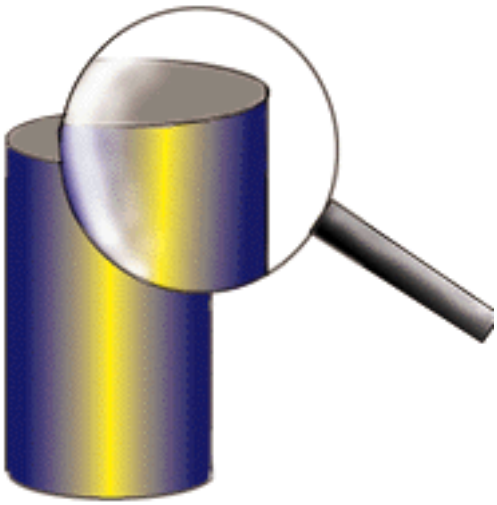
Knowledge Discovery In Databases: Tools and Techniques

by Peggy Wright
This work is funded by
U.S. Army Corps Engineers
Waterways Experiment Station
Vicksburg, MS 39180



Introduction

The amount of data being collected in databases today far exceeds our ability to reduce and analyze data without the use of automated analysis techniques. Many scientific and transactional business databases grow at a phenomenal rate. A single system, the astronomical survey application SCICAT, is expected to exceed three terabytes of data at completion [4]. Knowledge discovery in databases (KDD) is the field that is evolving to provide automated analysis solutions.

Knowledge discovery is defined as ``the non-trivial extraction of implicit, unknown, and potentially useful information from data" [6]. In [5], a clear distinction between data mining and knowledge discovery is drawn. Under their conventions, the knowledge discovery process takes the raw results from **data mining** (the process of extracting trends or patterns from data) and carefully and accurately transforms them into useful and understandable information. This information is not typically retrievable by standard techniques but is uncovered through the use of AI techniques.

KDD is a growing field: There are many knowledge discovery methodologies in use and under development. Some of these techniques are generic, while others are domain-specific. The purpose of this paper is to present the results of a literature survey outlining the state-of-the-art in KDD techniques and tools. The paper is not intended to provide an in-depth introduction to each approach; rather, we intend it to acquaint the reader with some KDD approaches and potential uses.

Background

Although there are many approaches to KDD, six common and essential elements qualify each as a

knowledge discovery technique. The following are basic features that all KDD techniques share (adapted from [5] and [6]):

- All approaches deal with large amounts of data
- Efficiency is required due to volume of data
- Accuracy is an essential element
- All require the use of a high-level language
- All approaches use some form of automated learning
- All produce some interesting results

Large amounts of data are required to provide sufficient information to derive additional knowledge. Since large amounts of data are required, processing efficiency is essential. Accuracy is required to assure that discovered knowledge is valid. The results should be presented in a manner that is understandable by humans. One of the major premises of KDD is that the knowledge is discovered using intelligent learning techniques that sift through the data in an automated process. For this technique to be considered useful in terms of knowledge discovery, the discovered knowledge must be interesting; that is, it must have potential value to the user.

KDD provides the capability to discover new and meaningful information by using existing data. KDD quickly exceeds the human capacity to analyze large data sets. The amount of data that requires processing and analysis in a large database exceeds human capabilities, and the difficulty of accurately transforming raw data into knowledge surpasses the limits of traditional databases. Therefore, the full utilization of stored data depends on the use of knowledge discovery techniques.

The usefulness of future applications of KDD is far-reaching. KDD may be used as a means of information retrieval, in the same manner that intelligent agents perform information retrieval on the web. New patterns or trends in data may be discovered using these techniques. KDD may also be used as a basis for the intelligent interfaces of tomorrow, by adding a knowledge discovery component to a database engine or by integrating KDD with spreadsheets and visualizations.

KDD Techniques

Learning algorithms are an integral part of KDD. Learning techniques may be supervised or unsupervised. In general, supervised learning techniques enjoy a better success rate as defined in terms of usefulness of discovered knowledge. According to [1], learning algorithms are complex and generally considered the hardest part of any KDD technique.

