Fifth International Workshop on Data Mining Standards, Services, and Platforms

Robert L. Grossman (Workshop Chair)
University of Illinois at Chicago
Chicago, IL
and Open Data Group
River Forest, IL
grossman@uic.edu

Abstract

This is the Fifth DM-SSP Workshop on Data Mining Standards, Services and Platforms. This year the workshop will focus on scaling data mining to data sets that are so large that they do not fit into memory and cannot be managed effectively by a database.

Categories and Subject Descriptors: C.2.4 [Distributed Systems]: Distributed applications - grid-enabled applications; D.2.11 [Software Architectures]: Domain specific architectures - service oriented architectures.

General Terms: Performance, Design, Measurement

Keywords: data mining systems, high performance data, mining, data mining standards, data mining grids

1 Large and Distant Data

The DM-SSP Workshops are a forum for those developing data mining systems and for those integrating data mining systems and applications into operational systems.

The 2007 DM-SSP Workshop is the fifth such workshop. The focus of the 2007 Workshop is on data mining systems that are designed to mine very large, and possibly distributed, data sets.

A simple but useful distinction is to divide data mining systems into three types: those designed to work with small, medium and large size data sets.

- **Small Data.** Most data mining systems are designed to work with data small enough to fit into the memory of a single computer.
- **Medium Data.** Some data mining systems are integrated with databases and are designed to work with data that is carefully managed by them.
- **Large Data.** Databases provide a lot of sophisticated functionality, but eventually run into scaling problems, as the size of the data grows. If data is too large to be managed by a database, then specialized systems designed to work with file based data can be used.

From this perspective, this year's workshop is focused on data mining systems for mining large data sets.

In practice, the situation is more complicated — systems designed to work with data in memory can always import data from a database or a file system, and systems designed to work with data in a database can always import data from a file system. On the other hand, in this case, there are always performance issues that arise when importing data in this way that generally do not arise for systems designed specifically to work with data in a database or in a (specialized) file system.

Another dimension of complexity for those designing data mining systems is whether the data: 1) is in one computer and its associated storage; 2) is in a cluster of machines co-located in a single location; or, 3) is in computers distributed over two or more (geographically) distributed locations. Think of this as local, near and distant data.

2 System Architectures

Data mining system designed to work with large data sets generally provide most of the following functions and capabilities.

- **Data access.** The data access layer accesses file based data and manages the data records and their attributes.
- **Resource management and monitoring.** The next layer acquires, manages and monitors the physical resources (computers, storage, etc.) required for the analysis.
- **Data mining components and services.** Components and services include those for preparing data, building statistical and data mining models, and evaluating the models produced.
- **Workflow.** For almost all projects, some type of workflow is required that links together several data mining components and services.

3 DM-SSP 07

The DM-SSP 07 Workshop contains invited and contributed presentations about several different systems designed to work with file based data that is too large to fit into memory or into a single database.

DM-SSP'07, August 12, 2007, San Jose, California, USA.