ENCYCLOPEDIA OF DATA WAREHOUSING AND MINING

Data Mining Technologies and Decision Support Systems for Business and Scientific Applications

Auroop R Ganguly* and Amar Gupta**

*Information Technologies & Quantitative Methods
Civil & Environmental Engineering
University of South Florida, Tampa, FL 33613
Email: auroop@alum.mit.edu

**MIT Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139
Email: agupta@mit.edu

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. The 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 on-line 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 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.

Tables 1 and 2 describe the state of the art in data mining technologies (or, DMT) and decision support systems (DSS) for science and business, and examples of their applications.

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
    - Operations Research
    - Data Assimilation, Estimation and Tracking
- Human Computer Interaction
    - Multidimensional OLAP and spreadsheets
    - Allocation and consolidation engine, alerts
    - 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

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

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 and Uthurusamy, 2002; Hand et al., 2001). The use of sophisticated and computationally intensive analytical methods are expected to become even more commonplace with recent research breakthroughs in computational methods and their commercialization by leading vendors (Smyth et al., 2002; Bradley et al., 2002; Grossman 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 ìtraditionalî 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 (e.g., Grossman et al., 2001; Ganguly, 2002b; Han et al., 2002; Kamath et al., 2002; Grossman and Mazzucco, 2002; Thompson et al., 2002; Curtarolo et al., 2003).

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, ìe-business applicationsî in the areas of supply chain planning, financial analysis and business forecasting, traditionally rely on decision support systems with embedded ìdata miningî, 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 (e.g., see Geoffrion and Krishnan, 2003). These applications can be custom built by utilizing software tools, or available as prepackaged ìe-business application suitesî from large vendors like SAP^{Æ}, PeopleSoft^{Æ} and Oracle^{Æ} as well as ìbest of breedî and specialized applications from smaller vendors like Seibel^{Æ} and i2^{Æ}. A recent report by the market research firm Gartner (Linden and Fenn, 2003) summarizes the relative maturity and current industry perception of ìadvanced analyticsî. 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 (e.g., Agosta et al., 2003; Geoffrion and Krishnan, 2003; Apte et al., 2002; Kohavi et al., 2002; Wang and 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 on 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 man-made 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 (e.g., see Shim et al., 2002; Carlsson and Turban, 2002; Wang and Jain, 2003; Yurkiewicz, 2003) usually need to read data from a variety of sources like on-line transactional processing (OLTP) systems, historical data warehouses and data marts, syndicated data vendors, legacy systems, or public domain sources like the Internet, as well as in the form of real-time or incremental data entry from external or internal collaborators, expert consultants, planners, decision makers and/or executives. Data from disparate sources are usually mapped to a predefined common data model, and incorporated through extraction, transformation and loading (ETL) tools. End users are provided GUI based access to define application contexts and settings, structure business workflows and planning cycles and format data models for visualization, judgmental updates or analytical and predictive modeling. The parameters of the embedded data mining models might be preset, calculated dynamically based on data or user inputs, or specified by a power user. The results of the data mining models can be automatically utilized for optimization and recommendation systems and/or can be used to serve as baselines for planners and decision makers. Tools like BI, Reports and OLAP (e.g., Hammer, 2003) are utilized to help planners and decision makers visualize key metrics and predictive modeling results, as well as utilize alert mechanisms and selection tools to manage by exception or by objectives. Judgmental updates at various levels of aggregation and their reconciliation, collaboration among internal experts and external trading partners as well as managerial review processes and adherence to corporate directives are aided by allocation and consolidation engines, tools for simulation, ad hoc and predefined reports, user defined business workflows, audit trails with comments and reason codes, and flexible information transfer and data handling capabilities. Emerging technologies include the use of automated DMT for aiding traditional DSS tasks ñ for example, the use of data mining to zero down on the cause of aggregate exceptions in multidimensional OLAP ìcubesî. The end results of the planning process are usually published in a pre-defined placeholder (e.g., a relational database table), which in turn can be accessed by execution systems or other planning applications. The use of elaborate mechanisms for user-driven analysis and judgmental or collaborative decisions, as opposed to reliance on automated DMT, remains a guiding principle for the current genre of business planning applications. The value of collaborative decision-making and global visibility of information is near axiomatic for business applications. However, future research needs to design better DMT applications that can utilize available information from disparate sources through advanced analytics, and account for specific domain knowledge, constraints or bottlenecks. Valuable and/or scarce human resources can be conserved by automating routine tasks and by reserving expert resources for high value added jobs (e.g., after a Pareto classification) or for exceptional situations (e.g., large prediction variance or situations of significant risks). In addition, certain research studies have indicated that judgmental overrides may not improve upon the results of automated descriptive and predictive models, on the (longer-term) average.

