IBM research in interactive data mining and scientific computing

EDWARD R. KRAL

IBM Almaden Research Center, San Jose, California

A number of research and development efforts from IBM are described in the areas of data mining and visualization of large data sets. Some useful principles for data mining and visualization are discussed, and a few research programs and available products are described. In addition, the current IBM line of hardware and support software for scientific computing is briefly described.

We are currently experiencing an explosion of data in the fields of business, finance, science, and engineering. Given large amounts of multivariate data, it can be difficult to identify meaningful patterns and relationships. Recent advances in data mining and visualization have developed in response to the need to filter and interpret these vast amounts of data. These include better algorithms for identifying patterns in the data, and better interactive tools for data analysis, manipulation, and presentation, which take advantage of principles of human perception and cognition.

Interactive Data Mining and Visualization

Data mining consists of a set of technologies for filtering and analyzing data in order to identify patterns among the many variables represented by the data. The analysis methodologies include many of the statistical, numerical, and computational techniques that have long been used in other fields of the sciences (Rogowitz, Rabenhorst, Gerth, & Kalin, 1996). Traditionally, because of large data sets and complex operations, data mining has not been an interactive process, but rather a batch process (Rogowitz et al., 1996). However, the tremendous increase in computational and graphics power that has occurred in the past few years has begun an evolution away from the traditional batch processes (Rogowitz et al., 1996), often undertaken on a centralized resource (mainframe), toward a more interactive model on decentralized workstations, and even more recently toward a mix of centralized and decentralized resources, referred to as the *client-server* model. The goal of interactive data mining is to present as much information to the user as possible in as interactive a mode as possible. Increased computational power will enable the user to interactively manipulate databases, analyze many scenarios in a short span of time, quickly evaluate the results of different algorithms, and perform "what-if" experiments on the data

(Rogowitz et al., 1996). In numerical simulations and in data mining analyses, the user will be able to interactively steer the calculation on the basis of intermediate results and find patterns that might otherwise have gone undiscovered.

In addition to data mining techniques, visualization tools are also necessary for extracting useful information from the data. Scientific visualization typically focuses on the presentation of physical quantities in terms of spatial variables, time, or other known physical quantities. However, nonphysical data, such as commercial or statistical data, often consist of vast tabular collections of loosely related events that have no natural coordinate system or intrinsic structure on which to formulate an analysis, and so their representation is necessarily abstract (Rogowitz et al., 1996). In this case, visual techniques such as the use of color, motion, orientation, and brightness may be used to naturally draw the user's attention to patterns in the data.

Data Mining Research and Tools

There are numerous ongoing projects involving data mining techniques within IBM. Several of these have evolved into supported products and custom solutions for particular needs. In general, all of the projects concern automatically extracting information from a collection of data to enable the user to act upon patterns and correlations that are not readily apparent. Some of these will be briefly described here. (More information about these projects may be accessed through the Decision Support Research web page <http://www.research.ibm.com/dss-grp/>)

Pattern discovery. The Quest data mining project (Agrawal et al., 1996) focuses on pattern recognition in large databases. Patterns are established by several means, some of which include association, in which one event implies another (e.g., 95% of customers who buy beer also buy pretzels); classification, in which events belonging to particular groups are used to develop a profile of that group; and sequential patterns, in which cause and effect (or at least temporally related) events are found (e.g., 30% of patients with condition X also develop condition Y). This research has led to the release of a data

The author would like to gratefully acknowledge the assistance of Chid Apte, Se June Hong, Rakesh Agrawal, and particularly Bernice Rogowitz in the preparation of this overview. Correspondence should be addressed to E. Kral, IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120-6099 (e-mail: ekral@almaden.ibm.com).

mining product, the IBM Intelligent Miner. Intelligent Miner utilizes these techniques and further proven techniques such as clustering and dependent value prediction to find relationships among the data. In addition, the package contains visualization functions to allow the user to easily interpret the results. (For more information on Quest, see <http://www.almaden.ibm.com/cs/quest/>; and for information on Intelligent Miner, <http://www.software.ibm.com/data/>)

Predictive modeling. The data abstraction research project focuses on developing robust techniques and systems for generating rule-based models for classification and regression from data. The result of the research is the Rule Abstraction for Modeling and Prediction (RAMP) system, a prototype package consisting of algorithms for developing effective rules for classification and regression (Apte, Hong, Lepre, Prasad, & Rosen, 1995). The system provides powerful new algorithms for selecting important and useful features, for optimally discretizing numeric features, and for generating minimal disjunctive normal form rules based on an R-MINI algorithm (Hong, in press). The system also provides many options for most effectively applying the rules by pruning and weighting and for evaluating the performance of the model for both predictive classification and rule-based regression. The RAMP system offers a user interface environment that encourages the user to experiment in seeking the best model for the end goal. (For more information, see <http://www.research.ibm.com/dar/>)

Clustering and text mining. Since the majority of the world's data is in the form of text (books, letters, manuals, etc.), tools to enable sorting through text-based data for patterns and correlations are an important aid in many areas of business, law, and education. Some products already available for this purpose include IBM InfoDetector and IBM Text Navigator, which utilize data segmentation (clustering) to separate data into clusters with similar characteristics. InfoDetector uses relational analysis, which may include factorial methods or hierarchical methods for data analysis, enabling the analysis of almost unlimited records that can be described by thousands of variables each. In addition, InfoDetector includes a visualization tool for analysis of results. (More information on text mining applications is on <http://direct.boulder.ibm.com/bi/tech/mining/textmine.htm>)

