KDD Workshop on Data Mining Standards, Services & Platforms (DM-SSP) 2006

Robert L. Grossman
Open Data Group
400 Lathrop Ave Suite 90
River Forest IL 60305
and National Center for Data Mining
University of Illinois at Chicago
851 S. Morgan Street (MC 249)
Chicago IL 60607
grossman@uic.edu

ABSTRACT

DM-SSP '06 was the fourth year there has been a conference on data mining standards, services and platforms and the sixth year that there has been a conference on the Predictive Model Markup Language (PMML) and related areas. The workshop consisted of five talks and two panels.

1. INTRODUCTION

DM-SSP '06 was the fourth year there has been a conference on data mining standards, services and platforms and the sixth year that there has been a conference on the Predictive Model Markup Language (PMML) and related areas.

The goal of the DM-SSP workshops is to bring together developers who are interested standards based frameworks for data mining systems and applications that a) prepare data for data mining, b) produce data mining models, or c) employ data mining models.

Today, there are a variety of standards, proposed standards, and specifications that are used in data mining, including PMML [1], SQL-based standards such as SQL/MM Data Mining [3], Java standards such as JSR 73: Data Mining API [2], and Microsoft's XML for Analysis (XMLA) web service specification [4].

There are also two other movements in the broader community of developers that are of interest to those developing data mining systems and applications. First, service oriented architectures are becoming more common, and this is beginning to impact how data mining is deployed. Second, there is an active community developing standards for grids, and some of these standards are relevant to data mining.

This year's workshop brought together developers, practitioners, and researchers from these various groups to discuss standards based systems, applications, and platforms for data mining.

2. WORKSHOP OVERVIEW

Historically, data mining has been mainly concerned with data that is static and at rest. Over the past few years, attention has also focused to data that is streaming. The talk "SPC: A Distributed, Scalable Platform for Data Mining" by First, Lisa Amini, Henrique Andrade, Ranjita Bhagwan, Frank Eskesen, Richard King, Philippe Selo, Yoonho Park, and Chitra Venkatramani discussed a stream based data mining system and the architecture emerging for these types of systems.

Because of the way many applications are deployed in operational systems or integrated with business processes, it is often useful to use one system for building or estimating statistical and data mining models (*model producers*) and another system for producing scores with these models using operational data (*model consumers*).

The Predictive Model Markup Language (PMML) is an XML standard for data mining models, as well as for many of the common operations required for preparing data for data mining, and is often used as an interchange format between models producers and consumers.

Over the past several years, PMML and related data mining standards have matured to the point that conformance of applications to PMML and interoperability have become relevant for many users and for many applications. The talk "Conformance Standard for the Predictive Model Markup Language" by Rick Pechter addressed this issue. This was followed by demonstrations of interoperability between various PMML consumers and producers from different vendors.

The paper "Augustus: The Design and Architecture of a PMML-Based Scoring Engine" by John Chaves, Chris Curry, Robert L. Grossman, David Locke and Steve Vejcik discussed an open source PMML-based scoring engine called Augustus and an associated PMML-repository that has been used in applications employing thousands to millions of separate PMML models.

Many applications can benefit by employing standards for specific types of data, such as text, images, and other unstructured information, as well as domain specific standards. The paper "PMML and UIMA Based Frameworks for Deploying Analytic Applications and Services" by David Ferrucci, Robert L. Grossman and Anthony Levas discussed how PMML could be integrated with the open UIMA framework for mining unstructured information, such as text, images, etc. The paper "MatML: XML for Information Exchange with Materials Property Data" by Aparna S. Varde, Edwin F. Begley, and Sally Fahrenholz-Mann discussed a markup language for materials called MatML and analytic applications built using it. MatML is useful for exchanging

materials property information.

3. SOME SUCCESS STORIES

Early data mining systems were generally closed in the sense that data was loaded into the system and reports were produced. To deploy models into operational systems required hand coding the required models and data preparation code. This approach for deploying models was labor intensive to say the least.

Later, some data mining systems began to support an export mechanism allowing models to be exported into SQL, C++ or Java. For some applications, it was helpful to have a format for data mining models that was independent of the language and of the application — PMML was introduced to to fulfill this requirement.

As a sign of the maturity of data mining standards, here are some examples from the workshop of how data mining is being deployed today:

- PMML models can be exported from SPSS or SAS and loaded into a MicroStrategy data warehouse and used to score data from the warehouse and to evaluate the results of scoring the data.
- A PMML model can be exported from a data mining application and loaded into a database such as IBM's DB2 so that data in the database can be scored using a data mining model from a third party application.
- Java-based enterprise applications incorporating data mining can be built that access JSR-73 compliant components.
- A PMML model can be loaded into a scoring engine that makes scoring available as a web service. Using this web service, a web mash up can be put together quickly and easily that includes data mining.
- Models repositories containing PMML models can be built to provide ready access to models, for compliance purposes, or when model providence is important.

4. RESEARCH OPPORTUNITIES

In this section, we discuss some research problems for standards based data mining platforms that were identified in the workshop and related discussions.

We begin with three broad challenges that have been actively discussed for the past several workshops. This year was no exception.

- How can standards that describe the data preparation required for data mining be improved?
- How can standards that describe the full environment required for deploying data mining models be improved?
- What is an appropriate workflow standard to integrate with PMML?

In addition, this year also saw active discussion of the following three research challenges related to standards based platforms for data mining:

- Develop standards based platforms and frameworks that integrate domain specific knowledge.
- Adapt and extend current data mining standards to apply to streaming and distributed data.
- Develop standards based platforms that integrate data mining services and grid services.

5. PAPERS PRESENTED

During the morning of the workshop, the following papers were presented:

- Rick Pechter, Conformance Standard for the Predictive Model Markup Language.
- Lisa Amini, Henrique Andrade, Ranjita Bhagwan, Frank Eskesen, Richard King, Philippe Selo, Yoonho Park, Chitra Venkatramani, SPC: A Distributed, Scalable Platform for Data Mining.
- David Ferrucci, Robert L. Grossman and Anthony Levas, PMML and UIMA Based Frameworks for Deploying Analytic Applications and Services.
- John Chaves, Chris Curry, Robert L. Grossman, David Locke and Steve Vejcik, Augustus: The Design and Architecture of a PMML-Based Scoring Engine.
- Aparna S. Varde, Edwin F. Begley, and Sally Fahrenholz-Mann, MatML: XML for Information Exchange with Materials Property Data.

During the afternoon of the workshop, there were several discussions regarding the upcoming release of PMML Version 3.2 (expected in the fourth quarter of 2006) and user requirements for future releases of PMML.

6. REFERENCES

[1] Data Mining Group. Predictive model markup language. *retrieved from www.dmg.org*, 2006.

[2] M. Hornick and et. a. Java specification request (jJSR) 73: Data mining api. *retrieved from jcp.org/en/jsr/detail?id=73*, 2006.

[3] J. Melton and A. Eisenberg. SQL multimedia and applications packages (SQL/MM). *ACM SIGMOD Record*, 30, 2001.

[4] Microsoft. XML for analysis. *retrieved from www.sqlserverdatamining.com*, 2006.