A Divide and Conquer Approach for Deriving Partially Ordered Sub-structures

S. Ben Yahia, Y. Slimani, and J. Rezgui

Département des Sciences de l’Informatique,
Faculté des Sciences de Tunis,
Campus Universitaire, 1060 Tunis, Tunisie
{sadok.benyahia, yahya.slimani}@fst.rnu.tn

Abstract. The steady growth in the size of data has encouraged the emergence of advanced main memory trie-based data structures. Concurrently, more acute knowledge extraction techniques are devised for the discovery of compact and lossless knowledge formally expressed by generic bases. In this paper, we present an approach for deriving generic bases of association rules. Using this approach, we construct small partially ordered sub-structures. Then, these ordered sub-structures are parsed to derive, in a straightforward manner, local generic association bases. Finally, local bases are merged to generate the global one. Extensive experiments carried out essentially showed that the proposed data structure allows to generate a more compact representation of an extraction context comparatively to existing approaches in literature.

1 Introduction

Classical approaches for extraction of such implicit knowledge suffer from the huge number of potentially interesting correlations (specially association rules) that can be drawn from a dataset. In order to limit the number of the reported rules, while conserving the "informativeness" property, a battery of results, provided by the mathematical foundations of the Formal Concept Analysis, yielded a compact and lossless subset of association rules, called generic bases of association rules [1]. In order to derive generic bases of association rules, the extraction of knowledge base problem can be reformulated as follows:(1) Discover two distinct "closure systems", i.e., sets of sets which are closed under the intersection operator,: the set of closed itemsets and the set of associated minimal generators. Also, the upper covers ($Cov^u$) of each closed itemset should be available. (2) From all the information discovered in the first step, i.e., two closure systems and the upper covers sets, derive generic bases of association rules (from which all the remaining rules can be derived). It is noteworthy that the recently proposed approaches advocate the use of advanced data structure, essentially based on tries structures, to store compactly in main memory input dataset [2, 3] or to store partial outputs (e.g.,[4]).

In this paper, we propose a new trie-based data structure called ITEMSET-TRIE. The ITEMSET-TRIE extends the idea claimed by the authors of FP-TREE [2]
and Cats [3] structures, aiming to improve storage compression and to allow (closed) frequent pattern mining without "explicit" candidate itemset generation step. Next, we propose an algorithm, falling in the characterization "Divide and Conquer" to extract frequent closed itemsets with their associated minimal generators. Hence, the derivation of approximative generic association rules is based on the exploration of such closed itemsets organized upon their natural partial order (also called precedence relation). That’s why we construct on the fly, concurrently with the closed itemsets discovery process, the local "iceberg lattice" [5]. Such local ordered sub-structures can be drawn quite naturally in a parallel manner. Then, these ordered sub-structures are parsed to derive, in a straightforward manner, local generic bases of association rules. Finally, local bases are merged to generate the global one. Such process can be recapitulated as follows: (i) Construct the Itemset-trie, (ii) Construct the local ordered structures, (iii) Merge the local generic association rules to derive a global one.

The remainder of the paper is organized as follows: In Section 2, we present the Itemset-Trie. Section 3 introduces the construction of the partially ordered structures topic¹. Section 4 discusses preliminary results on the practical performances of the presented algorithms. Section 5 concludes the paper and points out future directions to follow.

## 2 Itemset-Trie Data Structure

In the context of mining frequent (closed) patterns in transaction databases or many other kinds of databases, an important number of studies rely on Apriori-like "test-and-generate" approach². However, this approach suffers from a very expensive candidate set generation step, especially with long patterns or under low user-requirements. This drawback is reinforced with tediously repeated disk-stored database scans. To avoid the approach bottleneck, recent studies (e.g. the pioneering work of Han et al. and its FP-tree structure [2]) proposed to adopt an advanced data structure, where the database is compressed in order to achieve pattern mining. The idea behind the compact data structure FP-tree is that when multiple transactions share an identical frequent itemset, they can be merged into one with a registered number of occurrences. Beside a costly sorting step, the proposed FP-Tree structure is unfortunately not suited for an interactive mining process, in which a user may be interested in varying the support value. In this case, the FP-tree should be rebuilt since its construction is support dependent. Although the work presented in [3] tackles this insufficiency, the proposed structure, called Cats, in which a single item is represented in a node. That’s why we introduce a, support independent, more compact structure called Itemset-Trie, in which each node is composed by an itemset. To illustrate this compactness, let us consider the extraction context given by Figure 1(Up). Figure 1(a) depicts the associated FP-Tree, while Figure 1(b) represents the associated Itemset-Trie. Indeed, we remark that the

¹ Please note that algorithm pseudo-codes are omitted due to lack of available space.

² For a critical overview of these approaches, please refer to [6].