Invited talk at the *Sanken Symposium on Data Mining and Semantic Web*,
Osaka University, Japan, March 10-11, 2003

Knowledge Mining: A Proposed New Direction

Ryszard S. Michalski
Machine Learning and Inference Laboratory
School of Computational Sciences
George Mason University
and
Institute for Computer Science
Polish Academy of Sciences
Email: michalski@mli.gmu.edu

Summary

In the last several years, the field of data mining has been rapidly expanding, and attracting many new researchers and users. The underlying reason for such a rapid growth is a great need for systems that can automatically derive useful knowledge from vast volumes of computer data being accumulated worldwide. The field of data mining offers a promise for addressing this need. The major trust of research has been to develop a repertoire of tools for discovering both strong and useful patterns in large databases. The function performed by such tools can be succinctly characterized as a mapping:

$$\text{DATA} \rightarrow \text{PATTERNS} \qquad (1)$$

An underlying assumption is that the patterns are created solely from the data, and thus are expressed in terms of attributes and relations appearing in the data. Determining such patterns can be a problem of significant computational complexity, but of a relatively low conceptual complexity, and many efficient algorithms have been developed for this purpose (e.g., Breiman et al., 1984; Quinlan, 1993, Agrawal et al., 1996; Witten, Moffatt and Bell, 1999). This approach to the problem of deriving useful knowledge from databases has, however, some fundamental limitations, and new research should address several important tasks.

To assure that patterns are not only strong (i.e., represent frequently occurring relationships), but also useful to a specific user or group of users, the system cannot rely solely on the data, but must be able also to represent and understand user's goals. This requires a method for goal representation. To be able to derive patterns that are not only combinatorial combinations of concepts (attributes and relations) that are already present in the data, the system needs background knowledge that will allow it to reinterpret and/or combine concepts in the data into new concepts that can lead to more accurate and/or simpler patterns. To be able to re-use patterns determined in the previous analyses in the process of updating them in view of new data, the system needs to have a capability for incremental knowledge-based pattern discovery. The above-listed requirements create a

major new challenge: to integrating a knowledge base within a data mining system, and to develop methods for applying this knowledge during data mining.

Since there is a vast array of different tasks for which knowledge generated from data can be used, there are many different knowledge needs. Therefore, a data mining system has to use advanced knowledge representations and be able to generate many different types of knowledge from a given data source. This problem is being partially addressed by the growing inventory of available data mining programs. These programs are, however, often arranged into toolboxes, and individuals programs have to be manually invoked. Using such toolboxes can, therefore, be a very laborious and time consuming process, and may require considerable expertise. This problem is being partially addressed by the development of multistrategy data mining systems that integrate different data mining tools (e.g., Morik and Brockhausen, 1966). To automate further a data mining process, such tools need to be invokable through a high-level *knowledge generation language* (e.g., Michalski and Kaufman, 1998, 2000b). Since users want to understand data mining results, an important research direction is also the development *knowledge* visualization methods (e.g., Cervone and Michalski, 2003).

To address the research direction that aims at achieving all the above-mentioned tasks, we use the term *knowledge mining.* Knowledge mining can thus be characterized as concerned with developing and integrating a wide range of data analysis methods that are able to derive directly or incrementally new knowledge from large (or small) volumes of data using relevant prior knowledge. The process of deriving new knowledge has to be guided by criteria inputted to the system defining the type of knowledge a particular user is interested in. Algorithms for generating new knowledge must be not only efficient but also oriented toward producing knowledge satisfying the *comprehensibility postulate*, that is, easy to understand and interpret by the users (Michalski, 1983). Knowledge mining can be simply characterized by the following mapping:

$$\text{DATA} + \text{PRIOR\_KNOWLEDGE} + \text{GOAL} \rightarrow \text{NEW\_KNOWLEDGE} \qquad (2)$$

where GOAL is an encoding of the knowledge needs of the user(s), and NEW_KNOWLEDGE is knowledge satisfying the GOAL. Such knowledge can be in the form of decision rules, association rules, decision trees, conceptual or similarity-based clusters, equations, Bayesian nets, statistical summaries, visualizations, natural language summaries, or other knowledge representations.

