Frequent pattern mining under generalized subsumption

Jan Ramon and Jan Struyf

Frequent pattern mining (including the discovery of association rules) is an important task in data mining. Recently, there is increasing interest in mining relational databases. Up to now, most algorithms have focussed on a syntactical approach. However, the use of background knowledge would greatly improve the quality of the results. First, patterns and rules which are not equivalent from a syntactical point of view, may be semantically equivalent. Taking into account the semantical relationships between patterns improves the comprehensibility while decreasing the size of the discovered set of patterns. Second, while the use of background knowledge increases the expressivity and therefore comes with a cost, it also allows to better exploit the benefits of some optimizations.

While some special cases (such as taxonomies) have been investigated before in the propositional case, our work is the first to investigate this in the most general case, using generalized subsumption to determine semantical relationships. Our work focuses on three main topics. First, there is the problem of the representation of the information and generalizing definitions. We study the trie representation of set of relational pattern, the condensed representation of sets of association rules, extending (De Raedt & Ramon, 2004), the types of background knowledge that are compatible with certain optimizations such as the computation of canonical forms of patterns etc. Second, we investigate the efficiency issues in relation to this more general setting of frequent pattern mining. Here, relevant topics are the transformation of logical queries (cut, once, reorder etc.), the pack and ad-pack mechanisms to avoid re-evaluation of a common prefix of sets of queries, efficient ways to perform the monotonicity test and equivalence test, etc. Third, we are working on an efficient implementation and experiments.