AI Magazine Volume 11 Number 5 (1990) (© AAAI)

Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop

Gregory Piatetsky-Shapiro

The growth in the amount of available databases far outstrips the growth of corresponding knowledge. This creates both a need and an opportunity for extracting knowledge from databases. Many recent results have been reported on extracting different kinds of knowledge from databases, including diagnostic rules, drug side effects, classes of stars, rules for expert systems, and rules for semantic query optimization.

The importance of this topic is now recognized by leading researchers. Michie predicts that "The next area that is going to explode is the use of machine learning tools as a component of large scale data analysis'' *(AI Week,* March 15, 1990). At a recent NSF invitational workshop on the future of database research (Lagunita, CA, February 1990), "knowledge mining" was among the top five research topics.

The viability of knowledge discovery is also being recognized by business and government organizations. American Airlines is looking for patterns in its frequent flyer databases. Banks are analyzing credit data to determine better rules for credit assessment and bankruptcy prediction. General Motors is automatically constructing diagnostic expert systems from a database of car trouble symptoms and problems found. The IRS is looking for patterns of tax cheating in its databases. Those are only some of the examples.

Some of the research has matured enough to find its way into commercial systems for rule discovery in databases. Several such systems have appeared recently, including IXL™, BEAGLE™, KnowledgeSeeker™, AIM™, and KnowledgeMaker™.

Knowledge discovery in databases is an interesting topic, drawing from several fields including expert systems, machine learning, intelligent databases, knowledge acquisition, case-based reasoning and statistics. The *Knowledge Discovery in Databases* workshop, held on August 20, 1989 in Detroit, MI, during IJCAI-89, had succeeded in bringing together many leading researchers in Machine Learning, Expert Databases, Knowledge Acquisition, Fuzzy Sets, and other areas. The workshop had interesting presentations and lively panel discussion, with lots of interaction. It helped to remove some of the misconceptions that Machine Learning researchers have about databases—i.e. databases are not static tables of simple data, but complex entities with transactions, security, and updates. While those researchers just want to use the data, the database people want to integrate the acquired knowledge into the database system. The workshop also helped to educate the database researchers about the available wealth of approaches to machine discovery.

I was the chairman of the workshop. The program committee consisted of Jaime Carbonell, Carnegie Mellon University, William Frawley, GTE Laboratories, Kamran Parsaye, IntelligenceWare, Los Angeles, J. Ross Quinlan, University of Sydney, Michael Siegel, Boston University and MIT, and Ramasamy Uthurusamy, GM Research Laboratories.

The workshop generated a significant international interest. We received 69 submissions from 12 countries: USA (39), Canada (9), UK (3), P.R. China (3), Italy (3), France (2), Sweden (2), India (2), Belgium (2), Germany (2), Japan (1), and Australia (1). Thirty nine contributors from 9 countries were invited to attend the workshop. The workshop was also attended by Robert Simpson from DARPA and Y.T. Chien, Director of AI & Robotics at the NSF.

Nine excellent papers were presented in three sessions: Data-Driven Discovery Methods, Knowledge-Based Approaches, and Systems and Applications. The revised versions of the workshop papers will be included in the forthcoming collection on *Knowledge Discovery in Databases,* to be published by AAAI and MIT press in early 1991, and will not be discussed here.

The workshop concluded with a stimulating panel discussion by Pat Langley, Larry Kerschberg and J. Ross Quinlan.

There was general agreement that Knowledge Discovery is a promising research direction that will become more important as the number of domains for which there are no human experts increases. Applications to large real databases will require algorithms that are efficient and handle uncertainty well. More complex domains demand the use of more expressive (i.e., first-order) languages.

Use of Domain Knowledge in Discovery

There was, however, a spirited disagreement on usage of domain (background) knowledge, which can be defined as any information not explicitly present in the data. Tom Mitchell pointed out that some problems are so large (e.g. molecule segmentation in Meta-Dendral), that you need some domain constraints to limit the search space (e.g. double bonds don't break). Ross Quinlan suggested that if we can design efficient algorithms that search well, we should try to avoid such constraints, because they limit what we can find.

Jaime Carbonell said that he really liked constraints, and that in his latest domain—logistics planning—the size of search space may be 10^{300} without constraints like "Trucks don't drive over water". No system, no matter how efficient, can search a space that large.