The current research in the GMU Machine Learning Laboratory is concerned with developing the VINLEN system that aims of addressing the goals mentioned above. It is a multistrategy inductive database system for knowledge mining. While there are other efforts on developing systems under the name “inductive databases” (e.g., Mannila, 1997), VINLEN has many unique features. It aims at integrating a great variety of methods for pattern discovery, theory formation, clustering, data selection, statistical analysis, data and knowledge visualization, and others. A general VINLEN schema is presented in Fig. 1.

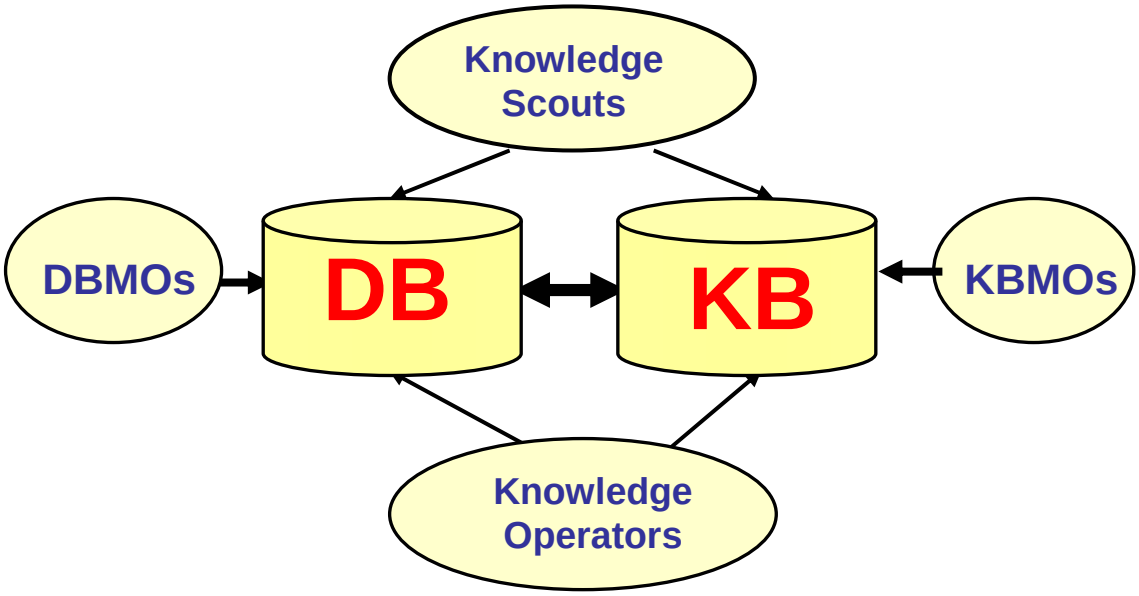

Figure 1. A general schema of the VINLEN inductive database.

DBMOs are conventional database management operators, and KBMOs are operators for managing knowledge representations in the knowledge base (KB). Knowledge operators invoke a range of knowledge mining programs, such as attributional rule learning in pattern discovery and theory formation modes, incremental attributional rule improvement, multi-head rule discovery, intelligent data generation, attribute selection, selection of the most representative datapoints, conceptual clustering, knowledge visualization via concept association graphs and generalized logic diagrams. These operators can be invoked individually, or via scripts in *Knowledge Query Language (KQL).* Such scripts, called *knowledge scouts*, can perform complex knowledge mining processes in the search for *target knowledge*—knowledge of interest to the given user. Knowledge base contains general purpose and domain knowledge, as well as knowledge generated in the previous data exploration experiments. The basic form of knowledge representation in VINLEN employs *attributional calculus*, a logic system that combines aspects of propositional logic, predicate logic, and many-valued logic (Michalski, 2003). The main module for pattern discovery and theory formation employs AQ learning (e.g., Kaufman and Michalski, 2000a).

