Mining Business Databases

Ad hoc techniques—no longer adequate for sifting through vast collections of data—are giving way to data mining and knowledge discovery for turning corporate data into competitive business advantage.

THE AMOUNT OF DATA COLLECTED AND WAREHOUSED IN ALL INDUSTRIES IS GROWING AT a phenomenal rate. From the financial sector to telecommunications operations, companies increasingly rely on analysis of huge amounts of data to compete. Although ad hoc mixtures of statistical techniques and file management tools once sufficed for digging through mounds of corporate data, the size of modern data warehouses, the mission-critical nature of the data, and the speed with which analyses need to be made now call for a new approach.

A new generation of techniques and tools is emerging to intelligently assist humans in analyzing mountains of data and finding critical nuggets of useful knowledge, and in some cases to perform analyses automatically. These techniques and tools are the subject of the growing field of knowledge discovery in databases (KDD) [5].

KDD is an umbrella term describing a variety of activities for making sense of data. We use the term to describe the overall process of finding useful patterns in data, including not only the data mining step of running specific discovery algorithms but also pre- and postprocessing and a host of other important activities. Our goal here is to provide a brief overview of the key issues in knowledge discovery in an industrial context and outline representative applications.

Tools and Users

The different data mining methods at the core of the KDD process can have different goals. In general, we distinguish two types:

- Verification, in which the system is limited to verifying a user's hypothesis, and
- Discovery, in which the system finds new patterns.

Discovery includes prediction, through which the system finds patterns to help predict the future behavior of some entities; and description, through which the system finds patterns in order to present the patterns to users in an understandable form.

Note that predictive models can be descriptive (to the degree they are understandable), and descriptive models can be used for prediction. Examples of key predictive methods include regression and classification (learning a function that maps a new example into one of a set of discrete classes). Key description methods include clustering, summarization, visualization, and change and deviation detection. Methods like dependency modeling (e.g., market basket analysis) can be either.

The knowledge discovery process in industry is performed mainly by analysts whose primary training and professional duties are in statistics and data analysis. The tools used are generally not expressly knowledge discovery tools but statistical analysis tools (e.g., S and SAS), graph- and chart-drawing software and spreadsheets, and database query engines. Direct programming in languages like C and AWK is typically used for complex analyses, usually on data selected from a database and dumped into a flat file for further manipulation.

Of the tools explicitly supporting knowledge discovery, some are generic, and some domain-specific; some are aimed at supporting single tasks and some at supporting multiple tasks.

There are dozens of generic, single-task tools available, especially for classification, using primarily decision-tree, neural-network, example-based, and rule-discovery approaches. Such tools mainly support only the core data mining step in the knowledge discovery process and require significant pre- and post-processing. The typical target user of these tools is a consultant or a developer who integrates them with other modules as part of a complete application.

Generic, multitask tools perform a variety of discovery tasks, typically combining classification (perhaps using more than one approach), visualization, query/retrieval, clustering, and more. Examples include Clementine [7], IMACS [4], MLC++, MOBAL, and Recon. These tools support more of the KDD process and simplify embedding of discovered knowledge into an application the business user can use. The target user of such tools is usually a sophisticated analyst who understands data manipulation. While these tools typically require special training and some tool customization, Clementine, in particular, is reported to have been widely used without customization by a variety of users ranging from business analysts to biochemists. Such use is made possible by a highly graphical user interface to data mining functions. IMACS uses a knowledge representation system to represent domain and task objects and integrate various aspects of the discovery process. IMACS allows the user to create new, more natural object-centered views of data stored in arbitrary ways (e.g., relational databases and flat files). Data can be segmented in a simple way using these views, and new segments can be defined easily from old ones. IMACS was a big step toward allowing business users to interact with data in terms they are familiar with, since the views (or concepts) were expressed entirely in user terms rather than being limited to database schemas created for other purposes.

Domain-specific tools support discovery in only a single domain, usually using the user's language; users need know little about the analysis process itself. Examples of such tools include Opportunity Explorer [1], which generates reports on changes in retail sales; IBM Advanced Scout, which analyzes NBA basketball game statistics and finds patterns of play coaches can

Ronald J. Brachman, Tom Khabaza, Willi Kloesgen,

Gregory Piatetsky-Shapiro, and Evangelos Simoudis

TERRY WIDENER

Domain-specific tools support discovery in only a single domain, usually using the user's language; users need know little about the analysis process itself.

use right away; and AT&T's Interactive Data Exploration and Analysis (IDEA) system [11], which focuses on understanding the market's reaction to promotions, new service offerings, and ongoing advertisements. IDEA also has an intuitive visual language for representing data mining procedures and is especially powerful for segmenting data, an operation very important in marketing applications.

