The Data-Mining Industry Coming of Age

Gregory Piatetsky-Shapiro, Knowledge Stream Partners

The information revolution is generating mountains of data, from sources as diverse as credit card transactions, telephone calls, Web clickstreams, space science, and human genome research. At the same time, faster and cheaper storage technology lets us store greater amounts of data online, and better database-management-system software provides easy access to those databases. Walmart, for example, built an 11-Tbyte database of customer transactions in 1998.

Traditional, interactive, ad hoc data-analysis methods cannot sufficiently handle such massive amounts of data. Researchers in data mining and knowledge discovery are creating new, more automated methods for discovering knowledge to meet the needs of the 21st century. This need for analysis will keep growing, driven by the business trends of one-to-one marketing, customer-relationship management, and Web personalization—all of which require customer-information analysis and customer-preferences prediction.

Shorter-term technology trends include migration of some mining functions from the server to the desktop and emerging tool interoperability through standards such as

- the cross-industry standard process for data mining (www.crisp-dm.org) model, for the data-mining process;
- the predictive modeling markup language specification, for the output of mining models (www.dmg.org); and
- the Microsoft OLE DB interface, for retrieving database information.[1]

If adopted, these standards will allow interoperability among different mining tools and integrate data mining with other applications, including database systems, spreadsheets, and decision support. Already, vendors such as Megaputer and Angoss have rolled out component-object model and distributed component-object model interfaces that let their tools function as add-ons to Microsoft Excel or Access.

The Web is expanding the focus of data mining beyond the traditional analysis of structured data. There are now huge amounts of information in free text, images, sounds, and video that are hard to analyze. Current systems, such as Altavista Photo and Media Finder (image.altavista.com) and MP3 search engines (mp3.lycos.com), can find an image or a sound based on the surrounding text, but they fail at more complex requests such as finding similar images or sounds. However, help is on the way with XML, which describes the Web document metadata in a computer-friendly way and will greatly facilitate analysis of previously unstructured data.

From tools to solutions

The first generation of what we now call data-mining systems appeared in the 1980s and consisted of research-driven tools focusing on single tasks. These tasks included building a classifier using a decision-tree tool[2] (for example, C4.5) or a neural network (for example, SNNS, www.informatik.uni-stuttgart.de/ipvr/bv/projekte/snns/snns.html), finding clusters in data (for example, Autoclass[3]), or data visualization (for example, Alfred Inselberg's parallel-coordinate approach[4]). Such tools addressed a generic data-analysis problem, and their intended user had to be technically sophisticated. Also, using more than one tool on the same data set was difficult and required significant data and metadata transformation.

Data-mining vendors developed the second-generation data-mining systems, called suites, around 1995. These tools were driven by the realization that the knowledge-discovery process requires multiple types of data analysis and most of the effort is spent in data cleaning and preprocessing.[5] The suites, such as SPSS's Clementine, Silicon Graphics' Mineset, IBM's Intelligent Miner, and The SAS Institute's Enterprise Miner, let the user perform several discovery tasks (usually classification, clustering, and visualization) and supported data transformation and visualization. Clementine (www.spss.com/clementine) pioneered the GUI, an important advance that let users visually build their knowledge-discovery process. By 1998, according to Herb Edelstein,[6] the data-mining tool market was at $45 million, a 100% improvement from 1997.

We can get another look at the tool trends by examining the tools listed at kdnuggets.com—a popular directory of data-mining and knowledge-discovery information that I maintain (see Table 1). The stabilization in the number of suites shows that the market for commercial tools is maturing. Still, the number of commercial tool vendors will probably decline as the industry consolidates. However, research is still active in association

and sequence algorithms, Bayesian networks, and data visualization. Finally, the newer areas of text and Web mining are growing rapidly.

