Knowledge Discovery in Databases: 10 years after

Gregory Piatetsky-Shapiro
Knowledge Stream Partners
148 State Street
Boston, MA 02109
gps@ksp.com

ABSTRACT

In this paper, we describe the past 10 years of KDD and outline predictions for the next 10 years.

Keywords

Knowledge Discovery in Databases, Data Mining, KDD, History.

1. THE PRE-HISTORY OF KDD

I have long been interested in topics of machine learning and discovery. For my Ph.D., I studied how a database system could optimize itself through machine learning. In 1986, while working at GTE Laboratories on CALIDA, an intelligent database assistant for integrating multiple databases, I found a query that could be optimized by a factor of 100 if the optimizer knew a simple pattern like

`file1` join with `file2` will always have `field1` = A

I have started looking at the ways of automatically discovering such patterns and attended Gio Wiederhold tutorial in 1987 at Int. Conf. on Database Engineering in LA, entitled "Extracting Knowledge from Data". Gio and his student R. Blum [1] have developed Rx, the first program that analyzed historical data from about 50,000 Stanford patients, and looked for unexpected side-effects of drugs. The program did discover some side effects that were unknown to its authors, and the approach looked very promising.

However, I could not quite convince GTE management that discovery in data was a good idea. One senior manager told me that he thought that data mining was a solved problem -- I could apply a decision tree (and building a decision tree was a solved problem, wasn't it?) to the database and presto -- I will have all the results I need.

Later, I attended a AAAI-88 workshop in Minneapolis on "Databases and Expert Systems". The workshop had interesting presentations and was a good way to get researchers to interact. Putting together a workshop seemed relatively easy (little did I know) and I decided to organize a workshop on discovery in data at next year's IJCAI-89. That would be a great way to stimulate more research in the field and to convince my management at GTE Laboratories that discovery in data was a good idea.

What should I call this workshop? The name "data mining" which was already used in the database community seemed unsexy, and besides statisticians used "data mining" as a pejorative term to criticize the activity. "Mining" is unglamorous and there is no indication what are we mining *for*. "Knowledge mining" and "knowledge extraction" did not seem much better, and "database mining™" was trademarked by HNC for their Database Mining Workstation™. So, I came up with "*Knowledge Discovery in Databases*", which emphasized the "discovery" aspect and the focus of discovery on "knowledge".

With encouragement and help from Jaime Carbonell (CMU), Bud Frawley (GTE), Kamran Parsaye (IntelligenceWare), Ross Quinlan (U. of Sydney), Michael Siegel (BU), and Sam Uthurusamy (GM Research), I put together a *Knowledge Discovery in Databases (*KDD-89) workshop at IJCAI-89 in Detroit.

The term "Knowledge Discovery in Databases" (KDD for short) became popular in the AI and Machine Learning community. However, the database researchers were on better speaking terms with the business folks and the press, and the term "data mining" became much more popular in the business press. As of Nov 1999, search on www.altavista.com gives about 100,000 pages for "data mining", compared to 18,000 for "knowledge discovery". Currently, both terms are used essentially as synonyms, as in the name of the main journal for the field -- "Data Mining and Knowledge Discovery" (Kluwer). Sometimes "knowledge discovery process" is used for describing the overall process, including all the data preparation and postprocessing while "data mining" is used to refer to the step of applying the algorithms to the clean data [3].

2. KDD-89 WORKSHOP

The KDD-89 workshop (www.kdnuggets.com/meetings/kdd89) was very successful, receiving 69 submissions from 12 countries. It was the largest workshop at IJCAI-89, with standing-room only attendance.

KDD-89 had 9 papers presented in 3 sessions, on Data-Driven Discovery, Knowledge-Based Approaches, and Systems and Applications and concluded with a summary panel discussion by Larry Kershberg, Ross Quinlan, and Pat Langley.

The main topics discussed at the KDD-89 workshop included:

- Expert Database Systems
- Scientific Discovery
- Fuzzy Rules
- Using Domain Knowledge
- Learning from Relational (Structured) Data
- Dealing with Text and other Complex Data
- Discovery Tools
- Better Presentation Methods
- Integrated Systems
- Privacy

3. AREAS OF SLOW PROGRESS

Some of these topics, like Expert Database Systems and discovery of fuzzy rules, which seemed like good ideas at the time, have disappeared from current research lexicon because they were examples of technology without a clear application.

Some important areas turned out to be much harder than we thought in 1989.

Learning from structured data is still very difficult and current best methods from the Inductive Logic Programming community (see http://www.cs.bris.ac.uk/~ILPnet2/) are still too slow to be used on large databases.

Interestingness of discovered patterns is still a hard problem, and it still requires significant amount of using domain knowledge. CYC [6], which held a lot of promise in 1989, did not produce the expected results. On the other hand, we now have the web that is the largest repository of general knowledge, although still with very imperfect query system.

Privacy, which was a concern 10 years ago, especially proper balancing between companies' desire to use personal information versus individual's desire to protect it, remains a thorny issue. Recent initiatives like P3P (www.w3.org/P3P), which is standard for encoding privacy preferences in a machine-readable way, may solve the technical issues of exchanging personal information. However, they do not solve the fundamental privacy concerns of individuals. In my opinion, the way to solve these concerns is to

1. allow people to opt-in into any list which will use their information
2. allow people to get some benefit in exchange for their personal info -- same way as people get supermarket discount coupons in exchange for using their frequent shopper cards.
3. allow opt-out at any time

4. GOOD PROGRESS

Great progress was achieved in faster hardware and bigger disks, enabling data miners to deal with much larger problems. Of course, the major development in computers over the last 10 years is the revolution brought by the Web. It has shifted the attention of data miners to problems of e-commerce and web personalization. It also brought much more attention to **Text Mining**, and resulted in a number of good text mining systems.