These systems (along with Clementine and IMACS) represent an important trend—moving knowledge discovery technology directly into the hands of business users. The key elements that help make the core statistical, machine learning, and other data mining technologies accessible to a mainstream user are:

- Putting the problem in the business user's terms, including viewing the data from a business model perspective (e.g., concepts and rules);
- Providing support for specific key business analyses (e.g., segmentation);
- Presenting results in a form geared to the business problem being solved; and
- Providing support for a protracted iterative exploratory process (discussed in the next section).

The Knowledge Discovery Process

The core of the knowledge discovery process is the set of data mining tasks used to extract and verify patterns in data. However, this core typically takes only a small part (estimated at 15%–25%) of the effort of the overall process. No complete methodology for this process exists yet, but knowledge discovery takes place in a number of stages [3]:

- Getting to know the data and the task: This stage is more significant than it sounds, especially when the data is to be pulled from multiple sources and when the analysis will not be done by the business user.
- Acquisition: Bringing the data into the appropriate environment for analysis.
- Integration and checking: Confirming the expected form and broad contents of the data and integrating the data into tools as required.
- Data cleaning: Looking for obvious flaws in the data and removing them, and removing records with errors or insignificant outliers.
- Model and hypothesis development: Simple exploration of the data through passive techniques and elaboration by deriving new data attributes where necessary; selection of an appropriate model in which to do analysis; and development of initial hypotheses to test.
- Data mining: Application of the core discovery procedures to reveal patterns and new knowledge or to verify hypotheses developed prior to this step.
- Testing and verification: Assessing the discovered knowledge, including testing predictive models on test sets and analyzing segmentation.
- Interpretation and use: Integration with existing domain knowledge, which may confirm, deny, or challenge the newly discovered patterns.

Throughout the process, visualization of results is also an integral activity. And at the end, we often have a business-need-specific application constructed from the results for everyday use by business users.

A key thing to note about a realistic knowledge discovery process is that it is not simple and linear, but thoroughly iterative and interactive. The results of analysis are fed back into the modeling and hypothesis derivation process to produce improved results on subsequent iterations. This activity takes time, and if applied to data generated on a regular basis (e.g., quarterly or yearly results), it can have a long lifespan. Systems like IDEA are beginning to support more of the infrastructure aspects of the process (e.g., it supports sequences of operations), allowing reuse of complex analyses on variant datasets.

Representative Applications

Knowledge discovery applications and prototypes have been developed for a variety of domains, including marketing, finance, banking, manufacturing, and telecommunications. A majority of the applications use a predictive modeling approach, although a few notable applications use other methods.

Marketing. A long-time user of statistical and other quantitative methods, marketing has been in the forefront of adopting knowledge-discovery techniques. Most marketing applications fall into the broad area called *database marketing* (*mailshot response* in Europe). This approach relies on analysis of customer databases, using such techniques as interactive querying, segmentation, and predictive modeling to select potential customers in a more precisely targeted way. *BusinessWeek* recently estimated that more than 50% of all U.S. retailers use or plan to use database marketing and that those using it have good results (e.g., a 10%–15% increase in credit card use, as reported by American Express).

An interesting application for predicting the size of television audiences using neural networks and rule induction was developed in the U.K. by Integral Solutions Ltd. for the BBC. Rule induction is used to identify the factors playing the most important roles in relating the size of a program's audience to its scheduling slot. The final models performed as well as human experts but were highly robust against change, because the models could be retrained from up-to-date data. (An example of a trained system is a rule-reduction learning system.)

Other applications are more descriptive, focusing on finding patterns that help market analysts make better decisions. Among the first systems developed and deployed in this area were Coverstory and Spotlight [2], which analyzed supermarket sales data and produced reports (using natural language and business graphics) on the most significant changes in a particular product volume and share broken down by region, product type, and other dimensions. In addition, such causal factors as distribution channels and price changes were also examined and related to changes in volume and share. These systems were quite successful. Spotlight later grew into the Opportunity Explorer system [1], which supports sales representatives of consumer packaged-goods companies in examining their business with individual retailers. This support is accomplished through presentations highlighting the advantages for the retailer if additional products are stocked or special promotions are performed. A new feature of Opportunity Explorer is generation of interactive reports with hyperlinks, allowing easy navigation between different report sections.

