Making Sense of your Data

Navigating the Sea of Data

Edmund W. Schuster

The Data Center, Massachusetts Institute of Technology, Building 35, Room 234, Cambridge, MA 02139-4307, USA

At Smart World 2004, Sunil Gupta of SAP paraphrased Samuel Taylor Coleridge in his opening statement by stating "data, data everywhere but not a bye to use." The problem of what to do with all the data gathered within organizations is becoming a critical management issue.

Forbes recently reported an interesting statistic about data [1]. For 2004, shipments of storage devices alone will equal four times the space needed to store every word ever spoken during the entire course of human history.

The explosion of storage device sales has gone largely unnoticed because the selling price has decreased sharply during the past few years. While the increase in unit sales has meant larger revenues for companies like EMC, the amount of the increase is not great enough to draw widespread attention to the rapid increase of data within organizations.

Businesses are truly undergoing a revolution in terms of the data available for decision-making. This is leading to new ideas about making the best use of vast quantities of data that involve the minute details of everyday operations. Emerging technologies such as Auto-ID, sensor networks, and the use of "loyalty cards" will add to the burden of handling and analyzing data.

For example, Harrah's Entertainment the world's largest casino operator gathers data each day for 45,000 slot machines through loyalty cards used by customers. Each spin of the slot machine is recorded. Harrah's uses this vast data stream to make detailed decisions about what customers should receive various awards such as free show tickets, dinner vouchers or room upgrades. This type of micro decision-making is a powerful tool to build brand loyalty [1].

Harrah's has plans to improve the integration between models and data so that the decision to reward select customers will happen just minutes after data collection about consumer behavior. This type of immediate feedback not only builds brand loyalty, but also provides competitive advantage.

Longer term, Harrah's uses its storehouse of data to make critical decisions about what machines to purchase and the optimum layout of a casino. All of these efforts are possible if the data is available.

Increasingly, companies from many types of industries want to look at as much data as possible, especially data associated with consumer behavior. Gaining specific insight about what motivates the consumer is already a top goal for many large major companies.

Though advances in storage devices have made it possible to amass large quantities of data, the efficient analysis of this data is still an open frontier. The only way to make sense of large quantities of data is through the application of mathematical models. Historically, the process of modeling has been slow. Managers often complain that models are complex, requiring specialists with many years of experience for model development and operation. Modeling can often lead to powerful insights, however, all too often models are developed for a single application and there is very little re-use or sharing within or outside of organizations.

Compounding matters, many types of data take a non-structured form making modeling on a large scale difficult. Unstructured data includes images, emails, and engineering designs [1]. In all of these cases, object representation requires more than a serial number that can be neatly stored in a database. It is only through a standard semantic that these types of objects can be accurately described for search and analytical purposes.

Semantics offer the opportunity for richer descriptions of unstructured data, allowing for this type of data to be matched to models quickly. One example involves the automotive industry. With over 100 years of experience, there are thousands of designs for various components of an automobile. Most of these designs are stored in CAD/CAM system databases. Organizing and searching these designs becomes problematic because a precise semantic does not exist for each design. This makes it impossible to combine designs in innovative ways.

Even though semantic methods to describe models and data represent potentially powerful techniques, the development of the computer languages and protocols to facilitate a true semantic based system are only in the early stages. With many years of combined experience in computer science and modeling, *The MIT Data Center* has an active research agenda to build an interoperable modeling system that will offer a structure for assignment of precise semantics to abstract objects like models and data. If successful, this effort will create an entirely new way to use the Internet as a tool for the analysis of data. With the vast quantities of data expected in the future, semantic based computer languages, protocols, and tools represent an interesting area for growth.

[1] Lyons, Daniel (2004), "Too Much Data," December 13.

NOTES

