Open Source Analytics: An Introduction to the Special Issue

Robert L Grossman
University of Illinois at Chicago
and Open Data Group
grossman@uic.edu

ABSTRACT

This special issue contains six articles on open source analytics. It includes an article describing the Weka data mining system, two articles on infrastructure to support analytics, an article on the PMML standard for statistical and data mining models, an article on how clouds are being used in analytics, and an article about an open source tool for cleaning data.

1. FIVE TRENDS IN OPEN SOURCE ANALYTICS

We introduce the articles in this special issue by discussing five trends in open source analytics.

Trend 1. The first trend is that open source software for statistics and data mining has matured and reached a critical mass. R (www.r-project.org), which was first released in 1996, is now the most widely deployed open source system for statistical computing. A recent article in the New York Times estimated that over 250,000 individuals use R regularly [4].

Another very popular open source system for data mining is the Weka system. Weka is described in an article in this issue by Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahrin, Peter Reutemann and Ian H. Witten called the "The WEKA Data Mining Software: An Update."

Trend 2. The second trend is that standards for statistics and data mining have matured sufficiently that they can can now serve as a foundation for open, standards based architectures for analytic applications.

The Predictive Model Markup Language or PMML [1] is an XML language for describing statistical and data mining models in an application and system independent fashion. With PMML, models can be exchanged easily between systems and persisted to repositories.

This issue contains an article about PMML by Rick Pechter called "What's PMML and What's New in PMML 4.0?"

Trend 3. It is difficult to analyze datasets without good tools for managing them. Databases have been used since they were introduced to help manage data for data mining. More recently, more specialized tools, such as Hadoop [2] have been developed. Hadoop is an open source system that has dramatically simplified the management and analysis of large datasets. The term *analytic infrastructure* is beginning to be used to describe the services, applications, utilities and systems that are used for either preparing data for modeling, estimating models, validating models, scoring data, or related activities.

This issue contains an article titled "KNIME — The Konstanz Information Miner" by Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Kilian Thiel and Bernd Wiswedel that describes a system which enables the visual assembly and interactive execution of a data analysis pipeline.

There is also an article that I wrote giving a brief introduction to analytic infrastructure.

Trend 4. A particularly important type of analytic infrastructure are clouds. The fourth trend is the growing importance of cloud-based services for analytics. Although there is no standard definition of clouds, clouds can be defined as on-demand services and resources available over a network, often the Internet, that are offered with the scale and reliablity of a data center. A recent publication by NIST [3] defines clouds and covers some of their main characteristics.

This issue contains an article by Alex Guazzelli, Kostantinos Stathatos and Michael Zeller about predictive analytics and cloud computing called "Efficient Deployment of Predictive Analytics through Open Standards and Cloud Computing."'

Trend 5. The final trend is the commoditization of data. Moore's law applies not only to CPUs, but also to the chips that are used in all of the digital devices that produce data. The result has been that the cost to produce data has been falling for some time. This, combined with the fact that the cost to store data has also been falling at the speed of Moore's law, is resulting in the increasing availability of datasets.

The availability of all this data, sometimes without charge, is enriching analytics. With the growing amount of data, data quality issues are critical.

The final article in this issue is an article by Peter Christen about an open source tool for cleaning data called "Development and User Experiences of an Open Source Data Cleaning, Deduplication and Record Linkage System."

2. REFERENCES

[1] Data Mining Group. Predictive Model Markup Language (pmml), version 4.0. http://www.dmg.org, 2009.

[2] Hadoop. Welcome to Hadoop! hadoop.apache.org/core/, 2009.

[3] Peter Mell and Tim Grance. Draft nist working definition of cloud computing. www.nist.gov, June 1, 2009.

[4] Ashlee Vance. Data analysts captivated by R's power. *New York Times*, January 6, 2009.