Another major advance was a holistic understanding of the entire Knowledge Discovery Process [2], which encompasses many steps from data acquisition, cleaning, preprocessing, to discovery step, to postprocessing of the results and their integration into operational systems.

Good progress was also achieved in Ensemble Classifiers (Boosting, Bagging); Association Rules; OLAP; and Data Visualization.

5. FROM GENERAL TOOLS TO DOMAIN SPECIFIC SOLUTIONS

In 1989 there were only a few data mining tools, produced by researchers to solve a single task, such as C4.5 decision tree [7] and SNNS neural network, or parallel-coordinate visualization [4]. These tools were difficult to use and required significant data preparation.

The second generation data mining systems, called suites, were developed by data mining vendors, starting from 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. The suites such as SPSS Clementine, SGI Mineset, IBM Intelligent Miner, or SAS Enterprise Miner allowed the user to perform several discovery tasks (usually classification, clustering, and visualization) and also supported data transformation and visualization. An important advance, pioneered by Clementine, was a GUI, which allowed users to build their knowledge discovery process visually.

By 1999, there are over 200 tools available for many different tasks (see `www.kdnuggets.com/software`). However, even the best data mining tools addressed only a part of the overall business problem. Data still had to be extracted from legacy databases, cleaned and preprocessed, and model results had to be delivered to the right channels and, most importantly, integrated with the specific application or business logic. Successful development of such applications in areas like direct marketing, telecom, and fraud detection, led to emergence of data-mining-based "vertical solutions".

Examples of such systems include HNC Falcon for credit card fraud detection, IBM Advanced Scout for basketball game analysis, and NASD KDD Detection system [5].

5. GROWTH OF KNOWLEDGE DISCOVERY FIELD

The Knowledge Discovery research community grew tremendously over the last 10 years. From one 1 workshop with about 50 attendees we have progressed to over 10 international conferences a year (see www.kdnuggets.com/meetings), dozens of courses, many publications, and several journals focused on the field.

In 1989, a really large database was 1 MB. Today, we have multi-terabyte databases that are being mined.

In 1989 there were a handful of companies providing data mining tools. In 1999 there were over 100 companies.

Other interesting trends could be observed by analyzing the subscribers to KDNuggets News, a popular newsletter on Data Mining and Knowledge Discovery topics that I am publishing (see `www.kdnuggets.com/news`. The subscriber base grew from 50 people who received the first issue in 1993 to about 8000 as of Nov 1999. KDNuggets subscriber list is still growing at 7% a quarter, but much slower than 40% a quarter in 1994 or 20% a quarter in 1997.

While researchers made the majority of the subscribers in the first few years, now the majority is from commercial domains. About half of subscribers are from .com and .net domains, but there are significant groups of data miners in Western Europe (especially UK, Germany, and France) and Pacific Rim (especially Australia, Japan and Singapore). Surprisingly, there are pockets of subscribers in over 80 countries, and on all continents except Antarctica.

6. EXPECTATIONS FOR THE FUTURE

Over the next 10 years, I expect to see continuing progress in faster CPU, bigger disks, along with ubiquitous and wireless net connectivity.

I expect standards to appear for different parts of the knowledge discovery process, and greatly facilitate industry growth. Already we have proposed standards like CRISP (www.crisp-dm.org) for the data mining process, PMML (www.dmg.org) for predictive model exchange, and Microsoft OLE DB.

Significant applications will appear in E-commerce, especially with real-time personalization. There will be significant use of intelligent agents.

I also expect great progress in pharmaceuticals and new drugs enabled by knowledge discovery and bioinformatics.

I think there will be tighter integration of knowledge discovery modules with a database system, and most database systems will include a set of discovery operations.

I expect also that the data mining industry will overcome the hype stage, and will merge with the database industry.

ACKNOWLEDGMENTS

My thanks to Rakesh Agrawal, John Elder, Usama Fayyad, Ross Quinlan, and Sam Uthurusamy who participated in KDD-99 Ten-year anniversary panel and made it interesting. The views expressed here should be blamed on me only.

7. REFERENCES

[1] Blum, Robert L. and Gio Wiederhold: "Studying Hypotheses on a Time-Oriented Clinical Database: An Overview of the RX Project"; Proc. of the 6th Symposium on Computer Applications in Medical Care, Washington, D.C., IEEE 82 , pages 725-735. 1981

[2] Brachman, R. and T. Anand. The Process of Knowledge Discovery in Databases: A Human-Centered Approach. In *Advances in Knowledge Discovery and Data Mining*, ed. U.. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI/MIT Press, 1996.

[3] Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. From Data Mining to Knowledge Discovery in Databases (a survey), AI Magazine, 17(3): Fall 1996, 37-54 .

[4] Inselberg, A. (1985). The plane with parallel coordinates. The Visual Computer, 1, 69-91.

[5] Kirkland, J. Inselberg, et al, The NASD Regulation Advanced-Detection System(ADS) , AI Magazine 20(1): 55-67.

[6] Lenat, D. B. "Cyc: A Large-Scale Investment in Knowledge Infrastructure." Comm. of the ACM 38, no. 11, November 1995

[7] J. Ross Quinlan: Induction of Decision Trees. Machine Learning, Volume 1, 1986

About the author:

Gregory Piatetsky-Shapiro is a founder and editor of KDNuggets™ News e-newsletter, the first and most popular newsletter for the Data Mining and Knowledge Discovery community, and the associated KDNuggets.com website.

Gregory is also a Vice-President and Chief Scientist at Knowledge Stream Partners, (www.ksp.com), a consulting and integration eCRM company, and a director of ACM SIGKDD.

Gregory organized the first KDD meeting in 1989.