Someone pointed out, however, that in the winter trucks can drive over water. This example, in a nutshell, illustrates the utility and pitfalls of using domain knowledge to constrain the search. Use of domain restraints narrows the search, enabling us to tackle larger problems and solve them faster. The danger is that the constraints will prevent the discovery of some unexpected solutions, such as where trucks drive over water. Depending on our objectives, we should be able to play both sides of this trade-off.

Quinlan suggested that it is OK to use domain knowledge verified by

the data. Langley said that a good way for dealing with this dilemma is to develop incremental learning systems, that can discover new things and reuse them in discovery, thereby boot-strapping themselves. Such systems can be started with little or no domain knowledge and eventually reach a good level. However, it was pointed out that it took scientists several hundred years to discover the necessary background knowledge for physics and chemistry, so the proposed incremental discovery system may have to run for quite a while.

Good Application Domains

The discussion then turned to what characterizes good domains for knowledge discovery.

Quinlan pointed that in order to learn something you have to almost know it already. A minimum requirement for a good domain is that you have measurements of the important parameters. As an example of a bad domain he gave off-shore oil drilling rigs that collect and send ashore enormous amounts of data, which are just measurements of various things on the platform. Nobody knows whether they are relevant to the production of the platform or not. To go looking in a database like that, which is very large, is probably not a good idea.

The reason medical domains are so appropriate for discovery is that (1) we have medical databases, (2) a considerable medical expertise goes into deciding what is recorded in the databases, and (3) it is very easy to outperform some doctors.

Kerschberg suggested applying discovery to "Legacy systems". These are old systems, developed in the 60s and 70s, using obsolete technology (some are still using punchcards!). The problem arises when the people who have maintained them are no longer available, while the systems still perform a useful business function. There are companies that maintain such legacy systems by doing reverse engineering of the old programs to extract the data model, and then forward engineering the code to transfer the old database into a new system. Such reverse engineering can use whatever knowledge can be discovered from the data itself.

Philip Schrodt from the University of Missouri described the Inter-University Consortium for Political and Social Research, which has about 2000 data sets—mostly social survey data for the last twenty years. He suggested there is plenty to discover there.

Langley observed that machine learning has had the most success with simple diagnostic tasks. Domains like medical diagnosis, or diagnosis of problems with cars, are where most of the initial high payoff applications are likely to be.

Barry Silverman suggested that there are many government databases with real discovery problems. For example, IRS databases contain many interesting patterns, although some researchers may not want to work on them on the grounds of the Fifth Amendment. The FBI has many years of records of all airplane crashes. Analysis of this database may contribute to improved safety in the air.

Many corporate databases are also good targets for discovery, with numerous such projects underway. However, the corporations (and especially banks and insurance companies) are not likely to publish if they find something of a competitive advantage.

Directions for Future Research

Finally, I asked the panel to describe good and bad directions for future research.

Quinlan said that there are many interesting directions, e.g. incremental algorithms vs batch algorithms, very fast algorithms. Almost anything in this area has some substance. A bad direction was to prove that some algorithm is better than ID3. A more serious example of bad research is trying to show how to squeeze out the last half of one percent of performance.

Langley addressed methodological concerns. He suggested building on what has been done, before coming up with new algorithms. One should not run a system on a couple of examples and see "if this is good". Rather, the discovered knowledge should be brought back into the database to see whether it is useful. Building tightly integrated systems is important as it will force us to generalize our theories.

Langley seconded a call by Bob Simpson for having more tools, tested and documented, made available to other people.[1] This, of course, will make it easier for other people to run comparative studies. They may show how bad your system is, but that is what science is all about.

Kerschberg took an engineering view. He suggested taking some of the existing algorithms and seeing how they scale up on very large databases. It is not so important how long it takes to discover knowledge, because it can be done off-line. But once the useful knowledge is discovered, it should be brought back into the database in some form, such as integrity rules, semantic optimization rules, or functional dependencies.

Pandora's Box?

An important issue that was only touched at the workshop is the appropriateness of discovery. A careless approach to discovery may open a Pandora's Box of unpleasant surprises.

Some kinds of discovery are actually illegal. The federal and state privacy laws limit what can be discovered about individuals. Use of drug trafficker's "profiles" by law enforcement agencies has been very controversial and some parts of the profile, such as race, have been ruled illegal.