While second-generation systems empower data analysts, they require significant knowledge of statistical theory to be used properly and cannot be used directly by the business users. Business users' needs led to the third generation of vertical data-mining-based applications and solutions in the late 1990s. These solutions were oriented toward solving a specific business problem, such as detecting credit card fraud or predicting cell phone customer attrition. They addressed the entire knowledge discovery in databases process, including access to legacy systems, and pushed the results into a front-end application. The interfaces were oriented to the business user and hid all the data-mining complexity. HNC Software's Falcon for credit card fraud detection, IBM's Advanced Scout for basketball game analysis, and NASD Regulation's Advanced-Detection System[7] exemplify such systems.

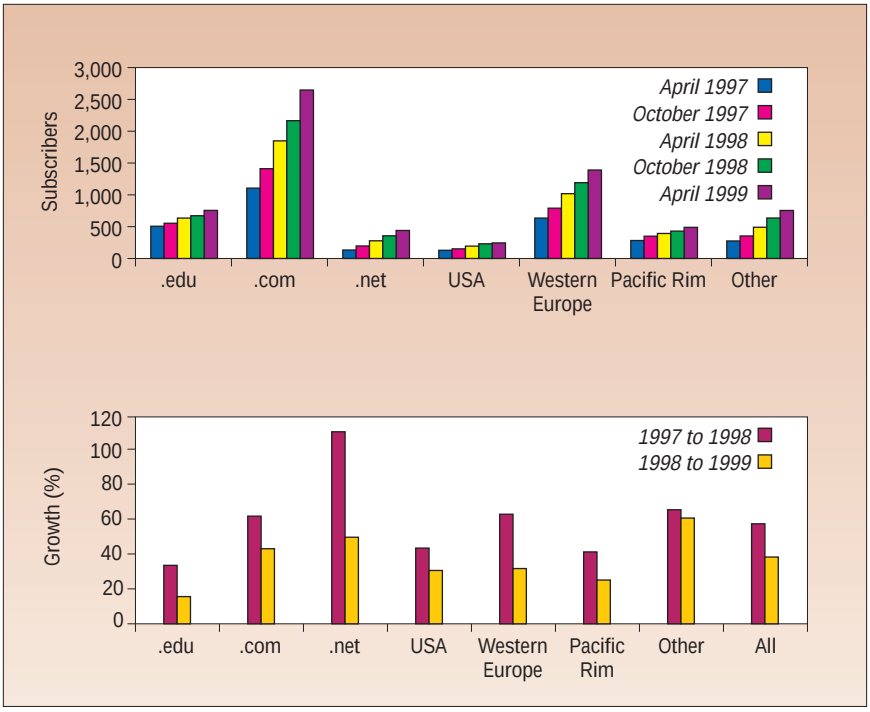

Figure 1. *KDNuggets News* subscribers by domain and region.

Applications and trends

Good data-mining application areas require knowledge-based decisions; have accessible, sufficient, and relevant data; have a changing environment; have suboptimal current methods; will not be obsoleted by imminent new technology; and provide a high payoff for the correct decisions.

Data-mining-based vertical applications were developed for numerous areas that fit these requirements—including banking and credit, bioinformatics, customer relationship management, name and address merging, health care and human resources, Internet advertising, e-commerce, insurance, investment, manufacturing, marketing, retail, sports and entertainment, and telecommunications. These applications solve the most popular business problems, including customer retention, cross-sell and up-sell, and fraud detection. (See kdnuggets.com for a catalog of solutions.)

Web mining presents a whole new set of opportunities and challenges, such as analyzing the clickstream data to select in real time the right pop-up advertisement or e-commerce offer. Because a Web site usually has much less history on a visitor than a bank has on its customer, researchers are developing new ways to make predictions. One promising technology is collaborative filtering, originally developed at MIT and implemented in systems such as Firefly Network (acquired by Microsoft in 1998) and NetPerceptions (www.netperceptions.com). These systems use information on what items previous users purchased or selected to predict and suggest what the current user might like. However, all such systems must protect user privacy. For example, Purchase Circles, launched by Amazon.com in August 1999, met with stiff resistance from users who did not want Amazon.com to broadcast information on what items people were ordering.

