Proceedings of national Seminar on Future Trends in Data Mining (NSFTDM-2010):-10th may, 2010, Organenised by
Department of Computer Science, Gandhi Engineering college, Bhubaneswar.

# Data Mining & Knowledge Discovery in Databases: An AI Perspective

Arabinda Nanda[1] Saroj Kumar Rout[2]

[1]Department of Computer science & Engineering, Krupajal Engineering College, Bhubaneswar, Orissa, Email:aru.nanda@rediffmail.com
[2]Department of Computer science & Engineering, Krupajal Engineering College, Bhubaneswar, Orissa, Email:rout_sarojkumar@yahoo.co.in

## *Abstract*

**Data mining and Knowledge discovery has several important application areas. Data mining and knowledge discovery have been topics considered at many AI, database and statistical conferences. Knowledge discovery generally refers to the process of identifying valid, novel and understandable patterns. Knowledge discovery from large databases, often called data mining, refers to the application of the discovery process on large databases or datasets. The discovery process can be broken into several steps, including: developing an understanding of the application domain; creating a target data set; data cleaning and preprocessing; finding useful features with which to represent the data; data mining to search for patterns of interest; and interpreting and consolidating discovered patterns. Data mining and knowledge discovery in databases have been attracting a significant amount of research, industry, and media attention of late. What is all the excitement about? This article provides an overview of this emerging field, clarifying how data mining and knowledge discovery in databases are related both to each other and to related fields, such as machine learning, statistics, and databases. The article mentions particular real-world applications, specific data-mining techniques, challenges involved in real-world applications of knowledge discovery, and current and future research directions in the field.**

***Key words***: Data mining, Knowledge database, KDD

## Introduction

Data is the raw fact. Processed data is called information. Fact of knowing about the world is called knowledge.Ex:- Cotton Produces Cloth .Cotton is the raw fact is called data. Produces by using some machine i.e. data in processing state is called information. The final output cloth is the knowledge. Knowledge is closely related with Intelligence. A person having more knowledge is called highly intelligence person. Data store in a data base where as knowledge store in a knowledge base. Across a wide diversity of fields, data are being collected and accumulated at a remarkable speed. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD). At an abstract level, the KDD field is concerned with the development of methods and techniques for making sense of data. The basic problem addressed by the KDD process is one of mapping low-level data (which are typically too voluminous to understand and digest easily) into other forms that might be more compact (for example, a short report), more abstract (for example, a descriptive approximation or model of the process that generated the data), or more useful (for example, a predictive model for estimating the value of future cases). At the core of the process is the application of specific data-mining methods for pattern discovery and extraction.

## Data Mining

Data mining is a logical process that is used to search through large amounts of information in order to find important data. The goal of this technique is to find patterns that were previously unknown. Once you have found these patterns, you can use them to solve a number of problems. Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is a powerful tool because it can provide you with relevant information that you can use to your own advantage. When you have the right knowledge, all you will need to do is apply it in the right manner, and you will be able to benefit. It is relatively easy to get information these days. But it is not so easy to get relevant information that can help you achieve a desired goal. This is where data mining becomes a powerful tool that you will want to become familiar with. It will give you the power to predict certain behaviors within a system.

Data mining has been defined in almost as many ways as there are authors who have written about it. Because it sits at the interface between statics, computer science, artificial intelligence, machine learning, database management and data visualization, the definition changes with the perspective of the user:

“Data Mining is the process of exploration and analysis by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.”
(M.J.A Berry and G.S Linoff).

“Data Mining is finding interesting structure (patterns, statical models, relationships) in database.” (U. Fayyad, S. Chaudhuri and P. Bradley)

“Data Mining is the application of statistics in the form of exploratory data analysis and predictive models to reveal patterns and trends in very large data sets.” (“Insightful Minor 3.0 user guide”).

## Need of KDD

The traditional method of turning data into knowledge relies on manual analysis and interpretation. For example, in the health-care industry, it is common for specialists to periodically analyze current trends and changes in health-care data, say, on a quarterly basis. The specialists then provide a report detailing the analysis to the sponsoring health-care organization; this report becomes the basis for future decision making and planning for health-care management. In a totally different type of application, planetary geologists sift through remotely sensed images of planets and asteroids, carefully locating and cataloging such geologic objects of interest as impact craters. Be it science, marketing, finance, health care, retail, or any other field, the classical approach to data analysis relies fundamentally on one or more analysts becoming intimately familiar with the data and serving as an interface between the data and the users and products.

## Data Mining and Knowledge Discovery in the Real World

A large degree of the current interest in KDD is the result of the media interest surrounding successful KDD applications, for example, the focus articles within the last two years in *Business Week*, *Newsweek*, *Byte*, *PC Week*, and other large-circulation periodicals. Unfortunately, it is not always easy to separate fact from media hype. Nonetheless, several well documented examples of successful systems can rightly be referred to as KDD applications and have been deployed in operational use on large-scale real-world problems in science and in business.

In science, one of the primary application areas is astronomy. In business, main KDD application areas includes marketing, finance (especially investment), fraud detection, manufacturing, telecommunications, and Internet agents.

## Data Mining and KDD

Historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. The term *data mining* has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities. It has also gained popularity in the database field. The phrase *knowledge discovery in databases* was coined at the first KDD workshop in 1989 (Piatetsky-Shapiro 1991) to emphasize that knowledge is the end product of a data-driven discovery. It has been popularized in the AI and machine-learning fields. In our view, KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. *Data mining* is the application of specific algorithms for extracting patterns from data. The distinction between the KDD process and the data-mining step (within the process) is a central point of this article. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data. Blind application of data-mining methods (rightly criticized as data dredging in the statistical literature) can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns.