Machine discovery is one of the earliest fields that has contributed to KDD [5]. While machine discovery relies solely on an autonomous approach to information discovery, KDD typically combines automated approaches with human interaction to assure accurate, useful, and understandable results.

There are many different approaches that are classified as KDD techniques. There are quantitative approaches, such as the probabilistic and statistical approaches. There are approaches that utilize visualization techniques. There are classification approaches such as Bayesian classification, inductive logic, data cleaning/pattern discovery, and decision tree analysis. Other approaches include deviation and trend analysis, genetic algorithms, neural networks, and hybrid approaches that combine two or more techniques.

Because of the ways that these techniques can be used and combined, there is a lack of agreement on how these techniques should be categorized. For example, the Bayesian approach may be logically grouped with probabilistic approaches, classification approaches, or visualization approaches. For the sake of organization, each approach described here is included in the group that it seemed to fit best. However, this selection is not intended to imply a strict categorization.

Probabilistic Approach

This family of KDD techniques utilizes graphical representation models to compare different knowledge representations. These models are based on probabilities and data independencies. They are useful for applications involving uncertainty and applications structured such that a probability may be assigned to each ``outcome'' or bit of discovered knowledge. Probabilistic techniques may be used in diagnostic systems and in planning and control systems [2]. Automated probabilistic tools are available both commercially and in the public domain.

Statistical Approach

The statistical approach uses rule discovery and is based on data relationships. An ``inductive learning algorithm can automatically select useful join paths and attributes to construct rules from a database with many relations'' [8]. This type of induction is used to generalize patterns in the data and to construct rules from the noted patterns. Online analytical processing (OLAP) is an example of a statistically-oriented approach. Automated statistical tools are available both commercially and in the public domain.

An example of a statistical application is determining that all transactions in a sales database that start with a specified transaction code are cash sales. The system would note that of all the transactions in the database only 60% are cash sales. Therefore, the system may accurately conclude that 40% are collectibles.

Classification Approach

Classification is probably the oldest and most widely-used of all the KDD approaches [11]. This approach groups data according to similarities or classes. There are many types of classification techniques and numerous automated tools available.

The **Bayesian Approach** to KDD ``is a graphical model that uses directed arcs exclusively to form an [sic] directed acyclic graph" [2]. Although the Bayesian approach uses probabilities and a graphical means of representation, it is also considered a type of classification.

Bayesian networks are typically used when the uncertainty associated with an outcome can be expressed in terms of a probability. This approach relies on encoded domain knowledge and has been used for diagnostic systems. Other pattern recognition applications, including the Hidden Markov Model, can be modeled using a Bayesian approach [3]. Automated tools are available both commercially and in the public domain.

Pattern Discovery and Data Cleaning is another type of classification that systematically reduces a large database to a few pertinent and informative records [7]. If redundant and uninteresting data is eliminated, the task of discovering patterns in the data is simplified. This approach works on the premise of the old adage, ``less is more". The pattern discovery and data cleaning techniques are useful for reducing enormous volumes of application data, such as those encountered when analyzing automated sensor recordings. Once the sensor readings are reduced to a manageable size using a data cleaning technique, the patterns in the data may be more easily recognized. Automated tools using these techniques are available both commercially and in the public domain.

The **Decision Tree Approach** uses production rules, builds a directed acyclical graph based on data premises, and classifies data according to its attributes. This method requires that data classes are discrete and predefined [11]. According to [5], the primary use of this approach is for predictive models that may be appropriate for either classification or regression techniques. Tools for decision tree analysis are available commercially and in the public domain.

Deviation and Trend Analysis

Pattern detection by filtering important trends is the basis for this KDD approach. Deviation and trend analysis techniques are normally applied to temporal databases. A good application for this type of KDD is the analysis of traffic on large telecommunications networks.

AT&T uses such a system to locate and identify circuits that exhibit deviation (faulty behavior) [12]. The sheer volume of data requiring analysis makes an automated technique imperative. Trend-type analysis might also prove useful for astronomical and oceanographic data, as they are time-based and voluminous. Public domain tools are available for this approach.