Visualization Research and Tools

Again, several ongoing projects have become well-established visualization products in the marketplace. Research in IBM encompasses visualization for data mining, chemistry, geophysics, pharmacology, and utilization of visual cues for data analysis. Two of the many projects will be described here. (For more information, refer to the Visualization Solutions link at <http://www.research.ibm.com/research/projects.html>)

Visualization Data Explorer. IBM Visualization Data Explorer is a full-function general-purpose visual-

ization package for use with scientific and statistical datasets. It utilizes a data-flow-driven execution model enabling the user to construct visualizations based on interactive mouse-driven construction of a "flow chart" of functional boxes. It runs on a client-server model, with the GUI running on the client. The server performs all of the computation and may reside on a single workstation or multiprocessor platform. Data Explorer also provides several analysis capabilities, such as arithmetic, transcendental, and Boolean expressions, univariate statistics, image processing functions, divergence, gradient, curl, and area and volume computations. A large number of references are available on the Data Explorer application. (These references and further information are available at <http://www.almaden.ibm.com/dx/>)

Visual cues (Diamond). An interesting body of research in visualization techniques has been performed by the Exploratory Visualization Group at the Watson Research Center. The focus of this work is the development of visualization tools based on visual and psychological principles in order to improve the interpretability of multivariate data (Rogowitz et al., 1996). The goal is to organize the data so as to make maximum use of the significant pattern recognition capabilities of the human brain. In one strategy, the data are organized by using visual cues that a human being will notice naturally, without conscious effort, in order to draw attention to interesting structures in the data. Another strategy is to give the user multiple linked representations of the data, so that all representations are updated dynamically whenever the user manipulates the data. For example, whenever a variable is transformed, or a subset of data is marked with a color, that transformation or color highlighting is propagated to all the representations.

The results of this research have been included in an interactive visual data mining package called Diamond, which is commercially available on OS/2, Windows 3.1, and UNIX from SPSS, Inc. (<http://www.spss.com>). An API version for integration with existing data mining applications is also under development. Various novel data representations are included—for example, a parametric snake plot, in which data points are connected in the order of another parameter, and a fractal foam plot, which is a glyph representation for interactively analyzing bivariate correlations among sets of variables (Rabenhorst, 1994). In addition to the various plot types that are available, other visual cues such as relative magnitude, color, motion, color mixing, and brightness are used to enhance the interpretive capabilities of the user.

Scientific Computing

In addition to the data mining and visualization packages introduced above, IBM offers a complete line of hardware and support software that provides the user with the computational power for meaningful interactive data analysis and visualization. The RS/6000 line of hardware provides platforms for every scale of analysis, from port-

ables utilizing the PowerPC 603e chip to the RS/6000 Scalable POWERParallel machine (SP), a scalable, distributed memory, multiple instruction multiple data (MIMD) massively parallel processing architecture utilizing the POWER2 chip. The processors in the RS/6000 line consist of the PowerPC 601, 603, and 604 chips, which offer excellent price/performance, to the multichip POWER2 and the new, integrated, single-chip POWER2 super chip (P2SC), which are competitive high-end workstation and parallel processing CPUs. All the workstations in the RS/6000 line run AIX, a robust, standards compliant, full-functioned UNIX operating system. The same operating system allows binary compatibility over the entire RS/6000 line, from the laptop to the supercomputer. In addition, a full line of programming languages and support tools is available, including Fortran 77 and 90, C, C++, a full MPI message-passing implementation, PVME message-passing, and debuggers and analysis tools.

Large-scale data mining and visualization computations can be run on the SP as a scalable server, with up to 512 nodes available. The nodes are connected by a high-performance (low-latency, high-bandwidth) flat-topology switch which allows reliable internode communication with no penalty for far-node messages. The distributed architecture and flat-topology switch allow scalable performance for a wide variety of applications. The SP supports a large number of both commercial and scientific applications, from data mining and transaction processing to computational geophysics and engineering. In addition, the RS/6000 line includes a full complement of graphical workstations to serve as clients and a line of symmetric multiprocessor (SMP) platforms for additional server tasks. (For more information on the RS/6000 line, see the RS/6000 home page at <http://www.rs6000.ibm.com/>)

Conclusion

Rapid increases in computational power are opening new frontiers of data analysis, heralding an age of truly interactive processing of large data sets and holding the promise of extracting more and more useful information from the vast amounts of data that currently exist and continue to increase. Greater computational power promises greater interactivity and the ability of the user to perform numerous data analysis tasks, perform experiments, and follow ideas with very little time penalty. The tools that will allow the user to harness this computational power and apply it to the data are being developed, with some already appearing in the market. These include data mining applications that will allow the user to extract hidden relationships and correlations from the data, and visualization tools that will reveal these relationships in a very easily understood manner.

REFERENCES

- AGRAWAL, R., ARNING, A., BOLLINGER, T., MEHTA, M., SHAFER, J., & SRIKANT, R. (1996). The Quest data mining system. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press.
- APTE, C., HONG, S. J., LEPRE, J., PRASAD, S., & ROSEN, B. (1995). *RAMP: Rules abstraction for modeling and prediction* (IBM Research Division Tech. Rep. No. RC-20271). White Plains, NY: IBM.
- HONG, S. J. (in press). R-MINI: An iterative approach for generating minimal rules from examples. *IEEE Transactions on Knowledge & Data Engineering*.
- RABENHORST, D. A. (1994). Interactive exploration of multidimensional data. *Proceedings of the International Society of Optical Engineering*, 2179, 277-286. Bellingham, WA: SPIE Press.
- ROGOWITZ, B. E., RABENHORST, D. A., GERTH, J. A., & KALIN, E. B. (1996). Visual cues for data mining. *Human Vision and Electronic Imaging: Proceedings of the International Society of Optical Engineering*, 2657, 275-301. Bellingham, WA: SPIE Press.