The development of VINLEN is major research and implementational task, and will take a considerable amount of time to complete. Some of the modules have already been implemented and partially integrated into the system (e.g., attributional rule learning in pattern discovery and theory formation modes, intelligent target data generation, conceptual clusterer, general graphical user interface with parallel access to modules). Preliminary results from their application to selected practical problems will be presented in the talk.

Acknowledgments

The author thanks Ken Kaufman for comments and collaboration on the development of VINLEN, and Mike Draminski for his contribution to the VINLEN implementation. This research was performed at the Machine Learning and Inference Laboratory at George Mason University. Laboratory's research activities are supported in part by the National Science Foundation under Grants No. IIS-9906858 and IIS-0097476, and in part by the UMBC/LUCITE #32 grant.

References

Agrawal, R., Mannila, H., Srikan R., Toivonen, H., and Verkamo, A. I., Fast Discovery of Association Rules, in *Advances in Knowledge Discovery and Data Mining*, U.M. Fayyad, G. Piatetsky-Shapiro, P., Smyth, and Uthurasamy, R. (eds.), AAAI Press, Menlo Park, CA. 1996.

Breiman, L., Friedman, J.H., Olshen, R.S., and Stone, C.J., Classifications and Regression Trees, Wadsworth Statistical Press, Belmont, CA, 1984;

Cervone, G. and Michalski, CAG1—A Program for Visualizing Concept Association Graphs, *Reports of Machine Learning and Inference Laboratory*, George Mason University, 2003, *to appear*.

Kaufman, K.A. and Michalski, R.S., "An Adjustable Rule Learner for Pattern Discovery Using the AQ Methodology," *Journal of Intelligent Information Systems*, 14, pp. 199-216, 2000a.

Kaufman, K.A. and Michalski, R.S., "A Knowledge Scout for Discovering Medical Patterns: Methodology and System SCAMP," *Proceedings of the Fourth International Conference on Flexible Query Answering Systems (FQAS'2000)*, Warsaw, pp. 485-496, 2000b.

Mannila, H., "Inductive Databases and Condensed Representations for Data Mining," *Proceedings of the International Logic Programming Symposium (ILPS'97)*, pp. 21-30, 1997.

Michalski, R. S., A Theory and Methodology of Inductive Learning, in *Machine Learning: An Artificial Intelligence Approach*, R. S. Michalski, J. Carbonell and T. Mitchell (Eds.), pp. 83-134, Morgan Kaufman Publishing Co., Palo Alto, 1983.

Michalski, R.S. and Kaufman, K., "Building Knowledge Scouts Using KGL Metalanguage," *Fundamenta Informaticae*, 40, pp. 433-447, 2000.

Michalski, R.S. and Kaufman, K.A., "Data Mining and Knowledge Discovery: A Review of Issues and a Multistrategy Approach," In *Machine Learning and Data Mining: Methods and Applications*, Michalski, R.S., Bratko, I. and Kubat, M. (eds.), London, John Wiley & Sons, pp. 71-112, 1998.

Michalski, R.S., "Attributional Calculus: A Representation System and Logic for Natural Induction," *Reports of the Machine Learning and Inference Laboratory*, MLI 03-01, George Mason University, 2003.

Morik, K. and Brockhausen, P., "A Multistrategy Approach to Relational Knowledge Discovery in Databases," *Proceedings of the Third International Workshop on Multistrategy Learning* (MSL-96), pp. 17-27, 1996.

Witten, I. H., Moffat, A., and Bell, T.C., Managing Gigabytes: Compressing and Indexing Documents and Images, 2nd ed., Morgan Kaufman, San Francisco, CA, 1999.