Scientists and engineers have traditionally utilized advanced quantitative approaches for making sense of observations and experimental results, formulating theories and hypotheses, and designing experiments. For users of statistical and numerical approaches in these domains, DMT often seems like the proverbial ìold wine in new bottlesî. However, innovative use of DMT include the development of algorithms, systems and practices that can not only apply novel methodologies but also scale to large scientific data repositories (e.g., see Han et al., 2002; Connover et al., 2003; Graves, 2003; Ramachandran et al., 2003; He et al., 2003). While scientific and business data mining have a lot in common, the incorporation of domain knowledge is probably more critical in scientific applications. When appropriately combined with domain specific knowledge about the physics or the data sources/uncertainties, DMT approaches have the potential to revolutionize the processes of scientific discovery, verification and prediction (e.g., Han et al., 2002; Karypis, 2002). This potential has been demonstrated by recent applications in diverse areas like remote sensing (e.g., Hinke et al., 2000), material sciences (e.g., Curtarolo et al., 2003), bioinformatics (e.g., Graves, 2003), and the earth sciences (e.g., Potter et al., 2003; Kamath et al., 2002; Thompson et al., 2002; Ganguly, 2002b; see also http://datamining.itsc.uah.edu/adam/). Besides physical and data dictated methods, human-computer interaction retains a significant role in real-world scientific decision-making. This necessitates the use of DSS, where the results of DMT can be combined with expert judgment and techniques from simulation, OR and other DSS tools. The ìReviews of Geophysicsî (a 1995 publication of the American Geophysical Union) provides a slightly dated discussion on the use of data assimilation, estimation and OR, as well as DSS, (e.g., http://www.agu.org/journals/rg/rg9504S/contents.html#hydrology). Examples of decision support systems and tools in scientific and engineering applications can also be found in dedicated journals like Decision Support Systems or Journal of Decision Systems (e.g., see vol. 8, number 2, 1998 as well as the latest issues), as well as in journals or web sites dealing with scientific and engineering topics (e.g., see the NASA air traffic control web site, http://www.asc.nasa.gov/aatt/dst.html; McCuistion and Birk, 2002; NASA research web sites, e.g., a global carbon DSS http://geo.arc.nasa.gov/website/cquestwebsite/index.html, and institutes like MIT Lincoln Laboratories, e.g., http://www.ll.mit.edu/AviationWeather/index2.html).

Business applications have focused on DSS, with embedded and scalable implementations of relatively straightforward DMT. Scientific applications have traditionally focused on advanced DMT in prototype applications with sample data. Researchers and practitioners of the future need to utilize advanced DMT for business applications and scalable DMT and DSS for scientists and engineers. This provides a perfect opportunity for innovative and multi-disciplinary collaborations.

CONCLUSION

The power of information technologies has been utilized to acquire, manage, store, retrieve and represent data in information repositories, and to share, report, process, collaborate on and move data in scientific and business applications. Database management and data warehousing technologies have matured significantly over the years. Tools for building custom and packaged applications, including but not limited to workflow technologies, web servers and GUI-based data entry and viewing forms, are steadily maturing. There is a clear and present need to exploit the available data and technologies to develop the next generation of scientific and business applications, which can combine data-dictated methods with domain specific knowledge. Analytical information technologies, which include DMT and DSS, are particularly suited for these tasks. These technologies can facilitate both automated (data-dictated) and human expert driven knowledge discovery and predictive analytics, and can also be made to utilize the results of models and simulations that are based on process ìphysicsî or business insights. If DMT and DSS were to be defined broadly, a broad statement can perhaps be made that while business applications have scalable but straightforward DMT embedded within DSS, scientific applications have utilized advanced DMT, but focused less on scalability and DSS. Multidisciplinary research and development efforts are needed in the future for maximal utilization of analytical information technologies in the context of these applications.

REFERENCES

Agosta, L., Orlov, L. M. and Hudso R. (2003). ìThe Future of Data Mining: Predictive Analytics.î *Forrester Brief*. November. 2 pages.

Apte, C., Liu, B., Pednault, E. P. D., Smyth. P. (2002). ìBusiness applications of data mining.î *Communications of the ACM*, 45(8): 49-53, August.

