

- Prof. Salem Benferhat (Examiner), Artois University, France
- Associate Prof., Sadok Ben Yahia (Co-supervisor), Tunis El Manar University, Tunisia
- Prof. Khaled Bsaies (Co-supervisor), Tunis El Manar University, Tunisia
- Prof. Marzena Kryszkiewicz (Reviewer), ICS-WUT, Poland
- Prof. Engelbert Mephu Nguifo (Co-supervisor), Blaise Pascal University, Clermont Ferrand, France
- Prof. Habib Ounelli (Reviewer), Tunis El Manar University, Tunisia

1. REFERENCES

[1] S. Ben Yahia, T. Hamrouni, and E. Mephu Nguifo. Frequent closed itemset based algorithms: A thorough structural and analytical survey. *ACM-SIGKDD Explorations*, 8(1):93–104, June 2006.

[2] T. Hamrouni, S. Ben Yahia, and E. Mephu Nguifo. Succinct minimal generators: Theoretical foundations and applications. *International Journal of Foundations of Computer Sciences*, 19(2):271–296, April 2008.

[3] T. Hamrouni, S. Ben Yahia, and E. Mephu Nguifo. GARM: Generalized association rule mining. In *Proceedings of the 6th International Conference on Concept Lattices and their Applications (CLA 2008), Olomouc, Czech Republic*, pages 145–156, October 2008.

[4] T. Hamrouni, S. Ben Yahia, and E. Mephu Nguifo. Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets. *Data & Knowledge Engineering, doi:10.1016/j.datak.2009.05.001*, 2009.

[5] T. Hamrouni, P. Valtchev, S. Ben Yahia, and E. Mephu Nguifo. About the lossless reduction of the minimal generator family of a context. In *Proceedings of the 5th International Conference on Formal Concept Analysis (ICFCA 2007), LNAI, Springer-Verlag, volume 4390, Clermont-Ferrand, France*, pages 130–150, February 2007.