Realistic Data for Testing Rule Mining Algorithms

Colin Cooper
Kings' College, UK

Michele Zito
University of Liverpool, UK

INTRODUCTION

The association rule mining (ARM) problem is a well-established topic in the field of knowledge discovery in databases. The problem addressed by ARM is to identify a set of relations (associations) in a binary valued attribute set which describe the likely coexistence of groups of attributes. To this end it is first necessary to identify sets of items that occur frequently, i.e. those subsets F of the available set of attributes I for which the *support* (the number of times F occurs in the dataset under consideration), exceeds some threshold value. Other criteria are then applied to these item-sets to generate a set of association rules, i.e. relations of the form $A \rightarrow B$, where A and B represent disjoint subsets of a frequent item-set F such that $A \cup B = F$. A vast array of algorithms and techniques has been developed to solve the ARM problem. The algorithms of Agrawal & Srikant (1994), Bajardo (1998), Brin, et al. (1997), Han *et al.* (2000), and Toivonen (1996), are only some of the best-known heuristics.

There has been recent growing interest in the class of so-called *heavy tail* statistical distributions. Distributions of this kind had been used in the past to describe word frequencies in text (Zipf, 1949), the distribution of animal species (Yule, 1925), of income (Mandelbrot, 1960), scientific citations count (Redner, 1998) and many other phenomena. They have been used recently to model various statistics of the web and other complex networks Science (Barabasi & Albert, 1999; Faloutsos *et al.*, 1999; Steyvers & Tenenbaum, 2005).

BACKGROUND

Although the ARM problem is well studied, several fundamental issues are still unsolved. In particular the evaluation and comparison of ARM algorithms is a very difficult task (Zaiane, et al., 2005), and it is often tackled by resorting to experiments carried out using data generated by the well established QUEST program from the IBM Quest Research Group (Agrawal & Srikant, 1994). The intricacy of this program makes it difficult to draw theoretical predictions on the behaviour of the various algorithms on input produced by this program. Empirical comparisons made in this way are also difficult to generalize because of the wide range of possible variation, both in the characteristics of the data (the structural characteristics of the synthetic databases generated by QUEST are governed by a dozen of interacting parameters), and in the environment in which the algorithms are being applied. It has also been noted (Brin, et al., 1997) that data sets produced using the QUEST generator might be inherently not the hardest to deal with. In fact there is evidence that suggests that the performances of some algorithms on real data are much worse than those found on synthetic data generated using QUEST (Zheng, et al., 2001).

MAIN FOCUS

The purpose of this short contribution is two-fold. First, additional arguments are provided supporting the view that real-life databases show structural properties that are very different from those of the data generated by QUEST. Second, a proposal is described for an alternative data generator that is simpler and more realistic than QUEST. The arguments are based on results described in Cooper & Zito (2007).

Heavy-Tail Distributions in Market Basket Databases

To support the claim that real market-basket databases show structural properties that are quite

different from those of the data generated by QUEST, Cooper and Zito analyzed empirically the distribution of item occurrences in four real-world retail databases widely used as test cases and publicly available from http://fimi.cs.helsinki.fi/data/. Figure 1 shows an example of such a distribution (on a log-log scale) for two of these databases. Results concerning the other two datasets are in Cooper and Zito (2007).

The authors suggest that in each case the empirical distribution may fit (over a wide range of values) a heavy-tailed distribution. Furthermore they argue that the data generated by QUEST shows quite different properties (even though it has similar size and density). When the empirical analysis mentioned above is performed on data generated by QUEST (available from the same source) the results are quite different from those obtained for real-life retail databases (see Figure 2).

Differences have been found before (Zheng *et al.*, 2001) in the transaction sizes of the real-life vs. QUEST generated databases. However some of these differences may be ironed out by a careful choice of the numerous parameters that controls the output of the QUEST generator. The results of Cooper and Zito may point to possible differences at a much deeper level.

A Closer Look at QUEST

Cooper and Zito also start a deeper theoretical investigation of the structural properties of the QUEST databases proposing a simplified version of QUEST whose mathematical properties could be effectively analyzed. As the original program, this simplified version returns two related structures: the actual database

Figure 1. Log-log plots of the real-life data sets along with the best fitting lines

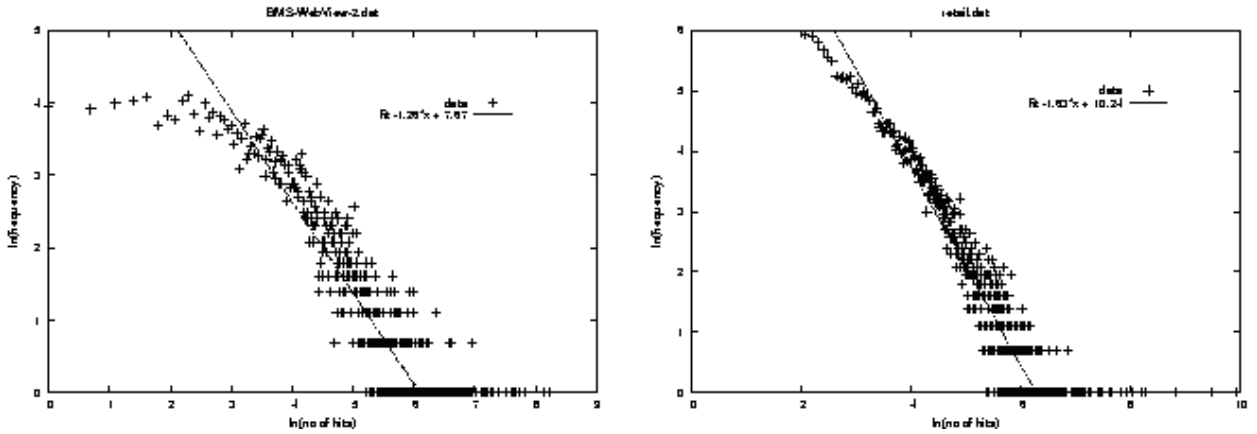

Figure 2. Log-log plots of the QUEST data sets along with the best fitting line