The Management Discovery Tool (MDT) system, a product being developed by AT&T and NCR, incorporates several other innovative ideas to allow business users to interact directly with data. MDT incorporates a set of business rules (encoded as metadata) that make it easy to set up monitors for detecting significant deviations in key business indicators. MDT also allows automatic report generation (in HTML form), helping users understand the causes of changes and to point and click to drill down into more detailed analyses. To accommodate mainstream business users, MDT provides a limited set of analysis types, including summarization, trend analysis, change analysis, and measure and segment comparison.

Another marketing area is market basket analysis—the study of retail stock movement data recorded at a point of sale—to support decisions on shelf-space allocation, store layout, and product location and promotion effectiveness. IBM offers tools to automatically find all interesting associations (the system incorporates notions of what is interesting); Lucent Technology's NicheWorks visualization system allows clustered purchases to be visualized intuitively.

Financial Investment. Many financial analysis applications employ predictive modeling techniques (e.g., statistical regression and neural networks) for such tasks as creating and optimizing portfolios and creating trading models. To maintain a competitive advantage, users and developers of such applications, which have been in use for several years, rarely publicize their precise details and effectiveness.

We can, however, point to a few examples. The Fidelity Stock Selector fund uses neural network models to select investments, performing quite well until recently. The output of these models is evaluated by the fund manager before the action is taken, so it is not entirely clear how to divide the credit between human and machine.

LBS Capital Management, a fund-management firm, uses expert systems, neural nets, and generic algorithms to manage portfolios worth $600 million; since its introduction in 1993, the system has outperformed the overall stock market [6].

Carlberg & Associates developed a neural network model for predicting the Standard & Poor's 500 Index, using interest rates, earnings, dividends, the dollar index, and oil prices. The model was surprisingly successful, explaining 96% of the variation in the Index from 1986 to 1995.

In these applications, predictive accuracy is paramount; the need to use the extracted knowledge to explain a recommended action is less important.

Thus, the main focus is ensuring that modeling methods do not overfit the data.

Fraud Detection. Many systems developed for fraud detection are not publicized, for obvious reasons, but several are worth mentioning. The FALCON fraud assessment system from HNC, Inc. was developed using a neural network shell and is now used by many retail banks to detect suspicious credit card transactions. The Financial Crimes Enforcement Network AI System (FAIS) system [12] from the U.S. Treasury's Financial Crimes Enforcement Network helps identify financial transactions that may indicate money-laundering activity. FAIS, which uses data from government forms, consists of a combination of off-the-shelf and custom components. Its use is expected to expand to a variety of government agencies concerned with detection of suspicious financial transactions. FAIS must overcome a difficult data quality problem because much of its data comes from handwritten notes.

In financial investment applications, predictive accuracy is paramount; the need to use the extracted knowledge to explain a recommended action is less important. The main focus is ensuring that modeling methods do not overfit the data.

AT&T developed a system for detecting international calling fraud by displaying calling activity in a way that lets users quickly see unusual patterns. The Clonedetector system, developed by GTE, uses customer profiles to detect cellular cloning fraud. If a particular customer suddenly starts calling in a very different way, fraud alert automatically kicks in. Another cellular fraud detection system is under development at NYNEX. The developers first mine the data to discover indicators of fraudulent use; subsequently, they automatically generate detection systems by feeding these indicators into a detector-constructor program, which uses the indicators to instantiate detector templates. Finally, the system learns to combine the detectors for optimal performance.

Manufacturing and Production. Controlling and scheduling technical production processes is a prospective KDD application field with a high potential for profit. Although large volumes of data generated during a production process are often only poorly exploited, in the long run it may be possible to control production processes automatically by discovering and using patterns indicative of high-quality products.

For example, one experiment being run in a large chemical company in Europe assists in the production process for polymeric plastics [8]. Control variables include the quantities of raw material and the heating parameters; process variables include temperatures and pressures measured at various locations, as well as chemical reaction times. The quality of the final product is assessed in a laboratory according to various criteria (e.g., degree of anhydrosity). Quality variables are determined several times a day and process and control variables nearly continuously. The key is determining the relationship among the three sets of variables. Even if time and location are represented only in a simple way, rule-inducing discovery methods can derive valuable insight into the basic relationships.

In a joint venture of General Electric and SNECMA, a troubleshooting system called CASSIOPEE was developed by Acknosoft based on the KATE discovery tool. The system is being applied by three major European airlines to diagnose and predict problems in Boeing 737 aircraft. Clustering methods are used to derive families of faults.

