

Knowledge



Discovery

We're interested in techniques that automatically find fundamental properties and principles that are original and useful.

TOSHINORI MUNAKATA,
Guest Editor

THE NOTION OF COMPUTERS AUTOMATICALLY FINDING useful information is an exciting and promising aspect of just about any application intended to be of practical use. Human civilization itself is the result of the vast accumulation of past discoveries, insights, inventions, and innovations. Replacing such creative human activities as sorting through it all with a computer represents a grand human endeavor. This special section presents an overview of knowledge discovery, or the computer techniques that, in the broadest sense, automatically find fundamental properties and principles.

The criteria we used to select the topics covered in these articles include the quality of the knowledge discovered and the practicality of that knowledge. For example, we would be interested in knowledge discovered by a machine that is original, non-trivial, fundamental, and simple, yet useful. We emphasize specific and convincing application examples, rather than mere promises. Background theory and general descriptions of the techniques employed in each article are generally omitted or minimal, left to other sources. Most authors also limited themselves to only a few application examples and a few references.

We hope the example applications give you a

sense of the state of knowledge discovery. You'll find more on the basics, as well as more on applications, in the references.

Since we cover such diverse domains, even just classifying the articles is not a simple matter. For convenience, they are organized into three parts. The first—General Domains—covers representative areas of knowledge discovery. Tom Mitchell of Carnegie-Mellon University explores applications and future perspectives of knowledge discovery in databases and data mining, emphasizing symbolic machine learning techniques. The example application he describes—inducing rules from the medical records of almost 10,000 pregnant women—represents a

Dealing with more comprehensive, non-oversimplified, real-world types of data, such as mixed data types, is a key direction for the future.

prototypical application of knowledge discovery in databases today. Although a number of techniques are also applied to such domains as management, science, and engineering, the idea is the same—distilling the underlying characteristics of a large volume of data.

Raúl Valdés-Pérez, also of Carnegie-Mellon University, emphasizes the aspects of knowledge discovery that are shared across the various fields of science. Although computers can't crank out major scientific findings every day, they promise to accelerate the rate of significant scientific discoveries. After explaining some basic concepts and identifying a niche for machine discovery in scientific practice, Valdés-Pérez describes example applications in medicine, mathematics, and chemistry.

The second part—Selected Techniques—covers four especially important techniques for knowledge discovery. Stephen Muggleton of the University of York in the U.K. explores induction by first-order predicate logic. Despite the general importance of first-order predicate logic in AI, it is not among the most common techniques in today's knowledge discovery applications. This article reveals its power in these applications.

LiMin Fu of the University of Florida discusses techniques and example applications of extracting underlying knowledge by using neural networks. Based on a simulation of the architecture and functioning of the human brain, neural networks are being applied to many types of problems, including pattern recognition, process control, and real-world

prediction of real-world phenomena. For example, given a large set of input-to-output mapping instances, a neural network can learn to repeat the mapping not only for the given training data but also for other similar patterns. A neural network can also identify a visual image, audio segments, and various types of signals. But looking closely at a neural network, we find that the only information in the network is a set of numeric weights associated with the interconnections of its neurons. Extracting or making sense of these numeric weights to come up with a higher level of knowledge has been and will continue to be a challenging problem.

Kenneth De Jong of George Mason University surveys knowledge discovery through evolutionary computation, also known as genetic algorithms. Planet Earth's five-billion-year history has produced increasingly complex forms of biological life. Evolutionary computation generally finds increasingly refined solutions based on the principles of genetics and evolution. Perhaps someday, new knowledge that other techniques has failed to find will emerge through evolutionary computation. De Jong introduces applications in which evolutionary computation has helped identify better strategies, heuristics, and solutions rarely obtained through ordinary search methods.

Wojtek Ziarko of the University of Regina in Canada reviews knowledge discovery applications using rough set theory—a relatively new and unknown technique. It has, however, been employed more widely than you might imagine, in such appli-

cations as data mining and reasoning about imprecise or incomplete data. It is already a valuable option for knowledge discovery in databases.

The third part—Specialized Topics—covers several specific application domains in knowledge discovery. Kevin Knight of the University of Southern California discusses knowledge discovery in natural-language processing, dealing with various forms of text. It is used extensively in such everyday computer applications as word processing, Internet communication, Web browsing, and machine translation. Moreover, in light of its range of applications in various types of data, often in huge volume, it holds especially practical implications in knowledge discovery. For example, future word processing software is likely to include knowledge discovery features or their derivatives, in addition to such “intelligent” tools as spelling and grammar checking.

Steffen Schulze-Kremer of the Max Planck Institute in Berlin discusses how knowledge discovery is used in molecular biology, particularly in the human genome project. As the field of molecular biology has advanced in recent years, using computers to help discover knowledge in biological data represents an important new tool in the field. Conversely, the knowledge discovered in nature may contribute to how information is processed in computer systems.

Finally, Murray Campbell of IBM describes Deep Blue, the first computer system to defeat a human world chess champion in an accredited match. Deep Blue unites enormous computational power with chess knowledge, enabling the system to determine moves at about the same level as the best humans. This approach, based on computational power and highly tuned knowledge, has great potential for developing new search strategies in many disciplines.

Please note that the areas and techniques covered in this special section are not exhaustive. For example, some common techniques in knowledge discovery in databases, such as statistical methods and database technology, are not included (see other issues of *Communications*, including November 1996 on data mining).

Future Directions

Knowledge discovery is a promising and challenging field, as the articles show. Some especially promising elements include:

Incorporation of background and associated knowledge. In any discipline, from scientific discovery

to business strategy, there is often a limit to the acquisition of knowledge from a single source of raw data, whatever the technique employed. Incorporating any associated knowledge significantly increases the efficiency of the process and the quality of the knowledge discovered.

More comprehensive types of data. Most knowledge discovery techniques today deal with relatively simple forms of target data. As Mitchell points out, dealing with more comprehensive, non-oversimplified, real-world types of data, such as ultrasound images and mixed data types, is a key direction for the future.

Human-computer interaction. Any human and any computer each reflect individual strengths and weaknesses. For example, a computer is incredibly fast and unbiased but lacks common sense and intuition. Deep Blue is a good example of human-computer interaction for delivering an advanced level of computation.

Hybrid systems. Each of the techniques discussed here also has advantages and weaknesses. Hybrid systems will help compensate for a particular system's weaknesses, thus creating new approaches to knowledge discovery.

When the two other special sections for which I was also the guest editor—both on AI (March 1994 and November 1995)—were published, practical applications of many intelligent techniques were beginning to be widespread. Today, the term “intelligent computing” is everywhere, from consumer products to industrial and scientific applications. Knowledge discovery in the 21st century will take a course in research and applications similar to that of AI in recent years.

In the future, techniques will continue to evolve and new methods will be devised, although we don't know yet what they are and when they may appear. **G**

TOSHINORI MUNAKATA (munakata@cis.scuohio.edu) is a professor in the Computer and Information Science Department at Cleveland State University in Ohio.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

© 1999 ACM 0002-0782/99/1100 \$5.00