UDC: 007.5
Review article

KNOWLEDGE DISCOVERY IN DATABASES: A COMPARISON OF DIFFERENT VIEWS

Eva Andrássyová

Department of Cybernetics and Artificial Intelligence, Technical University of Košice, Slovakia
E-mail: andrassy@tuke.sk

Ján Paralič

Department of Cybernetics and Artificial Intelligence, Technical University of Košice, Slovakia
E-mail: paralic@tuke.sk

The field of knowledge discovery in databases (KDD) is becoming very popular and it has grown quite a lot recently. The large amounts of data collected and stored may contain some information, which could be useful, but it is not easy to recognise it, nor is it trivial to obtain it. There is no human capable of sifting through such large amounts of data and even some of the existing algorithms are inefficient when trying to solve this task. KDD systems incorporate techniques from a large variety of related fields to utilise their strengths in the process of discovering knowledge.

Whilst working on the international GOAL[1] project (Geographic Information On-Line Analysis: GIS - Data Warehouse Integration) we have studied several publications to get an idea of what the KDD (process) is and also an idea of what it is not. We have studied those techniques that are applicable in this process, what tasks are to be solved and which particular steps the process should take. The interdisciplinary nature of KDD causes terminology use to vary from source to source.

The aim of this paper is ***to compare*** *the notions and definitions of KDD within the sources we studied and* ***to point out*** *their similarities and their differences. From all the steps of the KDD process, we will focus on the data mining step. (KDD is often misleadingly called data mining.) An attempt to link together the techniques and the methods as well as the tasks listed in each source under different names is presented here in the form of tables. We have made our conclusions hoping that we have chosen the best views for our later use.*

Keywords: knowledge discovery in databases (KDD), the process of KDD, data mining (DM).

1. INTRODUCTION

The large amounts of data collected from either manufacturing or business, (often as a side effect of computerisation), should be thoroughly analysed as they might contain some precious information for decision support. There is nothing new about analysing data, but it is the amount of data that is being looked at that has meant traditional methods are becoming inefficient. It is often misleadingly believed that **data mining** is a new and powerful technology. "The new is the confluence of (fairly) mature offshoots of such technologies as visualisation, statistics, machine learning and deductive databases, at a time when the world is ready to see their value." [7]

[1] INCO-COPERNICUS Project 977091.

2. THE PROCESS OF KDD

2.1. Definition

When studying literature concerning data mining we have encountered terms such as: **data mining, knowledge discovery in databases** or **abbreviation KDD**. In various sources those terms are explained in a rather different way. What follows next is a list that shows some of them and it also shows how they vary.

Quite clear definition of data mining is presented in [8]:

Data mining - the process of extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions.

A different view is presented in [6] where the definition is as follows:

Knowledge discovery in databases (often called data mining) aims at the discovery of useful information from large collections of data. In addition the author puts special emphasis on the fact that the goal of KDD is inherently interactive and iterative, and it is a process that contains several steps, one of which is data mining. (However in the rest of this article it is difficult to distinguish between KDD and DM.)

According to [5] *KDD* is an abbreviation of *knowledge discovery and data mining*, and this may lead to some confusion.

In our opinion the most sophisticated definition is the one given in [3], where the authors have determined that knowledge discovery in databases is an interactive and iterative process with several steps and that data mining is a part of this process. The process of KDD is defined as:

The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

The terms of the above definition are explained as follows:

pattern

- models or structure in data (traditional sense)
- an expression in some language, describing a subset of the data or a model applicable to that subset (the data comprises a set of facts)

process

- implies that there are many steps repeated in multiple iterations

nontrivial (process)

- it must involve a search for a structure, models, patterns, or parameters

valid

- discovered patterns should be valid for new data with some degree of certainty

novel

- at least to the system and preferably to the user

potentially useful

- for the user or task

understandable

- discovered patterns should be understandable - if not immediately, then at least after some postprocessing.

The authors suggest that this definition implies a way for defining quantitative measures for evaluation of extracted patterns, based on required and obtained notions. For **validity** we can define the measure of **certainty** or **utility** (a gain in some currency, due to a better prediction). Notions such as **novelty** and **understandability** are more subjective, in some cases understandability can be estimated by **simplicity** (the number of bits needed to describe a pattern). Interestingness is the name of the notion for an overall measure, which includes validity, novelty, usefulness and simplicity. Interestingness functions can be explicitly defined or manifested implicitly (the ordering of discovered patterns by the KDD system).