Bradley, P., Gehrke, J., Ramakrishnan, R., Srikant, R. (2002). ìScaling mining algorithms to large databases.î *Communications of the ACM*, 45(8): 38-43, August.

Carlsson, C., Turban, E. (2002). ìDSS: directions for the next decade.î *Decision Support Systems*, Elsevier, 33(2): 105-110, June.

Conover, H., Graves, S. J., Ramachandran, R., Redman, S., Rushing, J., Tanner, S., Wilhelmson, R. (2003). ìData Mining on the TeraGrid.î Poster Presentation, *Supercomputing Conference* Phoenix, AZ, Nov. 15, 2003.

Curtarolo, S., Morgan, D., Persson, K., Rodgers, J. and Ceder, G. (2003). ìPredicting crystal structures with data mining of quantum calculations.î *Physics Review Letters*, 91(13).

Fayyad, U., Uthurusamy, R. (2002). ìEvolving data mining into solutions for insights.î *Communications of the ACM*, 45(8): 28-31, August.

Geoffrion, A. M. and Krishnan, R. (eds.) (2003). ìE-business and management science ñ Mutual impacts (Parts 1 and 2).î *Management Science*, 49(10-11), October-November.

Ganguly, A. R. (2002a). ìSoftware Review ñ Data Mining Componentsî, Editorial review, *ORMS Today*. The Institute for Operations Research and the Management Sciences (INFORMS), 29(5), 56-59, October.

Ganguly, A. R. (2002b). ìA Hybrid Approach to Improving Rainfall Forecasts.î *Computers in Science and Engineering*. IEEE Computer Society and American Institute of Physics, 4(4): 14-21. July/August.

Graves, S. J. (2003). ìData Mining on a Bioinformatics Grid.î *SURA BioGrid Workshop*, Raleigh, N.C., Jan. 28 - 30, 2003.

Grossman, R., Kamath, C., Kegelmeyer, W., Kumar, V., and Namburu, R. (eds.) (2001). *Data Mining for Scientific and Engineering Applications*, Kluwer, September.

Grossman, R. L., Hornick, M. F., Meyer, G.. (2002). ìData mining standards initiative.î Communications of the ACM, 45(8): 59-61, August.

Grossman, R. L., Mazzucco, M. (2002). ìDataSpace: A Data Web for the Exploratory Analysis and Mining of Data.î *Computers in Science and Engineering*. IEEE Computer Society and American Institute of Physics, 4(4): 44-51. July/August.

Hammer, J. (ed.) (2003). ìAdvances in online analytical processing.î *Data & Knowledge Engineering*, Elsevier, 45(2), 127-256, May.

Han, J., Altman, R. B., Kumar, V., Mannila, H., Pregibon, D. (2002). ìEmerging scientific applications in data mining.î Communications of the ACM, 45(8): 54-58, August.

Hand, D., Mannila, H., Smyth, P. (2001). *Principles of Data Mining*, MIT Press, Cambridge, MA.

He, Y., Ramachandran, R., Li, X., Rushing, J., Conover, H., Graves, S. J., Lyatsky, W., Tan, A., Germany, G. (2003). ìFramework for Mining and Analysis of Space Science Data.î *SIAM International Conference on Data Mining* (2003), San Francisco, CA, May 1-3, 2003.

Hinke, T., Rushing, J., Ranganath, H. S., Graves, S. J. (2000). ìTechniques and Experience in Mining Remotely Sensed Satellite Data.î *Artificial Intelligence Review*, 14 (6): Issues on the Application of Data Mining, pp 503-531, December.

Kamath, C., Cant˙-Paz, E., Fodor, I. K., Tang, N. A. (2002). ìClassifying of Bent-Double Galaxies.î *Computers in Science and Engineering*. IEEE Computer Society and American Institute of Physics, 4(4): 52-60. July/August.

Karypis, G. (2002). ìGuest Editor's Introduction: Data Mining.î *Computers in Science and Engineering*. IEEE Computer Society and American Institute of Physics, 4(4): 12-13. July/August.

Kohavi, R., Rothleder, N. J., Simoudis, E. (2002). ìEmerging trends in business analytics.î *Communications of the ACM*, 45(8): 45-48, August.

Linden, A. and Fenn, J. (2003). ìHype Cycle for Advanced Analytics, 2003.î *Gartner Strategic Analysis Report*. May. 11 pages.