Other Approaches

Neural networks may be used as a method of knowledge discovery. Neural networks are particularly useful for pattern recognition, and are sometimes grouped with the classification approaches. There are tools available in the public domain and commercially. Genetic algorithms, also used for classification,

are similar to neural networks although they are typically considered more powerful. There are tools for the genetic approach available commercially.

Hybrid Approach

A hybrid approach to KDD combines more than one approach and is also called a multi-paradigmatic approach. Although implementation may be more difficult, hybrid tools are able to combine the strengths of various approaches. Some of the commonly used methods combine visualization techniques, induction, neural networks, and rule-based systems to achieve the desired knowledge discovery. Deductive databases and genetic algorithms have also been used in hybrid approaches. There are hybrid tools available commercially and in the public domain.

Conclusions and Future Directions

KDD is a rapidly expanding field with promise for great applicability. Knowledge discovery purports to be the new database technology for the coming years. The need for automated discovery tools had caused an explosion in the number and type of tools available commercially and in the public domain. The S*i*ftware web site [9] is updated frequently and is intended to be an exhaustive listing of currently available KDD tools.

It is anticipated that commercial database systems of the future will include KDD capabilities in the form of intelligent database interfaces. Some types of information retrieval may benefit from the use of KDD techniques. Due to the potential applicability of knowledge discovery in so many diverse areas there are growing research opportunities in this field. Many of these opportunities are discussed in [10], a newsletter which has regular contributions from many of the best-known authors of KDD literature. A fairly comprehensive list of references and applicable websites are also available from the Nugget site. These sites are updated very frequently and have the most current information available. An international conference on KDD is held annually. The annual KDD conference proceedings provide additional sources of current and relevant information on the growing field of Knowledge Discovery in Databases.

References

1

Brachman, R.J., and Anand, T. The Process Of Knowledge Discovery In Databases: A Human-Centered Approach. In *Advances In Knowledge Discovery And Data Mining* , eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA., 1996, pp. 37-57.

2

Buntine, W. Graphical Models For Discovering Knowledge. In *Advances In Knowledge Discovery And Data Mining*, eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R.

Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA., 1996, pp. 59-82.

3

Buntine, W. "A Guide To The Literature On Learning Probabilistic Networks From Data." *IEEE Transactions on Knowledge and Data Engineering* 8, 2 (Apr. 1996), 195-210.

4

Fayyad, U.M., Djorgovski, S.G., and Weir, N. Automating The Analysis And Cataloging Of Sky Surveys. In *Advances In Knowledge Discovery And Data Mining* , eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA., 1996, pp. 472-493.

5

Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. From Data Mining To Knowledge Discovery: An Overview. In *Advances In Knowledge Discovery And Data Mining* , eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA., 1996, pp. 1-34.

6

6. Frawley, W.J., Piatetsky-Shapiro, G., and Matheus, C. Knowledge Discovery In Databases: An Overview. In *Knowledge Discovery In Databases*, eds. G. Piatetsky-Shapiro, and W. J. Frawley, AAAI Press/MIT Press, Cambridge, MA., 1991, pp. 1-30.

7

Guyon, I., Matic, N., and Vapnik, V. Discovering Informative Patterns And Data Cleaning. In *Advances In Knowledge Discovery And Data Mining*, eds. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA., 1996, pp. 181-203.

8

Hsu, C.N., and Knoblock, C.A. Using Inductive Learning To Generate Rules For Semantic Query Optimization. In *Advances In Knowledge Discovery And Data Mining*, eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA., 1996, pp. 425-445.

9

Piatetsky-Shapiro, G. *S*i*ftware: Tools For Data Mining And Knowledge Discovery*. World Wide Web URL: http://www.kdnuggets.com/siftware.html.

10

Piatetsky-Shapiro, G., and Beddows, M. *Knowledge Discovery Mine -- Data Mining And Knowledge Discovery Resources.* World Wide Web URL:http://www.kdnuggets.com.

11

Quinlan, J.R. *C4.5: Programs For Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

12

Sasisekharan, R., Seshadri, V., and Weiss, S.M. Data Mining And Forecasting In Large-Scale Telecommunication Networks. *IEEE Expert: Intelligent Systems & Their Applications* 11, 1 (Feb. 1996), 37-43.