In the rest of the sources that were looked at there are no measures considered for discovered patterns evaluation (namely [8]). Some people feel that ([6] for instance) evaluation should be left to the user who is given all the patterns that satisfy the user specifications and occur frequently enough in the data. In the author's opinion this is an advantage of such a system as every user has a different subjective measure for interestingness, according to his prior knowledge.

In most of the sources read the term **Data Mining** (DM) is often used to name a field of knowledge discovery. This confusing use of terms KDD and DM is due to historical reasons and due to the fact that the most of the work is focused on refinement and the applicability experiments of ML and AI algorithms for the data mining step. Preprocessing is often included in this step as a part of the mining algorithm.

2.2. The steps of the KDD process

According to the definition above, KDD is an interactive and iterative process. It means that at any given stage the user should have the opportunity to make changes (for instance to choose a different task or technique) and repeat the following steps to achieve better results. In Table 1 we have listed those steps of KDD where we have compared the terms for different sources. The table is organised in such a way that the terms in the row refer to the same action.

Data mining (the dark grey coloured row) gets the most attention in research and therefore the same can be said for publications. These mostly focus on learning algorithms and some methods combine data mining with previous data preparation (the light grey coloured row), and this is usually a dataset reduction.

The KDD process according to [1] is outlined in Figure 1. The first two steps of the KDD process, namely *task discovery* and *data discovery,* produce the first input (a goal of the KDD process). The following steps in the KDD process are *data cleaning*, *model development, data analysis* and *output generation.* In what follows the inputs and steps of a KDD process according to [1] will be described in more detail.

Table 1. The process of KDD - list of steps

Simoudis [8]	Mannila [6]	Fayyad et al. [3]	Brachman & Anand [1]
	Understanding the domain	learning the application domain	task discovery
data selection		creating a target dataset	data discovery
data transformation	preparing the data set	data cleaning and preprocessing	data cleaning
		data reduction and projection	model development
		choosing the function of data mining	
data mining	Discovering patterns (data mining)	choosing the data mining algorithm(s)	data analysis
		data mining	
result interpretation	Postprocessing of discovered patterns	interpretation	output generation
	putting the results into use	using discovered knowledge	

Task Discovery is one of the first steps of KDD. The client has to state the problem or goal, which often seems to be clear. Further investigation is recommended such as trying to get acquainted with the customer's organisation after having spent some time at the place and then to sift through the raw data (to understand its form, content, organisational role and the sources of data). Then the real goal of the discovery will be found.

Data Discovery complements the step of task discovery. In the step of data discovery, we have to decide whether the quality of data is satisfactory for the goal (what the data does or does not cover).

The Domain Model plays an important part in the KDD process, though it often remains in the mind of the expert. A data dictionary, integrity constraints and various forms of metadata from the DBMS can possibly contribute to the retrieval of background knowledge for KDD purposes as well as some analysis techniques. These can take advantage of formally represented knowledge when fitting data to a model (for example ML techniques such as explanation-based learning that are integrated with inductive learning techniques).