Political, ethical and moral considerations may affect other discoveries. The FBI proposal for setting up a nationwide database of criminal suspects was shut down after congressional objections on invasion of privacy.

A pattern that involves racial or ethnic characteristic is likely to be controversial. The FDA ban in April 1990 on blood donations by people from Haiti and Sub-Saharan Africa is a good example. The discovered pattern of high incidence of AIDS in those groups was protested as being racially motivated, because there was also a high incidence of AIDS in another geographical group (men from New York and San Francisco), who were not forbidden to donate blood.

However, from the media reports on this controversy it was not clear what is the strength of those patterns, and what additional factors were considered. To avoid purely emotional arguments in such cases it is desirable to give more detailed information such as

> Based on a sample of size *S,* people in group *Z* have 33 to 41 % likelihood of developing the disease *X*. *P*% of all people fall into group *Z*. The nationwide risk of developing *X* is 10 to 12 %.

This will allow the public a better perception of the pattern and will decrease the controversy.

Summary

The workshop confirmed that knowledge discovery in databases is an idea whose time has come. Some of the important research issues addressed in the workshop were:

Domain knowledge.

It should be used to reduce search space, but used carefully so as not to prevent un-anticipated discoveries. While a specialized learning algorithm will outperform a general method, a desirable compromise is to develop a framework for augmenting the general method with the specific domain knowledge.

Dealing with Uncertainty.

Databases typically have missing, incomplete or incorrect data items. Thus any discovery algorithm must deal with noise. Rules discovered in noisy data will necessarily be approximate.

Efficiency

Exponential and even high-order polynomial algorithms will not scale for dealing with large volumes of data. Efficient linear or sublinear (using sampling) algorithms are needed.

Incremental Approach.

Incremental algorithms are desirable for dealing with with changing data. An incremental discovery system that can re-use its discoveries may be able to boot-strap itself.

Interactive Systems.

Perhaps the best practical chance for discovery comes from systems, where a "knowledge analyst" uses a set of intelligent, visual and perceptual tools for data analysis. Such tools would go far beyond the existing statistical tools and significantly enhance the human capabilities for data analysis. What tool features are necessary to support effective interaction? Algorithms need to be re-examined from this point of view (e.g. a neural network may need to generate explanations from its weights).

The incremental, interactive discovery methods may transform the static databases of today into evolving information systems of tomorrow. Caution is required for discovery on demographic databases, to avoid findings that are illegal or unethical.

Some of the research issues that were little addressed in this workshop, but are likely to become more important in the future are:

Discovery Tools.

Deductive and object-oriented database systems can provide some of the needed support for induction on large volumes of data. Parallel hardware may be effectively used. What additional operations should be provided by the tools to support discovery?

Complex Data

Dealing with more complex (not just relational) data, including text, geographic information, CAD/CAM, and visual immages.

Better Presentation.

The discovered knowledge can be represented not only as rules, but as text, graphics, animation, audio patterns, etc.. Research on visualization and perceptual presentation is very relevant here.

Finally, discovery systems should be applied to real databases and judged on whether or not they can make useful and significant discoveries. Some successful applications have already been reported and more are on the way!

Acknowledgements.

I am very grateful to all the workshop participants for making it all possible. I thank Bud Frawley, Chris Matheus, Larry Kerschberg, and Wray Buntine for their insightful comments on this report, and Shri Goyal for his encouragement. The expressed opinions and mistakes remaining are mine only.

Note

1. This is the goal of the recently started *MACHINE LEARNING TOOLBOX* project at the University of Aberdeen in Scotland, UK, run in cooperation with Esprit.

About the Author

Gregory Piatetsky-Shapiro is a Senior Member of Technical Staff at GTE Laboratories, 40 Sylvan Road, Waltham, MA 02254 (gps0@gte.com). He leads a project on building intelligent tools for Discovery in Databases, using a combination of machine learning, statistical and visual methods. He has been both a researcher and a practitioner in AI & Database fields for the last twelve years and has over a dozen publications. He and William Frawley are editors of the forthcoming collection on Knowledge Discovery in Databases (AAAI/MIT press, Spring 1991). Gregory received his M.S. ('79) and Ph.D. ('84) from New York University.