We can see other interesting trends by analyzing the *KDNuggets* newsletter subscriber list (kdnuggets.com/news). The subscriber base grew from 50 people in 1993 to over 8,000 as of November 1999. The subscriber list is still growing at 7% a quarter—much more slowly than the 40% a quarter in 1994 and the 20% a quarter in 1997. Researchers made up the majority of the subscribers in the first few years, but now most are from commercial domains. The region analysis shows widespread acceptance of data-mining technologies in the US and Western Europe, followed by the Pacific Rim, Eastern Europe, and Latin America (see Figure 1).

Table 1. The growth of commercial and public-domain tools in different categories, according to kdnuggets.com.

Tool type	Feb. 1997	Feb. 1998	Nov. 1998	Aug. 1999
Suites	18	30	37	37
Classification:				
Multiapproach	4	8	8	10
Decision tree	14	18	18	19
Rule discovery	10	11	12	14
Neural network	50	60	76	100
Bayesian	0	0	10	20
Other	6	7	12	13
Total	84	104	126	176
Associations and sequences	0	0	6	13
Clustering	6	5	10	12
Visualization	5	12	26	31
Text and Web mining	0	5	7	15

Challenges for the 21st century

A big challenge for 21st-century miners will be to come up with widely accepted standards on data-mining processes and models. If adopted, they will stimulate major industry growth. Another challenge is to come up with real-time algorithms for Web mining and to better use the rich textual and multimedia Web data. Many companies strive for seamless integration of data mining into business processes. Finally, these advances must be balanced with privacy protection.[8]

The Y2K problem will continue to haunt data miners long after we consume the millennium champagne. Although most business have converted their operational systems to handle Y2K problems, very few went back to correct their historical data, where two-digit years are still common. A data miner in 2012 looking at a year marked "11" will need to decide if it refers to 1911 or 2011; the correct answer will depend on the application.

Despite all these challenges, I am optimistic that the data-mining industry will overcome its growing pains and become as accepted as the database industry. Data mining will become just another service provided by a good database system. ■

References

1. S. Willett, "Microsoft's Plan to Bring Data Mining to Masses," *Computer Reseller News*, 25 May 1999; www.techweb.com/wire/story/TWB19990525S0001 (current Nov. 1999).
2. J.R. Quinlan, "Induction of Decision Trees," *Machine Learning*, Vol. 1, 1986, pp. 81–106.
3. P. Cheeseman et al., "AUTOCLASS: A Bayesian Classification System," *Proc. ICML-86*, Morgan Kaufmann, San Francisco, 1988, pp. 54–64.
4. A. Inselberg, "The Plane with Parallel Coordinates," *The Visual Computer*, Vol. 1, 1985, pp. 69–91.
5. R.J. Brachman and T. Anand, "The Process of Knowledge Discovery in Databases," *Advances in Knowledge Discovery and Data Mining*, U. Fayyad et al., eds., AAAI/MIT Press, Cambridge, Mass., 1996, pp. 37–57.
6. "Two Crows Releases 1999 Technology Report," *Data Mining News*, Vol. 2, No. 18, 10 May 1999.
7. J.D. Kirkland et al., "The NASD Regulation Advanced-Detection System," *AI Magazine*, Vol. 20, No. 1, Spring 1999, pp. 55–67.
8. J. Markoff, "The Rise of Little Brother," *Upside*, Apr. 1999; http://www.upside.com/texis/mvm/story?id=36d4613c0 (current Nov. 1999).

Gregory Piatetsky-Shapiro is the chief scientist at Knowledge Stream Partners. He also runs the kdnuggets.com Web site and e-newsletter. His interests include knowledge discovery and its practical applications in business and e-commerce. His received his MS and PhD in computer science from New York University. He founded the Knowledge Discovery in Data meetings in 1989; chaired KDD-91, KDD-93, and KDD-98; and is currently the ACM SIGKDD director. He is also a member of the IEEE Computer Society, ACM, and AAAI. Contact him at Knowledge Stream Partners, 148 State Street, Boston, MA 02109; gshapiro@ksp.com.