D

Data Mining and Decision Support for Business and Science

Auroop R. Ganguly
Oak Ridge National Laboratory, USA

Amar Gupta
University of Arizona, USA

Shiraj Khan
University of South Florida, USA

INTRODUCTION

Analytical Information Technologies

Information by itself is no longer perceived as an asset. Billions of business transactions are recorded in enterprise-scale data warehouses every day. Acquisition, storage, and management of business information are commonplace and often automated. Recent advances in remote or other sensor technologies have led to the development of scientific data repositories. Database technologies, ranging from relational systems to extensions like spatial, temporal, time series, text, or media, as well as specialized tools like geographical information systems (GIS) or online analytical processing (OLAP), have transformed the design of enterprise-scale business or large scientific applications. The question increasingly faced by the scientific or business decision-maker is not how one can get more information or design better information systems but what to make of the information and systems already in place. The challenge is to be able to utilize the available information, to gain a better understanding of the past, and to predict or influence the future through better decision making. Researchers in data mining technologies (DMT) and decision support systems (DSS) are responding to this challenge. Broadly defined, data mining (DM) relies on scalable statistics, artificial intelligence, machine learning, or knowledge discovery in databases (KDD). DSS utilize available information and DMT to provide a decision-making tool usually relying on human-computer interaction. Together, DMT and DSS represent the spectrum of analytical information technologies (AIT) and provide a unifying platform for an optimal combination of data dictated and human-driven analytics.

Table 1. Analytical information technologies

Data-Mining Technologies
• Association, correlation, clustering, classification, regression, database knowledge discovery
• Signal and image processing, nonlinear systems analysis, time series and spatial statistics, time and frequency domain analysis
• Expert systems, case-based reasoning, system dynamics
• Econometrics, management science
Decision Support Systems
• Automated analysis and modeling
o Operations research
o Data assimilation, estimation, and tracking
• Human computer interaction
o Multidimensional OLAP and spreadsheets
o Allocation and consolidation engine, alerts
o Business workflows and data sharing

Table 2. Application examples

Science and Engineering
• Bio-Informatics
• Genomics
• Hydrology, Hydrometeorology
• Weather Prediction
• Climate Change Science
• Remote Sensing
• Smart Infrastructures
• Sensor Technologies
• Land Use, Urban Planning
• Materials Science
Business and Economics
• Financial Planning
• Risk Analysis
• Supply Chain Planning
• Marketing Plans
• Text and Video Mining
• Handwriting/Speech Recognition
• Image and Pattern Recognition
• Long-Range Economic Planning
• Homeland Security

BACKGROUND

Tables 1 and 2 describe the state of the art in DMT and DSS for science and business, and provide examples of their applications.

Researchers and practitioners have reviewed the state of the art in analytic technologies for business (Apte et al., 2002; Kohavi et al., 2002; Linden & Fenn, 2003) or science (Han et al., 2002), as well as data mining methods, software, and standards (Fayyad & Uthurusamy, 2002; Ganguly, 2002a; Grossman et al., 2002; Hand et al., 2001; Smyth et al., 2002) and decision support systems (Carlsson & Turban, 2002; Shim et al., 2002).

MAIN THRUST

Scientific and Business Applications

Rapid advances in information and sensor technologies (IT and ST) along with the availability of large-scale scientific and business data repositories or database management technologies, combined with breakthroughs in computing technologies, computational methods, and processing speeds, have opened the floodgates to data-dictated models and pattern matching (Fayyad & Uthurusamy, 2002; Hand et al., 2001). The use of sophisticated and computationally-intensive analytical methods is expected to become even more commonplace with recent research breakthroughs in computational methods and their commercialization by leading vendors (Bradley et al., 2002; Grossman et al., 2002; Smyth et al., 2002).

Scientists and engineers have developed innovative methodologies for extracting correlations and associations, dimensionality reduction, clustering or classification, regression, and predictive modeling tools based on expert systems and case-based reasoning, as well as decision support systems for batch or real-time analysis. They have utilized tools from areas like traditional statistics, signal processing, and artificial intelligence, as well as emerging fields like data mining, machine learning, operations research, systems analysis, and nonlinear dynamics. Innovative models and newly discovered patterns in complex, nonlinear, and stochastic systems encompassing the natural and human environments have demonstrated the effectiveness of these approaches. However, applications that can utilize these tools in the context of scientific databases in a scalable fashion have only begun to emerge (Curtarolo et al., 2003; Ganguly, 2002b; Grossman & Mazzucco, 2002; Grossman et al., 2001; Han et al., 2002; Kamath et al., 2002; Thompson et al., 2002).

Business solution providers and IT vendors, on the other hand, have focused primarily on scalability, process automation and workflows, and the ability to combine results from relatively simple analytics with judgments from human experts. For example, e-business applications in the areas of supply-chain planning, financial analysis, and business forecasting traditionally rely on decision-support systems with embedded data mining, operations research and OLAP technologies, business intelligence (BI), and reporting tools, as well as an easy-to-use GUI (graphical user interface) and extensible business workflows (Geoffrion & Krishnan, 2003). These applications can be custom built by utilizing software tools or are available as prepackaged e-business application suites from large vendors like SAP®, PeopleSoft®, and Oracle®, as well as best of breed and specialized applications from smaller vendors like Seibel® and i2®. A recent report by the market research firm Gartner (Linden & Fenn, 2003) summarizes the relative maturity and current industry perception of advanced analytics. For reasons ranging from excessive (IT) vendor hype to misperceptions among end users caused by inadequate quantitative background, the business community is barely beginning to realize the value of data-dictated predictive, analytical, or simulation models. However, there are notable exceptions to this trend (Agosta et al., 2003; Apte et al., 2002; Geoffrion & Krishnan, 2003; Kohavi et al., 2002; Wang & Jain, 2003; Yurkiewicz, 2003).

Solutions Utilizing DMT and DSS

For a scientist or an engineer, as well as for a business manager or management scientist, DMT and DSS are tools used for developing domain-specific applications. These applications might combine knowledge about the specific scientific or business domain (e.g., through the use of physically-based or conceptual-scientific models, business best practices and known constraints, etc.) with data-dictated or decision-making tools like DSS and DMT. Within the context of these applications, DMT and DSS can aid in the discovery of novel patterns, development of predictive or descriptive models, mitigation of natural or manmade hazards, preservation of civil societies and infrastructures, improvement in the quality and span of life as well as in economic prosperity and well being, and development of natural and built environments in a sustainable fashion. Disparate applications utilizing DMT and DSS tools tend to have interesting similarities. Examples of current best practices in the context of business and scientific applications are provided next.

Business forecasting, planning, and decision support applications (Carlsson & Turban, 2002; Shim et al.,