McCuistion, J. D. and Birk, R. (2002). ìFrom Observations to Decision Support: The New Paradigm for Satellite Data.î *NASA Technical Report*. Available at http://www.iaanet.org/symp/berlin/IAA-B4-0102.pdf.

Potter, C., Klooster, S., Steinbach, M., Tan, P., Kumar, V., Shekhar, S., Nemani, R. and R. Myneni, (2003). ìGlobal Teleconnections of Ocean Climate to Terrestrial Carbon Flux.î *Journal of Geophysical Research*, American Geophysical Union, 108 (D17), 4556.

Ramachandran, R., Rushing, J., Conover, H., Graves, S. J., Keiser, K. (2003). ìFlexible Framework for Mining Meteorological Data.î *American Meteorological Society's (AMS) 19th International Conference on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Long Beach, CA, Feb. 9-13, 2003.

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., Carlsson, C. (2002). ìPast, present, and future of decision support technology.î *Decision Support Systems*, Elsevier, 33(2): 111-126, June.

Smyth, P., Pregibon, D., Faloutsos, C. (2002). ìData-driven evolution of data mining algorithms.î *Communications of the ACM*, 45(8): 33-37, August.

Thompson, D. S., Nair, J. S., Venkata, S. S. D., Machiraju, R. K., Jiang, M., Craciun, G. (2002). ìPhysics-Based Feature Mining for Large Data Exploration.î *Computers in Science and Engineering*. IEEE Computer Society and American Institute of Physics, 4(4): 22-30. July/August.

Wang, G. C. S. and Jain, C. L. (2003). *Regression Analysis: Modeling and Forecasting*. Institute of Business Forecasting, 299 pages.

Yurkiewicz, J. (2003). ìForecasting software survey: Predicting which product is right for you.î *ORMS Today*, Institute for Operations Research and the Management Sciences (INFORMS), February.

Terms and Definitions

Analytical Information Technologies (AIT): Information technologies that facilitate tasks like predictive modeling, data assimilation, planning or decision-making, through automated data-driven methods, numerical solutions of physical or dynamical systems, human-computer interaction, or a combination. AIT includes DMT, DSS, BI, OLAP, GIS, and other supporting tools and technologies.

Business Intelligence (BI): Broad set of tools and technologies that facilitate management of business knowledge, performance and strategy, through automated analytics or human-computer interaction.

Business and Scientific Applications: End-user modules which are capable of utilizing AIT along with domain specific knowledge (e.g., business insights or constraints, process physics, engineering know-how). Applications can be custom built or pre-packaged and are often distinguished form other information technologies by their cognizance of the specific domains for which they are designed. This can entail the incorporation of domain specific insights or models, as well as pre-defined information and process flows.

Data Mining Technologies (DMT): Broadly defined, these include all types of data-dictated analytical tools and technologies that can detect generic and interesting patterns, scale (or can be made to scale) to large data volumes and help in automated knowledge discovery or prediction tasks. These include determining associations and correlations, clustering, classifying and regressing, as well as developing predictive or forecasting models. The specific tools used can range from ìtraditionalî or emerging statistics and signal or image processing to machine learning, artificial intelligence and knowledge discovery from large databases, as well as econometrics, management science and tools for modeling and predicting the evolutions of nonlinear dynamical and stochastic systems.

Data Assimilation: Statistical and other automated methods for parameter estimation, followed by prediction and tracking.

Decision Support Systems (DSS): Broadly defined, these include technologies that facilitate decision-making. These can embed DMT and utilize these through automated batch processes and/or user-driven simulations or what-if scenario planning. The tools for decision support include analytical or automated approaches like data assimilation and operations research, as well as tools that help the human experts or decision-makers manage by objectives or by exception like OLAP or GIS.

Geographical Information Systems (GIS): Tools that rely on data management technologies to manage, process and present geo-spatial data, which in turn can vary with time.

On-Line Analytical Processing (OLAP): Broad set of technologies that facilitate drill-down or aggregate analyses, as well as presentation, allocation and consolidation of information along multiple dimensions (e.g., product, location and time). These technologies are well-suited for management by exceptions or objectives, as well as automated or judgmental decision-making.

Operations Research (OR): Mathematical and constraint programming, and other techniques for mathematically or computationally determining optimal solutions for objective functions in the presence of constraints.

Predictive Modeling: The process through which mathematical or numerical technologies are utilized to understand or reconstruct past behavior, and predict expected behavior in the future. Commonly utilized tools include statistics, data mining and operations research, as well as numerical or analytical methodologies that rely on domain-knowledge.