Towards Data Mining Benchmarking: A Test Bed for Performance Study of Frequent Pattern Mining

Jian Pei and Runying Mao
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada
{peijian, rmao}@cs.sfu.ca

Kan Hu and Hua Zhu
DBMiner Technology Inc.
Burnaby, BC, Canada
{kanhu, hzhua}@dbminer.com

Performance benchmarking has played an important role in the research and development in relational DBMS, object-relational DBMS, data warehouse systems, etc. We believe that benchmarking data mining algorithms is a long overdue task, and it will play an important role in the research and development of data mining systems as well.

Frequent pattern mining forms a core component in mining associations, correlations, sequential patterns, partial periodicity, etc., which are of great potential value in applications. There have been a lot of methods proposed and developed for efficient frequent pattern mining in various kinds of databases, including transaction databases, time-series databases, etc. However, so far there is no serious performance benchmarking study of different frequent pattern mining methods.

To facilitate an analytical comparison of different frequent mining methods, we have constructed an open test bed for performance study of a set of recently developed, popularly used methods for mining frequent patterns in transaction databases and mining sequential patterns in sequence databases, with different data characteristics. The testbed consists of the following components.

- A synthetic data generator, which can generate large sets of synthetic data in various kinds of data distributions. A few large data sets from real world applications will also be provided.

- A good set of typical frequent pattern mining methods, ranging from classical algorithms to recent studies. The method are grouped into three classes: *frequent pattern mining, max-pattern mining*, and *sequential pattern mining*. For frequent pattern mining, we will demonstrate Apriori, hashing, partitioning, sampling, TreeProjection, and FP-growth. For maximal pattern mining, we will demonstrate MaxMiner, TreeProjection, and FP-growth-max. For sequential pattern mining, we will demonstrate GSP and FreeSpan.

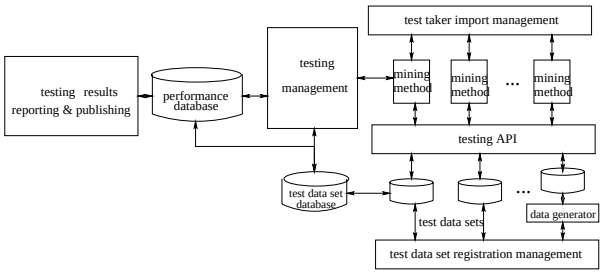

Figure 1: The architecture of the open test bed.

- A set of performance curves. These algorithms, their running speeds, scalabilities, bottlenecks, and performance on different data distributions, will be compared and demonstrated upon request. Some performance curves from our pre-conference experimental evaluations will also be shown.

- An open testbed. Our goal is to construct an extensible test bed which integrates the above components and supports an open-ended testing service. Researchers can upload the object codes of their mining algorithms, and run them in the test bed using these data sets. The architecture is shown in Figure 1.

This testbed is our first step towards benchmarking data mining algorithms. By doing so, performance of different algorithms can be reported consistently, on the same platform, and in the same environment. After the demo, we plan to make the testbed available on the WWW so that it may, hopefully, benefit further research and development of efficient data mining methods.

MOD 2000, Dallas, TX USA