Quest: A Project on Database Mining

R. Agrawal M. Carey C. Faloutsos S. Ghosh M. Houtsma
T. Imielinski B. Iyer A. Mahboob H. Miranda R. Srikant A. Swami

IBM Almaden Research Center
San Jose, CA 95120

1 Background

Several organizations have collected massive amounts of data. These data sets are usually stored on tertiary storage and are very slowly migrating to database systems. One of the reasons for the limited success of database systems in this area is that current database systems do not provide the necessary functionality for a user interested in taking advantage of this information.

Database mining refers to the efficient construction and verification of models of patterns embedded in large databases, and is emerging as a major application area for databases. The goal of the Quest project is to enhance database technology to address this problem.

2 Prototype

In the demonstration at Sigmod94, we will show some of the database mining technology that we have developed. In particular, we will demonstrate mining of association rules over sales data captured by retail organizations, such as department stores, supermarkets and catalog companies. An example of such a rule is that 98% of customers that purchase tires and auto accessories also get automotive services done.

The interesting aspect of our software is that it is *not* verification driven, which is the current state of art in industry. We ask the user to simply provide two input parameters driven by business considerations: i) minimum confidence, and ii) minimum support. 98% was the confidence in the above example, and it indicates the fraction of cases in which when the antecedent holds, the consequent of the rule also holds. The support of a rule is the fraction of total transactions in which the rule holds. By specifying minimum confidence and support, the user is asking for all the rules that have confidence above minimum confidence and that are present at least in minimum support fraction of transactions. We do not then require any more human intervention, and we generate all the rules that satisfy these constraints.

We will also demonstrate mining of sequential patterns in sales transactions. That is, we will show how to find what items customers buy over a series of visits in sequence (e.g. an order for sheets and pillowcases, followed by a comforter, followed by ruffles and shams). Again, we only require the user to specify minimum support, i.e. the minimum fraction of customer transaction sequences in which the pattern is required to be present. Our software then finds all sequential patterns that have minimum support. Note that a term of the sequence can have more than one item (e.g. sheets and pillowcases) and an item (or set of items) can appear multiple times.

References

[1] R. Agrawal, T. Imielinski, and A. Swami, "Database Mining: A Performance Perspective", *IEEE Transactions on Knowledge and Data Engineering*, Special Issue on Learning and Discovery in Knowledge-Based Databases, Dec. 1993.

[2] R. Agrawal, S. Ghosh, T. Imielinski, B. Iyer, and A. Swami, "An Interval Classifier for Database Mining Applications", *VLDB-92*, Vancouver, British Columbia, Canada, 1992, 560–573.

[3] R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", *SIGMOD-93*, Washington D.C., May 1993.

[4] R. Agrawal, C. Faloutsos, A. Swami: "Efficient Similarity Search in Sequence Databases", *4th Int'l Conf. on Foundations of Data Organization and Algorithms*, Chicago, Oct. 1993.