Report on the SIGKDD 2001 Conference Panel "New Research Directions in KDD"

Johannes Gehrke
Department of Computer Science
Cornell University
johannes@cs.cornell.edu

Prediction is very difficult, especially about the future. — Niels Bohr

Data mining as a discipline has matured considerably, and there exists a multitude of scalable algorithms that transform oceans of bits in very large databases into interpretable patterns and predictive models. At the panel on *New Research Directions in KDD* at the recent SIGKDD 2001 Conference, visionaries from academia, industry, and government debated challenges in data mining research over the next decade. Panelists were asked for the views about what they think the KDD community should (and should not) be working on, what are important problems that are being neglected, what are emerging directions, and what a decade of KDD research teaches us for the future.

Panelists included Rakesh Agrawal from IBM Almaden Research Center, Tom Mitchell from Whizbang! Labs and Carnegie Mellon University, Daryl Pregibon from AT&T Research Labs, and Ted Senator from the Defense Advanced Research Projects Agency.

Daryl started the panelists' presentations by pointing out that KDD is the synergy of three subjects: artificial intelligence, database systems, and statistics. He argued that current KDD methods are low on artificial ingelligence, and that there is a lot of interesting research to be done. Similarly, he argued that data mining research exists at a confluence of theory, algorithms, and applications, and that the current literature is low on data mining theory.

In terms of application areas, Daryl maintained that current research emphasizes "e-verything"[1] (e-commerce, e-business) too much, and that there are existing challenges in applications of data mining to scientific datasets, medicine, manufacturing, and the communications industry. Daryl also pointed out that data mining needs to strike a balance between interpretation and predictive accuracy, and that exploration of this balance is an interesting topic for further research.

Other topics mentioned by Daryl as being underrepresented is research into fundamental theories of data mining and explorations of the foundations of data mining; automation of the KDD process; and mining high-speed data streams. Daryl concluded by emphasizing the importance of the education of the next generation of data miners — both to the research community and data miners in application domains. He argued that universities need to step up to the challenge of creating courses in the foundations of KDD as a transtition to the development of whole programs in KDD.

Ted Senator opened his presentation by pointing out the importance of mining link-structured or *relational* data consisting of interrelated records. Relational data has intrinsic explicit or implicit links between individual records, and the challenge is to mine patterns and find structure in such data that exploints the existing link structure. Challenges for linked data include variations in distributions and data characteristics, non-independence of the records in training datasets, and the ability to discover structure only if the link structure between records is explicitly taken into account. Finding rare, but important, correlated events (Ted called this type of mining "'micro-mining") in data is another challenging research topic. Ted mentioned that there exist ample examples of relational data such as dna data, data about human communties, and the Web.

From the perspective of a data mining analyst with the goal to create a KDD infrastructure in an institution, Ted emphasized the importance of thinking of KDD as an overall project, not only as a process. This entails the creation of an infrastructure for KDD that permits continuous mining with repeatable results, the adaptation of models to changes in the data, and tight integration with existing data management technology. Ted also underlined the importance of research in metrics for evaluating data mining tasks, research in the foundations of data mining, and the use of background knowledge.

As candidates for problems that have received sufficient attention from the research community he mentioned algorithms for traditional problems such as association rules, clustering, and decision trees. He concluded by posing the question whether data mining should be a commodity product, and what it would take to push data mining technology towards crossing the chasm and to become ubiquituous. He predicted that there will be a flood of low-capability addons to existing products, rather than powerful stand-along KDD applications.

Tom Mitchell started his presentation by pointing out that KDD is at an exciting stage, and he illustrated KDD-external and internal trends as corroborating evidence:

- An explosion in online availability of information. One such domain is the data about individuals such as online governmental and commercial databases, and geospatial data from small sensors (such as in your cell phone). Another domain is scientific data; well-known

[1] Taken verbatim from Daryl's slides.

examples include databases with biological data, but this growth extends far beyond biology. Examples include the sloan sky survey, neurological databases, and databases with environmental information.

- A fusion of ideas from database systems, statistics and machine learning.
- Formation of educational programs in KDD throughout the world.

Tom outlined his top picks for future research topics as follows:

- Integration of human/computer data mining. (Tom called this "mix-initiative data mining".)
- New types of training data. Current research assumes that there is only labeled data (for example, classification and regression problems), or only unlabeled training data (for example, clustering and association problems). What is the space in between these two extrema?
- Mining datasets with redundant information, and using redundant information to improve the accuracy of the data mining algorithm.

Rakesh Agrawal descibed in his overview presentation a list of research problems:

1. Foundations. What is data mining, what are its foundations and its scope?
2. Privacy implications. Can we build accurate data mining models while preserving privacy of individual records? Are personalization and privacy in direct conflict?
3. Web mining. What type of analysis can we do beyond click streams analysis? Can we mine knowledge basis from the web (maybe topic specific), and what does completeness and accuracy in this setting mean? How can we deal with malicious spam that tries to subvert mining activities and intentionally produces noise in the data that is made to look like part of patterns?
4. Beyond hyperlinks. What type of social behavior (beyond hubs and authorities) exists on the web? Rakesh mentioned viral marketing as one example.
5. Actionable patterns. Most pattern learning algorithms output several orders of magnitude more patterns than humans can examine. What are the good patterns? How can we use domain knowledge to distinguish between good and bad patterns?
6. Simultaneous mining over multiple data types; examples include the integration of pictures, text, and other data sources.

Rakesh mentioned also other important problems such as online algorithms for data mining and mining data streams; long sequential patterns; algorithms for mining richter patterns (for example, tree-structured patterns); the creation of data mining benchmarks; and automatic, data-dependent selection of data mining parameters and algorithms. He suggested that applications that are most likely to benefit from new data mining technology are web applications and bioinformatics.

As inhibitors Rakesh mentioned an insufficient skill base of people who really understand the technology, and as associated problem, the difficult usage of current data mining technology.

After the presentations, several remarks from the audience added to some of the main directions discussed, such as the importance of work on foundations of data mining, and automatic data mining task and parameter selection.