The main advantage of knowledge discovery applications in this area is the high cost savings achievable when the results are used to control an expensive production or operation process. Key challenges are the representation and exploitation of time and location, as well as model levels, such as quality, process, and control.

Network Management. Another application area with a strong temporal component is the management of telecommunication networks. These large and com-

plex networks produce many alarms daily, sequences of which contain implicit information about the behavior of the network. Using data mining, valuable knowledge about the overall system and its performance can be extracted. Regularities in the alarms can be used in fault management systems for filtering redundant alarms, locating problems in the network, and predicting severe faults. In this vein, the Telecommunication Alarm Sequence Analyzer (TASA) was built at the University of Helsinki in cooperation with a manufacturer of telecommunications equipment and three telephone networks [9]. The system uses a novel framework for locating frequently occurring alarm episodes from the alarm stream and presenting them as rules. The system discovered rules that were then integrated into the alarm-handling software of the telephone networks.

Other Areas. Health care is an information-rich and high-payoff area, ripe for data mining. One of the first applications in this area was KEFIR [10], which performs an automatic drill-down through data along multiple dimensions to determine the most interesting deviations of specific quantitative measures relative to their previous and expected values. It explains key deviations through their relationships to other deviations in the data and, where appropriate, generates recommendations for actions in response to these deviations. KEFIR uses a Web browser to present its findings in a hypertext report, using natural language and business graphics.

Data quality is another promising area for data mining applications; data mining tools have helped verify financial trading data, detecting errors impossible to detect through conventional means. Yet another area generating excitement is that of discovering knowledge on the Internet, using autonomous programs to gather and collate information available from the Web and other sources.

Application Development

While much data mining technology is well developed, its practical application in industry is affected by a number of issues:

- Insufficient training: Graduates of business schools are familiar with verification-driven analysis techniques, occasionally with predictive modeling but rarely with other discovery techniques.
- Inadequate tool support: Most available data mining tools support only one of the core discovery techniques, typically prediction. Other methods, such as clustering, deviation detection, visualization, and summarization, are also needed, as are methods for dealing with exceptions (rare cases) that may be significant in some applications. The tools must also support the complete knowledge discovery process and provide a user interface suitable for business users rather than for other technologists.
- Data unavailability: For a given business problem, the required data is often distributed across the organization in a variety of formats, and the data is often poorly organized or maintained. For this reason, data acquisition and preprocessing usually play a significant part in any knowledge discovery project. Data warehousing is becoming widespread and can potentially alleviate such problems.
- Overabundance of patterns: When the search for patterns has a wide scope, a large number of patterns can be discovered. Proper statistical controls are needed to avoid discoveries due to chance, while domain knowledge can help the system focus on the interesting findings. Rule refinement and other generalization methods could be used to further compress findings.
- Changing and time-oriented data: Many applications deal with behavior that changes significantly over time (e.g., stock market fluctations). Such applications are more challenging, because common algorithms suitable for flat tables do not work well with sequential and other time-oriented patterns. But such applications can become especially successful, since it is easier to retrain a system than to retrain a human. A few recently developed data mining methods are designed for handling deviation detection and time-oriented data.
- Spatially oriented data: Other applications, especially in manufacturing, biology, and geographically oriented systems, deal with spatial data in which there are special types of patterns requiring special algorithms.
- Complex data types: Other types of information, including text, images, audio, video, and anything related to the Web, present an even grander challenge, with potentially great rewards.
- Scalability: Current tools cannot handle truly vast quantities of data. Data warehouses starting at 200GB are no longer rare, yet, at best, our tools can deal with 1GB at a time. However, progress is being made toward using massively parallel and high-performance computing systems to help deal with large databases.

The Potential for KDD Applications

Domains suitable for knowledge discovery are information-rich, have a changing environment, do not already have existing models, require knowledge-based decisions, and provide high payoff for the right decisions. Given a suitable domain, the costs and benefits of a potential application are affected by the following factors:

- Alternatives: There should be no simpler alternative solutions.
- Relevance: The key relevant factors need to be present in the data.
- Volume: There should be a sufficient number of cases (several thousand at least). On the other hand, extremely large databases may be a problem when the results are needed quickly.
- Complexity: The more variables (fields) there are, the more complex the application. Complexity is also increased for time-series data.
- Quality: Error rates should be relatively low.
- Accessibility: Data should be easily accessible; accessing data or merging data from different sources increases the cost of an application.
- Change: Although dealing with change is more difficult than not dealing with change, it can be more rewarding, since the application can be automatically and regularly retrained on up-to-date data.
- Expertise: The more expertise available, the easier the project. It should be emphasized that expertise on the form and meaning of the data is as important as knowledge of problem-solving in the domain.

