From: KDD-96 Proceedings. Copyright © 1996, AAAI (www.aaai.org). All rights reserved.

Harnessing the Human in Knowledge Discovery

Georges G. Grinstein

University of Massachusetts at Lowell
Institute for Visualization and Perception Research
Lowell MA
and
The MITRE Corporation
Bedford MA
ggg@mitre.org or grinstein@cs.uml.edu

Knowledge discovery is the process of discovering interesting, non-trivial patterns in data [1]. In the sub-field called knowledge discovery in databases (KDD) the discovery process targets data repositories, and often includes metrics on the results it has achieved, measuring how good the discoveries are with respect to, for example, non-trivialness, novelty, or extent.

Knowledge, the primary goal of data analysis and exploration, is most often discovered by generating information (structure) from data, and then abstracting non-trivial patterns (rules or associations for example) from the information. The discovery process can be done using numerous means that share the same goal: visualization, data mining, statistics, neural networks, or mathematical modeling and simulation [2].

Visualization Approaches

Visualization is different from the rest, however, in that it is also the actual mechanism by which the analyses and their results can be presented to the user. Visualization, in other words, harnesses the perceptual and cognitive capabilities of the human user, who is still the most powerful pattern recognizer and inference engine.

Visualizations can be divided into three classes [3, 4]]: exploratory, confirmatory or production. Exploratory visualization is dynamic and relatively unpredictable. The user typically does not know what to look for, or has minimal direction. The emphasis is on organizing, testing, developing concepts, looking for trends, and defining hypotheses.

In confirmatory visualization, on the other hand, the user has some sense of a goal, or some hypothesis to be confirmed. The visualization process is more stable and predictable. The user often selects predetermined system parameters, and typically is looking to confirm or refute hypotheses.

Production visualization is the most stable and predictable of the visualization approaches. Typically the user already has a validated hypothesis, and is looking to, for example, display data to emphasize a particular point. System parameters are set, but require fine tuning, perhaps using color map selections or layout formats.

Just as there are three classes of visualizations, there are three steps to the knowledge discovery process. The user initially defines concepts of interest and uses them to define domain structure in the data. Then one invokes algorithms that use the concepts to mine the data for non-trivial patterns. Finally the results are presented to the user, and the process is iterated.

Visualization in Knowledge Discovery

Where and how, in these three steps of knowledge discovery, can visualization be used?

In the first step, preliminary concepts or initial key points that need to be defined are typically presented as text files for the user to select and refine. In this initial part of the process today's user has few visualizations or presentations to call upon. This is therefore an open field for exploration. How might we visually represent concepts in the early selection stage, or visually have users define such concepts for interaction?

Successful presentations at this stage will help the user formalize concepts, and will make it easier to focus the analytic tools.

In the second and third steps of knowledge discovery, the user ideally interacts continually with the selected information-setting parameters, fine-tuning, sub-selecting, and so on-based on examination of the preliminary information presented or visualized by the analytic tools. This iterative activity is fundamental to the discovery process, whether in KDD, or statistics, or modeling and simulation. And it offers further opportunities for innovative visualization technique. While many tools today summarize information visually, few offer the kind of intermediary visualizations that could reduce the number of iterations necessary to reach a result. We have to ask ourselves, then, how one can visually present the resulting datasets and summaries.

Visualization of databases, and more generally datasets, is in its infancy. Best known are statistical representations of data, such as statistical computations plotted in a variety of ways: histograms, time series, scatterplots. Modern approaches have extended the scatterplot concept. Whereas the pixel is a visual object whose representation on the screen is driven by three values (RGB or HLS, for example), an icon may be driven by an arbitrary number of values, permitting presentation of icons en masse on the screen [5]. Such generalizations enable us to display larger, higher-dimensional datasets.

It is interactive displays of this kind that will provide the mechanism, in effect the breakthrough necessary, to harness our human perceptual and cognitive capabilities. The user will then finally be able to interact visually with the information, and thus more fully participate with all the numerous components and steps of the knowledge discovery process.

Presentation

We will present a brief history of alternative visualizations and how they have been applied to various data visualization problems. The emphasis will be on how visualization, in particular exploratory visualization, can support the knowledge discovery process, including concept development for database management, database visualizations, and minimally structured dataset visualizations.

References

[1] Frawley, W.J., Piatetsky-Shapiro, G. and Matheus, C.J., 1991, Knowledge Discovery in Databases: An Overview. In Knowledge Discovery in Databases, Piatetsky-Shapiro and Frawley, W.J., Editors, AAAI/MIT Press, Vol 19, pp 1-27.

[2] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., 1996, Editors, Advances in Knowledge Discovery and Data Mining, AAAI Press.

[3] Grinstein, G., and John P. Lee, 1996, Describing Visual Interactions to the Database: Closing the Loop Between Users and Data, Proceedings of the Second SPIE'96 Visual Data Exploration and Analysis Conference, San Jose, pp 93-103.

[4] Lee, J.P. and Grinstein, G., 1995, An Architecture for Retaining and Analyzing Visual Explorations of Databases, 1995 IEEE Visualization Conference Proceedings, Nielson and Silver (Editors), pp 101-108.

[5] Erbacher, R., Grinstein, G., Levkowitz, H., Masterman, L., Pickett, R. , Smith, S., 1995, Exploratory Visualization Research at the University of Massachusetts at Lowell, Computers and Graphics Journal, Special Issue on Visual Computing, Vol 19, No 1, pp 131-139.