Data Mining Standards, Services, and Platforms 2004 (DM-SSP 2004)

Workshop Report

Robert Grossman
University of Illinois at Chicago
Chicago, Illinois
and Open Data Partners
Oak Park, IL
grossman@uic.edu

ABSTRACT

This is a summary of the workshop on Data Mining Standards, Services and Platforms, which was held at KDD 2004. The Workshop contained talks on the Predictive Model Markup Language (PMML) and the Java standard for data mining JSR-73. It also contained talks on emerging service-based architectures for data mining. Finally, infrastructures for privacy preserving data mining were also discussed.

1. INTRODUCTION

KDD 2004 was the fourth year that there has been a KDD workshop on the Predictive Model Markup Language (PMML) and related areas and the second year of a broader conference with the theme of Data Mining Standards, Services and Platforms.

Data mining and knowledge discovery has matured rapidly since the first KDD conferences in the late 1980s, drawing from the related fields of statistics, machine learning, databases, and high performance computing. Recall the fundamental role played by the relational database model [1] and the SQL standard in the development of databases, and the similar role played by the message passing model and the MPI standard in high performance computing [2]. There is not yet a similar model or standard in data mining. The role of this workshop is to move closer to such a model and supporting standards.

The online proceedings can be found at:

`www.ncdm.uic.edu/workshops/dm-ssp04.htm.`

An invited talk was given by Jamie MacLennan of Microsoft with the title Vectors on Data Mining: How standards and platforms will impact the near future of Data Mining.

2. PMML PRODUCERS AND CONSUMERS

Part of this role is played by the Predictive Model Markup Standard or PMML developed by the Data Mining Group [3] and [4]. One of the goals of PMML is to create a standard interface between producers of models, such as statistical or data mining systems, and consumer of models, such as scoring engines, applications containing embedded models, and other operational systems. There are now quite a few vendors shipping scoring engines, which is an important measure of success in this area.

Distinguishing between PMML producers and consumers is important for at least a couple of reasons. First, PMML producers and PMML consumers in general have quite different requirements. For example, a statistician might use a data mining system and work for several months to produce a PMML model. On the other hand, a PMML consumer might be a light weight scoring engine embedded in an operational system and used for real time scoring. Second, an XML interface between PMML producers and consumers provides a simple mechanism for cleanly separating the modeling phase in the data mining process from the deployment phase.

For the same reason, it is useful to have a clean and well defined interface between the data preparation phase and the modeling phase of the data mining process. To this end, for the past several years, the developers of PMML have been working to define common transformations, and other operations, such as compositions, that are required for data preparation. This is one of the themes of this year's workshop.

As a standard architecture for scoring and a standard architecture for data preparation emerges, we are one step closer to a standard infrastructure for data mining.

PMML Version 3.0 was released in October of 2004, two months after the workshop. An overview of PMML Version 3.0 was given by Stefan Raspl of IBM.

PMML Version 3.0 introduces a mechanism for composing common data mining operations. This was described in the talk: A Simple Strategy for Composing Data Mining Operations by Gregor Meyer of IBM and by Robert Grossman of the University of Illinois at Chicago and Open Data Partners.

Bill Hosken and Bernard Scherer of SPSS presented a customer case study of a high performance PMML-based scoring engine in a talk titled: Distributed Scoring Using PMML.

3. SERVICE BASED ARCHITECTURES

Data mining started as a stand alone application; more recently data mining has been embedded in databases and distributed Java-based architectures have been developed.

Another theme in the DM-SSP 2004 Workshop is the maturation of service-based architectures for data mining to complement XML-based producer/consumer models.
Robert Chu of SAS gave an introduction to web service based approaches to data mining in a talk called: Web Services Standards for Data Mining.
Data mining has never limited itself to just small data sets. On the other hand scaling web services to working with large remote and distributed data sets faces several technical challenges. Some of these were described, along with some possible solutions, in a talk by David Hanley of the University of Illinois at Chicago and Robert Grossman of the University of Illinois at Chicago and Open Data Partners with the title: Experimental Studies Scaling Web Services For Data Mining.

4. OTHER THEMES

PMML has served as a neutral common ground between various other data mining standards, such as Java-based standards and SQL-based standards. Mark F. Hornick of Oracle gave a talk describing the recently released Java standard for data mining (JSR-73) with the title: Java Data Mining (JSR-73): Status and Overview.
Finally, the importance of privacy preserving data mining has grown enormously during the past few years. A fourth theme in the workshop was the beginning the process to develop standards in this area. Stanley R. M. Oliveira and Osmar R. Zaiane of the University of Alberta gave a talk about this with the title: Toward Standardization in Privacy-Preserving Data Mining.

5. REFERENCES

[1] E. F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, 1970.

[2] W. Gropp, E. Lusk, and A. Skjellum. *Using MPI: Portable Parallel Programming with the Message Passing Interface, 2nd Edition*. MIT Press, 1999.

[3] R. L. Grossman, S. Bailey, A. Ramu, B. Malhi, P. Hallstrom, I. Pulleyn, and X. Qin. The management and mining of multiple predictive models using the predictive model markup language (pmml). *Information and Software Technology*, 41:589–595, 1999.

[4] R. L. Grossman, M. Hornick, and G. Mayer. Data mining standards initiatives. *Communications of the ACM*, 45(8):59–61, 2002.