Overall, knowledge discovery technology promises to improve industry's ability to cope with and exploit its ever-growing abundance of data. ◻

Acknowledgments

We thank Usama Fayyad and Sam Uthurusamy for encouraging us to write this article, Padhraic Smyth for his ideas on data mining tasks, and Robert Golan for help on financial applications. Colin Shearer also contributed much useful material.

References

1. Anand, T. Opportunity Explorer: Navigating large databases using knowledge discovery templates. *J. Intell. Inf. Syst. 4*, 1 (Jan. 1995), 27–38.
2. Anand, T., and Kahn, G. Focusing knowledge-based techniques on market analysis. *IEEE Expert 8*, 4 (Aug. 1993), 19–24.
3. Brachman, R., and Anand, T. The process of knowledge discovery in databases: A human-centered approach. In *Advances in Knowledge Discovery and Data Mining,* U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI Press/The MIT Press, Cambridge, Mass., 1996, pp. 37–57.
4. Brachman, R., Selfridge, P., Terveen, L., Altman, B., Borgida, A., Halper, F., Kirk, T., Lazar, A., McGuinness, D., and Resnick, L. Integrated support for data archaeology. *Int. J. Intell. Coop. Inf. Syst. 2*, 2 (June 1993), 159–185.
5. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., Eds. *Advances in Knowledge Discovery and Data Mining*. AAAI Press/The MIT Press, Cambridge, Mass., 1996.
6. Hall, J., Mani, G., and Barr, D. Applying computational intelligence to the investment process. In *Proceedings of CIFER-96: Computational Intelligence in Financial Engineering*. (New York, March 1996). IEEE Press, Piscataway, N.J., 1996.
7. Khabaza, T., and Shearer, C. Data mining with Clementine. In the digest of the *IEE Colloquium on Knowledge Discovery in Databases*, Digest 1995/021(B) (London, Feb. 1995).
8. Kloesgen, W. Tasks, methods, and applications of knowledge extraction. In *New Techniques and Technologies for Statistics*, W. Kloesgen, P. Nanopoulos, and A. Unwin, Eds. IOS Press, Amsterdam, 1996, 163–182.
9. Mannila, H., Toivonen, H., and Verkamo, A.I. Discovering frequent episodes in sequences. In *Proceedings of 1st International Conference on Knowledge Discovery and Data Mining* (Montréal, Aug. 1995). AAAI Press, Menlo Park, Calif., 1995, pp. 210–215.
10. Matheus, C., Piatetsky-Shapiro, G., and McNeill, D. Selecting and reporting what is interesting: The KEFIR application to healthcare data. In *Advances in Knowledge Discovery and Data Mining*, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI Press/The MIT Press, Cambridge, Mass., 1996, pp. 495–516.
11. Selfridge, P.G., Srivastava, D., and Wilson, L.O. IDEA: Interactive Data Exploration and Analysis. In *Proceedings of SIGMOD-96* (Montréal, June 1996). ACM Press, New York, 1996, pp. 24–34.
12. Senator, T.E., Goldberg, H.G., Wooten, J., Cottini, M.A., Khan, A.F.U., Klinger, C.D., Llamas, W.M., Marrone, M.P., and Wong, R.W.H. The Financial Crimes Enforcement Network AI System (FAIS): Identifying potential money laundering from reports of large cash transactions. *AI Mag. 16*, 4 (Winter 1995), 21–39.

Additional references for this article can be found at http://www.research.microsoft.com/research/datamine/CACM-DM-refs/.

RONALD J. BRACHMAN is vice president of Information Systems and Services Research at AT&T Laboratories. He can be reached at rjb@research.att.com.

TOM KHABAZA is the head of Consultancy Services in the Data Mining Division of Integral Solutions Ltd. He can be reached at tomk@isl.co.uk.

WILLI KLOESGEN is a senior researcher in the Knowledge Discovery Group of the German National Research Institute for Information Technology (GMD). He can be reached at kloesgen@gmd.de.

GREGORY PIATETSKY-SHAPIRO is a principal member of the technical staff at GTE Laboratories, Inc. and the principal investigator of the company's Knowledge Discovery in Databases Project. He can be reached at gps@gte.com.

EVANGELOS SIMOUDIS is vice president of data mining and decision support solutions at IBM's Almaden Research Center. He can be reached at simoudis@almaden.ibm.com.