# Real World Association Rule Mining

Adriano Veloso, Bruno Rocha, Márcio de Carvalho, and Wagner Meira Jr.

Computer Science Department, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
{adrianov,gusmao,mlbc,meira}@dcc.ufmg.br

**Abstract** Across a wide variety of fields, data are being collected and accumulated at a dramatic pace, and therefore a new generation of techniques has been proposed to assist humans in extracting usefull information from the rapidly growing volumes of data. One of these techniques is the association rule discovery, a key data mining task which has attracted tremendous interest among data mining researchers. Due to its vast applicability, many algorithms have been developed to perform the association rule mining task. However, an immediate problem facing researchers is which of these algorithms is likely to make a good match with the database to be used in the mining operation. In this paper we consider this problem, dealing with both algorithmic and data aspects of association rule mining by performing a systematic experimental evaluation of different algorithms on different databases. We observed that each algorithm has different strengths and weaknesses depending on data characteristics. This careful analysis enables us to develop an algorithm which achieves better performance than previously proposed algorithms, specially on databases obtained from actual applications.

## 1 Introduction

The capabilities of both generating and collecting data have rapidly increased during the last years. Advances in commercial and scientific data collection have generated a flood of data. Advanced database management systems and data warehousing technology allow to gather the flood of data and to transform it into databases of an enormous size.

Due to this situation, a number of techniques with the ability to intelligently and automatically transform the processed data into knowledge have been proposed. The most widely studied of these techniques is the association rule mining, which has an elegantly simple problem statement, that is, to find the sets of all subsets of items that frequently occur as database transactions, and to infer rules showing us how a subset of items influences the presence of another subset. The prototypical application is the analysis of sales on *basket* data, but besides this example, association rules have been shown to be useful in domains such decision support, university enrollments, e-commerce, etc.

Since its origin in [1], several algorithms [2,3,4,5,6,7,8,9,10,11] have been proposed to address the association rule mining task, and we observed that, although these algorithms share basic ideas, they present different strengths and weakness depending on database characteristics. Neglecting this fact, little work has been devoted to the data aspect in the knowledge discovery process, but for real world applications practitioners have to face the problem of discovery knowledge from real databases with different characteristics. Furthermore, the extant algorithms were bench-marked on artificial

B. Eaglestone, S. North, and A. Poulovassilis (Eds.): BNCOD 2002, LNCS 2405, pp. 77–89, 2002.

datasets, but the performance improvements reported by these algorithms on artificial data do not generalize when they are applied to real datasets, and these artificial improvements can also cause problems to practitioners in real world applications.

In this paper we deal both with the algorithmic and data aspects of association rule mining. We conduct an extensive experimental characterization of these algorithms on several artificial and real databases, presenting a systematic and realistic set of results showing under which conditions an algorithm is likely to perform well and under what conditions it does not perform well. We also study the differences between the artificial datasets, generally used as benchmarks, and real datasets from actual applications (i.e., e-commerce), showing how much these differences can affect the performance of the algorithms. Furthermore, we realized that not only do there exist differences between real and artificial data, but there also exist differences between real datasets. This careful analysis enables us to develop a superior algorithm, which achieves better runtimes than the previous algorithms, especially on real datasets.

The paper is organized as follows. The problem description and related works are given in the next section. In Section 3, we systematize the most common algorithms and the strategies used by each one. Section 4 presents our new algorithm, ADARM (ADaptive Algorithm for Association Rule Mining). In Section 5 we discuss and present the analysis of the algorithms described in the preceding section. Section 6 summarizes our paper and closes with interesting topics for future work.

### 1.1 Research Contributions of this Paper

Our first contribution is a complete experimental comparison against the state-of-the-art algorithms for mining association rules. In our experiments we employed not only synthetic datasets. In fact, as opposed to the great majority of previous works, we also employed different real datasets from actual applications. Understanding how each technique behaves on different types of data is the first step of developing efficient algorithms.

The other important contribution is ADARM, the algorithm developed during this work. ADARM combines relevant features of other algorithms, which highlight the advantages of each one depending on data characteristics. These features combined together, provide to ADARM a better performance over different types of databases.

## 2 Problem Description and Related Works

The goal of association rule mining is to discover if the occurrence of certain items in a transaction can imply the occurrence of other items, or in other words, to find associative relationships between items. If indeed such interesting relationships are found, they can be put to various profitable uses such as personalization, recommendations, etc. Given a set of items, we must predict the occurrence of another set of items with a certain degree of confidence. This problem is far from trivial because of the exponential number of ways in which items can be grouped together. To state this problem we need some definitions. Let $I = \{I_1, I_2, ..., I_m\}$ be a set of $m$ attributes or items. Let $D$ be a set of transactions where each transaction is a subset of $I$, and the transaction is uniquely identified by a $tid$. Let $C$ be a subset of $I$, also called an itemset. If $C$ has $k$ items it is