MIT Technology Review

MIT Technology Review 10 BREAKTHROUGH TECHNOLOGIES 2001

Data Mining

M. MITCHELL WALDROP
Saturday, February 1, 2003

Usama Fayyad

"Hello again, Sidney P. Manyclicks. We have recommendations for you. Customers who bought this title also bought..."

Intrusive? A touch of personal attention in the sterile world of e-shopping? Both, perhaps-but definitely a tour de force of database technology. Conventional databases sort though a few megabytes of structured data to find answers to specific queries. But compiling a simple recommendation list requires a system that can burrow through gigabytes of Web site visitor logs in search of patterns no one can anticipate in advance.

Welcome to data mining, also known as knowledge discovery in databases (KDD): the rapidly emerging technology that lies behind the personalized Web and much else besides. The emphasis here is on "emerging," says Usama Fayyad, who should know: data mining didn't exist as a field until he helped pioneer it.

In 1987, the Tunisian-born computer scientist was a graduate student at the University of Michigan. He had taken a summer job with General Motors, which was compiling a huge database on car repairs. The idea, he says, was to enable any GM service technician to ask the database a question based on the model of car, the engine capacity, and so on, and get a quick, appropriate response. Sounds straightforward. But, recalls Fayyad, "there were hundreds of millions of records-no human being could go through it all." The pattern recognition algorithm he devised to solve that problem became his 1991 doctoral dissertation, which is still among the most cited publications in the data-mining field.

Data mining proved to have surprisingly broad application. Fayyad left Michigan for NASA's Jet Propulsion Laboratory, where he applied his techniques to astronomical research. In particular, his algoithm helped in automatically determining which of some two billion observed celestial objects were stars and which were galaxies. The tool also helped find volcanoes on Venus from the huge number of radar images being transmitted from space probes. A geologist could retrieve the image of a previously identified volcano; the computer would then examine the picture for patterns and search through other images for similar patterns. That worked so well, Fayyad says, that "pretty soon the military intelligence people were all over us, wanting to use it. And so were doctors, who wanted to do automatic searches of radiology images." In 1995, in response to this widening interest, Fayyad and his colleagues planned a full-scale international conference on KDD. The conference drew about 500 participants, more than double what had been expected. (KDD 2000 drew 950.)

By this time, with the Internet gushing information onto everyone's desktop, the urgency for data mining was becoming evident in the corporate world. IBM and other industry giants sensed a market--and wanted in. Microsoft set its sights on Fayyad and enticed him to join the company's research labs. "They suggested that I take a look at databases in the corporate world," says Fayyad. "It was pretty sad. In many companies, the 'data warehouses' were actually 'data tombs': the data went in and were never looked at again." Fayyad joined Microsoft in 1996 and organized a new research group in data mining. "We looked at new algorithms for scaling up to very large databases-gigabytes or larger," he says.

By decade's end, Fayyad had caught the entrepreneurial bug sweeping through computer science labs. "I realized that even the organizations that loved the idea of data mining were having trouble just maintaining their data." What they needed, he reasoned, was a company to

host their databases for them, and provide data-mining services on top of that. The result was digiMine, a Kirkland, Wash., startup that opened for business in March 2000 with Fayyad as CEO.

And the future of data-mining technology? Wide open, says Fayyad-especially as researchers begin to move beyond the field's original focus on highly structured, relational databases. One very hot area is "text data mining": extracting unexpected relationships from huge collections of free-form text documents. The results are still preliminary, as various labs experiment with natural-language processing, statistical word counts and other techniques. But the University of California at Berkeley's LINDI system, to take one example, has already been used to help geneticists search the biomedical literature and produce plausible hypotheses for the function of newly discovered genes.

Another hot area, says Fayyad, is "video mining": using a combination of speech recognition, image understanding and natural-language processing techniques to open up the world's vast video archives to efficient computer searching. For instance, when Carnegie Mellon University's Informedia II system is given an archive of, say, CNN news clips, it produces a computer-searchable index by automatically dividing each clip into individual scenes accompanied by transcripts and headlines.

Fayyad hopes that ultimately the techniques of data mining will become so successful and so thoroughly integrated into standard database systems that they will no longer be thought of as exotic. "People will just assume that their database software will do what they need."

Others in Data Mining

Organization	Project
Howard Wactlar (Carnegie Mellon)	Search very large video collections
Marti Hearst (University of California, Berkeley)	Automated discovery of new information from large text collections
Nokia Research Center (Helsinki, Finland)	Finding recurrent episodes in event sequence data
Raghu Ramakrishnan (University of Wisconsin)	Visual exploration of data on the Web