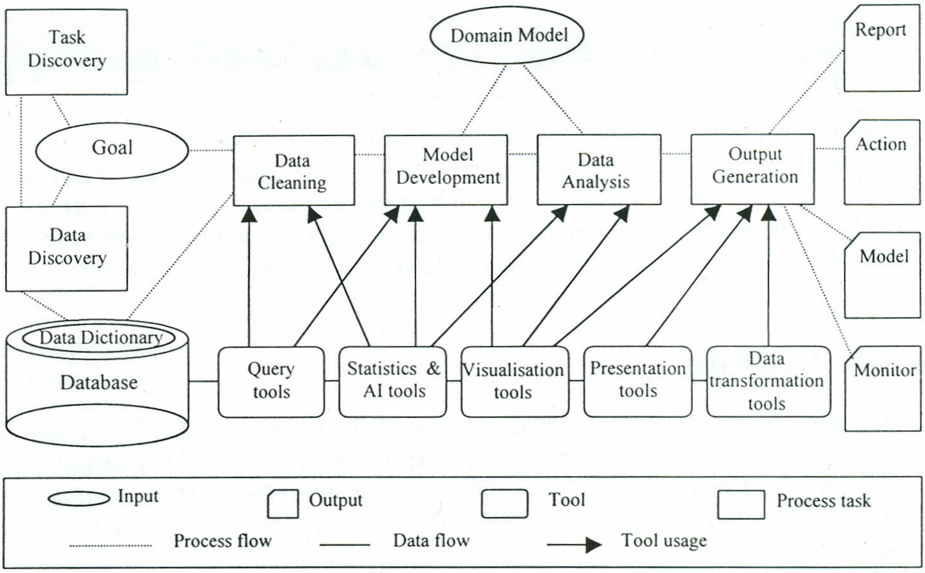

Figure 1. Schema of the KDD process

Data Cleaning is often necessary though it may happen that something removed by cleaning can be an indicator of some interesting domain phenomenon (outlier or a key data point?). The analyst's background knowledge is crucial in the data cleaning that is provided by comparisons of multiple sources. Another way of cleaning is to clean the data before loading it into the database by using editing procedures. Recently, the data for KDD has come form data warehouses which contain data that has already been cleaned in some way.

Model Development is an important phase of KDD that must precede the actual analysis of the data. Interaction with the data leads analysts to the formation of a hypothesis (it is often based on experience and background knowledge). Sub-processes of this model development are:

- data segmentation (unsupervised learning techniques, for example clustering)
- model selection (choosing the best type of model after having explored several different types)
- parameter selection (the parameters of a chosen model).

Data Analysis in general is the wish to understand why certain groups of entities behave the way they do and it is the search for laws or rules for this type of behaviour. The first thing to be analysed should be those areas where such groups have already been identified. Sub-processes in data analysis are:

- model specification - some formalism is used to denote a specific model
- model fitting - when necessary the specific parameters are determined (in some cases the model is independent of the data, and in other cases the model has to be fitted to the training data)
- evaluation – the model is evaluated against the data

- model refinement – the model is refined in iterations according to the evaluation results.

As mentioned previously model development and data analysis are complementary, so it often leads to oscillation between those two steps.

Output Generation - output can be in various forms. The simplest form is a report with the analysis results. Another, more complicated form, is graphs and in some cases it is desirable to obtain action descriptions which might be taken directly as outputs. Or there could be a monitor as an output, which should trigger an alarm or action under a certain condition. Output requirements might determine the task of the designed KDD application.

3. THE TASKS OF DM

For the title of this section we have used the term the tasks of DM, although we looked at many different terms and these are shown in Table 2. This table is organised in such a way that the tasks in a row (from different sources) refer to the same task. This organisation is based on a particular description of the table items available in a particular source.

We accepted the list of tasks in the first column as a standard and a brief description of the DM tasks is as follows (For more details see [4].):

- **the discovery of SQO rules** - to perform a syntactical transformation of the incoming query to produce a more efficient query by adding or removing conjuncts; it is characteristic of SQO rules that the query processing time (derived from the access method and the indexing scheme of DBMS) is taken into account as the cost of an attribute
- **the discovery of database dependencies** - in this case the term refers to the relationships among attributes of relations
- **the discovery of association rules** – the relationship of sets of items, i.e. those that are assigned by support and the confidence factor
- **dependence modeling** - dependencies among attributes are in the form of if-then rules as in "if (A is true)-then (C is true)"
- **deviation detection** - focuses on the discovery of significant deviations between the actual contents of a data subset and its expected contents
- **clustering** – a classification scheme, where the classes are unknown
- **causation modeling** – the relationship of cause and effect among attributes
- **classification** - each tuple belongs to a class, one of the pre-defined set of classes
- **regression** - similar to classification, the predicted value is rather continuous
- **summarisation** – a kind of summary, describing some properties shared by most of the tuples belonging to the same class.