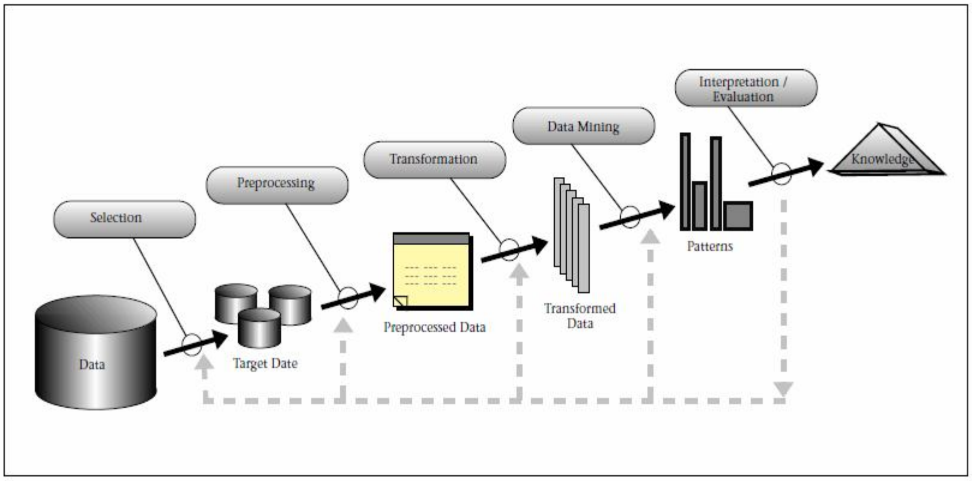

*Figure 1. An Overview of the Steps That Compose the KDD Process.*

## Concluding Remarks: The Potential Role of AI in KDD

In addition to machine learning, other AI fields can potentially contribute significantly to various aspects of the KDD process. We mention a few examples of these areas here:

**Natural language** presents significant opportunities for mining in free-form text, especially for automated annotation and indexing prior to classification of text corpora. Limited parsing capabilities can help substantially in the task of deciding what an article refers to. Hence, the spectrum from simple natural language processing all the way to language understanding can help substantially. Also, natural language processing can contribute significantly as an effective interface for stating hints to mining algorithms and visualizing and explaining knowledge derived by a KDD system.

**Planning** considers a complicated data analysis process. It involves conducting complicated data-access and data-transformation operations; applying preprocessing routines; and, in some cases, paying attention to resource and data-access constraints.

Typically, data processing steps are expressed in terms of desired post conditions and preconditions for the application of certain routines, which lends itself easily to representation as a planning problem. In addition, planning ability can play an important role in automated agents (see next item) to collect data samples or conduct a search to obtain needed data sets.

**Intelligent agents** can be fired off to collect necessary information from a variety of Sources. In addition, information agents can be activated remotely over the network or can trigger on the occurrence of a certain event and start an analysis operation. Finally, agents can help navigate and model the World-Wide Web (Etzioni 1996), another area growing in importance.

**Uncertainty in AI** includes issues for managing uncertainty, proper inference mechanisms in the presence of uncertainty, and the reasoning about causality, all fundamental to KDD theory and practice. In fact, the KDD-96 conference had a joint session with the UAI-96 conference this year (Horvitz and Jensen 1996).

**Knowledge representation** includes *ontologies,* new concepts for representing, storing, and accessing knowledge. Also included are schemes for representing knowledge and allowing the use of prior human knowledge about the underlying process by the KDD System. These potential contributions of AI are but a sampling; many others, including human computer interaction, knowledge-acquisition techniques, and the study of mechanisms for reasoning, have the opportunity to contribute to KDD.In conclusion, we presented some definitions of basic notions in the KDD field. Our primary aim was to clarify the relation between knowledge discovery and data mining. We provided an overview of the KDD process and basic data-mining methods. Given the broad spectrum of data-mining methods and algorithms, our overview is inevitably limited in scope: There are many data-mining techniques, particularly specialized methods for particular types of data and domain. Although various algorithms and applications might appear quite different on the surface, it is not uncommon to find that they share many common components. Understanding data mining and model induction at this component level clarifies the task of any data mining algorithm and makes it easier for the user to understand its overall contribution and applicability to the KDD process. This article represents a step toward a common framework that we hope will ultimately provide a unifying vision of the common overall goals and methods used in KDD. We hope this will eventually lead to a better understanding of the variety of approaches in this multidisciplinary field and how they fit together.

# References

Agrawal, R., and Psaila, G. 1995. Active Data Mining. In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), 3–8. Menlo Park, Calif.: American Association for Artificial Intelligence.

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; and Verkamo, I. 1996. Fast Discovery of Association Rules. In *Advances in Knowledge Discovery and Data Mining*, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 307–328. Menlo Park,Calif.: AAAI Press.

Apte, C., and Hong, S. J. 1996. Predicting Equity Returns from Securities Data with Minimal Rule Generation. In *Advances in Knowledge Discovery and Data Mining*, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 514–560. Menlo Park, Calif.: AAAI Press.

Basseville, M., and Nikiforov, I. V. 1993. *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs, N.J.: Prentice Hall.

Berndt, D., and Clifford, J. 1996. Finding Patterns in Time Series: A Dynamic Programming Approach. In *Advances in Knowledge Discovery and Data Mining*,eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 229–248. Menlo Park, Calif.: AAAI Press.

Berry, J. 1994. Database Marketing. *Business Week*, September 5, 56–62.

Brachman, R., and Anand, T. 1996. The Process of Knowledge Discovery in Databases: A Human-Centered Approach. In *Advances in Knowledge Discovery and Data Mining*, 37–58, eds. U. Fayyad, G. Piatetsky- Shapiro, P. Smyth, and R. Uthurusamy. Menlo Park, Calif.: AAAI Press.