Report on the SIGKDD-2002 Panel
The Perfect Data Mining Tool: Interactive or Automated?

Mihael Ankerst

The Boeing Company

P.O. Box 3707 MC 7L-70,

Seattle, WA 98124

mihael.ankerst@boeing.com

1 INTRODUCTION

Commercial and academic data mining tools range from being fully automated to highly interactive. We will discuss the role of human involvement in the data mining process. On the one hand, providing interactivity/visualization enables domain knowledge transfer and the use of the human's perceptual capabilities. On the other hand, the vast amount of data to be mined today makes real-time interactivity hard to achieve and unnecessarily burdens the user to perform tasks that may be done automatically. Questions to be discussed include: What are the ideal roles of the computer and of the user in the data mining process? Which data mining methods (clustering, classification, association rules,…) can be improved by more human involvement? What kind of applications requires more human involvement? What kind of applications requires little or no human involvement?

The panel was organized by Mihael Ankerst (The Boeing Company) and the participants were Surajit Chauduri (Microsoft Research), Georges Grinstein (University of Massachusetts Lowell & AnVil, Inc.), Jiawei Han (University of Illinois at Urbana-Champaign), and Gregory Piatetsky-Shapiro (KDnuggets).

2 THE POSITION STATEMENTS

Surajit Chauduri

Anyone who has ever used a computer knows that interactive tools are indispensable.

But, that is different from saying that generic data visualization tools are very effective for data mining. I think that while they are useful, they do not help us solve our primary challenge in mining enterprise data. Enterprise data is split across many tables and the most difficult task is specifying the query/view that defines the relevant data over which data mining is done. Specifying such queries/views requires application knowledge. Try visualizing Wal Mart's Data Warehouse to find trends without preprocessing!

Georges Grinstein

Getting rid of the human in the loop? Wrong decision!

For the last decade I have argued for an increase in human participation through visualization in the data exploration and knowledge discovery processes. At first there was resistance from the AI community. After all, computation is precise and humans imprecise. The AI and especially the KDD community later adopted a more reasonable position reminiscent of Greek geometry principles: you can use a drawing for guidance or illustration but no proof (or algorithm) can depend on a drawing. This is quite limiting. Further, the community often argues that imagery is misleading and that an algorithm can compute anything that an image shows (clusters, outliers, trends, ...). That is true. After I see a trend in an image, I can write an algorithm that identifies and discovers that trend. After I see a patterned structure or cluster in data I can write code to segregate that data into these clusters. What is the goal of data mining? It is a stage in the discovery process leading to providing knowledge for decision, most often human decision. The discover process itself consists of numerous such stages where human decision can speedup and facilitate the discovery process.

Imagine a black box capable of answering any question it is asked. Any question. Will this eliminate our need for human participation as many suggest? Quite the opposite. The fundamental problem still comes down to a human interface issue. How do I phrase the question correctly? How do I set up the parameters to get a solution that is applicable in the particular case I am interested in? How do I get the results in reasonable time and in a form that I can understand. Note that all the questions connect the discovery process to me, for my human consumption.

Thus I will again argue that visualization is necessary at all stages of the discovery process: at the front end for data awareness, understanding, massaging, and audit; at the back end for presentation of results (either in confirmatory or presentation visualization); and in the middle stages for monitoring and understanding the computational elements, an area still under-visualized.

Jiawei Han

What is an algorithm without visualizations and constraints?

My view of an attractive data mining tool is not a fully automated one but a user-friendly, interactive one, using a high-level graphical user interface to specify and control mining primitive as well as various kinds of visualization tools. The reason is that different users at different occasions may like to mine on different portions of data, for different kinds of knowledge, and with different requirements and constraints, and moreover, they would like to interactively refine their mining queries and perform drilling/dicing to progressively deepen their mining process, based on the preliminary mining results. This should be done in a highly interactive manner.

Gregory Piatetsky-Shapiro

Visualize results, not the data!

The human eye is an excellent tool for spotting natural patterns. Much progress has been made in developing very powerful visualization tools that allow many, perhaps too many, types of visualization. However, when visualization is used as a part of the data mining process to help the user spot the patterns manually, the visualization tools are frequently too complex. The most powerful visualizations are very complex to understand, burden the user with too many choices, and require long and special training.

We argue that the goal should be to simplify visualization as much as possible to help humans make more accurate decisions. Less effort should be spent on visualizing data, and more on visualizing the results of data mining and helping the users to understand them. The visualization tools should not confuse an average user with an overwhelming set of choices for visualization. Instead, they should guide the user towards the most appropriate visualizations for the task. A possible long-term goal could be to get rid of visualization altogether and to automate the decision process.

3 THE PANEL DISCUSSION

Some statements from the slides of the panelists:

Surajit Chauduri

- Advanced metaphors of visualization are over-rated. Many visualizations are too complex with many theorists and too few practitioners.
- Visualization is not a solution for auto-parameterization of algorithms.
- There are far more interesting and high-impact challenges for us in KDD (e.g. mining enterprise data warehouses, “row” and “column” extractors for data reduction)

Georges Grinstein

- Current state-of-the-art DM tools are automated, but the perfect DM tool is highly interactive and participatory.
- Some tools require no interaction or very little (real time decision systems, manufacturing monitoring), some tools require a tremendous amount of interaction (protein function determination).
- In the ideal, automation is where we’re heading but interaction is a phase through which we must pass but which we always need in complex situations. All DM methods should contain more visualization.

Jiawei Han

- The degree of interaction depends on the stage of mining.
- Data selection and viewing of mining results should be fully interactive, the mining process should be more interactive than the current state-of-the art and embedded application should be fairly automated.
- Routine mining in finance, insurance, manufacturing may need more automation, whereas fraud detection, CRM may need more interaction.

Gregory Piatetsky-Shapiro

- The perfect DM tool should have a more automated mode for beginners and a more interactive mode for experts.
- Automate as much as possible but not more.
- More interaction is needed when sample applications are more varied (consumer data). More automation is needed when sample applications are more similar (science, bioinformatics, manufacturing).