Table 2. The tasks of DM

Freitas [4]	Simoudis [8]	Fayyad et al. [3]	Fayyad et al. [2]	Mannila [6]
Tasks	Operations	model functions	tasks	problem
	query and reporting			
Discovery of SQO rules	Multidimensional analysis			
Discovery of database dependencies	link analysis (associations or relations between the records)	link analysis		finding keys or functional dependencies
Discovery of association rules				association rules
Dependence modeling		Dependency modeling	dependency modeling	
Deviation detection	deviation detection		change and deviation detection	
Clustering	Database segmentation (clustering)	Clustering	clustering	
Causation modeling		sequence analysis		finding episodes from sequences
Classification	Predictive modelling (C4.5, NN)	Classification	classification	
Regression		Regression	regression	
Summarisation	statistical analysis (EDA[2])	Summarisation (EDA)	summarisation (EDA)	

4. CONCLUSION

In this paper we aimed to provide an introductory overview of the field of knowledge discovery in databases with the emphasis on one part of it - data mining. As most researchers agree, KDD is a process made up of several steps, where data preparation is as important as the knowledge extraction itself. Less attention is given to the evaluation and the usage of this extracted knowledge and in this area there clearly is the potential for further research.

ACKNOWLEDGEMENTS

This work is supported by the European Commission within the INCO-COPERNICUS Project 977091 "Geographic Information On-Line Analysis: GIS - Data Warehouse Integration (GOAL)".

[2] Exploratory Data Analysis

REFERENCES

[1] R. J. Brachman and T. Anand. The process of knowledge discovery in databases. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, eds. *Advances in Knowledge Discovery & Data Mining.* AAAI/MIT Press, Cambridge, Massachusetts, 1996.

[2] U. M. Fayyad, G. Piatetsky-Shapiro and P. Smyth. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery & Data Mining.* AAAI/MIT Press, Cambridge, Massachusetts, 1996.

[3] U. M. Fayyad, G. Piatetsky-Shapiro and P. Smyth. The KDD process for extracting useful knowledge from volumes of data. *Comunications of the ACM,* Vol.39, No.11, 1996., pp. 27-34.

[4] A. A. Freitas. *Generic, Set-Oriented Primitives to Support Data-Parallel Knowledge Discovery in Relational Database Systems.* Ph.D. Thesis, University of Essex, UK, 1997.

[5] S. R. Hedberg. Searching for the mother lode: tales of the first data miners. *IEEE EXPERT,* Vol.11, No.5, 1996, pp. 4-7.

[6] H. Mannila. Methods and problems in data mining. In F. Afrati, P. Kolaitis and P. Delphi, eds. *The Proceedings of International Conference on Database Theory.* Springer-Verlag, Berlin, 1996.

[7] B. Mark. Data mining - Here we go again? *IEEE EXPERT,* Vol. 11, No.5, 1996.

[8] E. Simoudis. Reality check for data mining. *IEEE EXPERT,* Vol.11, No.5, 1996.

Received: 8 October 1999
Accepted: 10 December 1999

Eva Andrássyová
Ján Paralič

OTKRIVANJE ZNANJA U BAZAMA PODATAKA: USPOREDBA RAZLIČITIH GLEDIŠTA

Sažetak

Područje otkrivanja znanja u bazama podataka postaje sve zanimljivije jer sustavi za otkrivanje znanja sadrže tehnike iz različitih povezanih područja. Tijekom rada na međunarodnom projektu GOAL (projekt se odnosi na on-line analizu zemljopisnih informacija integracijom zemljopisnog informacijskog sustava i skladišta podataka), ustanovljeno je da interdisciplinarna priroda otkrivanja znanja u bazama podataka uzrokuje različitost terminologija koje se upotrebljavaju u ovom području. Istraživanje se odnosilo na tehnike koje se primjenjuju u tom procesu, na zadatke koje koje je trebalo rješavati i na pojedinačne korake koje je pri tom trebalo provesti. Svrha istraživanja bila je usporediti različite nazore i odgovarajuće definicije ovog procesa, te naglasiti sličnosti i razlike među njima. Poseban naglasak stavljen je na rudarenje podataka kao jedan korak u tom procesu. Istaknuto je da se često proces otkrivanja znanja u bazama podataka pogrešno poistovjećuje s procesom rudarenja podataka.

Ključne riječi: otkrivanje znanja u bazama podataka, proces otkrivanja znanja u bazama podataka, rudarenje podataka.