Boris Brešić

Knowledge Acquisition in Databases

Article Info:

Management Information Systems,
Vol. 7 (2012), No. 1,
pp. 032-041

Received 13 November 2011
Accepted 15 February 2012

UDC 005.94 ; 004.65 ; 004

Summary

Knowledge discovery and acquisition in databases features as a separate discipline of business intelligence. Generally speaking, it denotes the process of analysing large quantities of data, its goal being to discover new information and knowledge, and apply them in resolving business problems. More specifically, it implies to data acquisition – or mining – as the initial step, using data-driven learning algorithms, and is aimed at establishing their mutual correlations. Due to its popularisation, significantly supported by computing technology development, the notion of data mining has gradually come to be equated with knowledge discovery. This article deals with knowledge acquisition from databases, i.e. data mining with OLAP (On-Line Analytical Processing) tools, which enable multidimensional data management and graphic presentation thereof. Such an approach is mostly applicable when analysing smaller data sets, i.e. in the phase of data preparation for mining, in order to better understand or select variables included in the business information process.

Keywords

business intelligence, database, knowledge discovery, data mining, OLAP tools

1. Introduction

Discovering, or better to say, acquiring knowledge in databases as a discipline of business intelligence in its broader sense refers to the process of analysing large amounts of data with the aim of discovering new information and knowledge, and its application in resolving business problems. In its more specific tense, it refers to mining data as the initial step of this process, using data-driven learning algorithms, with the aim of discovering correlations between them. The term “knowledge discovery in databases” (KDD) was introduced by Usama Fayaad at the first international conference on knowledge discovery and data mining (Giudici, 2003, p. 2). According to Fayaad (1996), the process comprises six steps, where data mining is the step applying one of the algorithms on transformed data (Figure 1). The CRISP-DM process model contains basically the same steps, only extended and elaborated for application in industrial environment. Popularising the term “data mining” gradually led to equating it to the term “data discovery”, so that many authors and experts consider these two terms synonymous.

Dissemination and popularization of knowledge discovery were preceded by the development of computing technology, especially in the segment of data storage and relational database development.

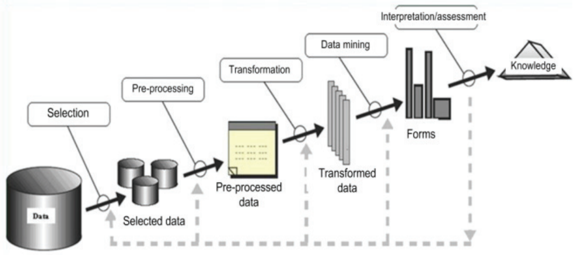

Figure 1 Knowledge discovery process (Fayyad, Piatetsky-Shapiro, & Smyth, 1996)

When discussing knowledge in databases, one primarily refers to data mining. However, what must not be forgotten is knowledge discovery in databases using tools for ad hoc queries and analyses, as well as reporting. Actually, knowledge is also discovered by means of application of Structured Query Language (SQL) queries and relational databases, that is, compiling reports. Such a procedure starts from an assumption, i.e. setting a hypothesis.

The user sets a hypothesis, and then tries to confirm or verify it by shaping and executing queries. Such knowledge discovery procedure is even more simple and efficient with OLAP (On-Line Analytical Processing) tools, which enable multidimensional data handling and graphic presentation thereof. In a such top-down knowledge discovery method, criteria and correlations between data are implied by the user who tries to confirm them. Such an approach is applicable when analysing smaller data sets, but in the case of large amounts of data, it is not always simple and easy to confirm or refute the given hypothesis. In that case, using the data mining technique as a bottom-up knowledge discovery method, algorithms break down data and reach the goal, i.e. the discovery of patterns or regularities hidden by the data.

The classical queries and analyses in databases cannot substitute data mining as a knowledge discovery method in general. These are two mutually supplementing methods, so that, for instance, OLAP tool can be applied in the phase when data is prepared for mining so that the latter can be understood better, that is, for selecting variables to be included in the mining procedure.

2. OLAP

2.1. Transaction Processing Systems

In 1970, British mathematician E.F. Codd set up a relation data model that was to become the basis of the currently most common relational databases. Operative systems supporting parts or entire business processes save the data, which they subsequently process into relational databases. Such systems are also referred to as transactional systems, or On-Line Transactional Processing (OLTP) systems. Their main feature is that they require a high level of accessibility and response to user requirements. Adherence to normal forms (Database normalization, n.d.) when designing relational models results in logical consistency of data and robustness of the data model. Some examples of OLTP systems include systems for charging for prepaid services used by mobile telecommunication service providers and banking transaction systems.

For the purpose of reporting and analysing data left by users after every transaction, executing queries on relational bases of OLTP is not recommendable for several reasons. Data extraction for reports often includes multiple criteria and numerous tasks, which is a great burden, whereas given the inherent feature of the relational data model (normalised models), the Relational Database Management System (RDBMS) is under great burden as well. On the other hand, response time to complex analytic queries into the base can be extremely long, which naturally depends on query complexity, and is also the greatest problem. Finally, OLTP bases do not store historic data.

2.2. Data Warehouses

Due to the reasons above, late 1980s and early 1990s saw the development of the concept of data storage. Unlike transactional databases, built on the relational model, data warehouses are built on the dimensional data model[1]. The dimensional data model enables fast response, because data from relational bases are denormalised and adjusted (summarized and aggregated) to the needs of analyses and reporting. Such data organization is the foundation of the OLAP system. The data warehouse also contains all historic data, and it is organized around the ratios of business processes and dimensions to be observed. Dimensional data models can also be applied in classical relational databases through the star scheme, which is the dominant form of data warehouse organization.

Figure 2 shows a simplified dimensional model adjusted to application in the relational database model. It is a case of tracking orders of various products over time. The *orders_fact* fact table records the quantity of ordered products according to the following dimensions: *product*, *supplier*, *employee*, *customer* and *time*.

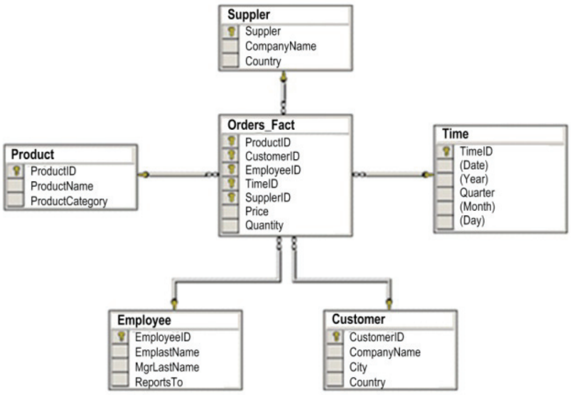

Figure 2 An example of dimensional model in the relational database (star scheme).

Data from the relational data bases of the OLTP system are loaded periodically into data warehouses, mostly in periods when system burden are the lowest. The data entry model is called *extract-transform-load* process. In a nutshell, the data is extracted from transactional bases and transform – which includes coding and deriving new valuesand variables, their aggregation and similar activities, and, finally, transformed values are loaded into the warehouse structure.

[1] Besides the denormalised dimensional data warehouse model propagated by Ralph Kimball (2002), there is also the the normalised model concept (the third dormal form) advocated by Bill Inmon.

The process is illustrated as follows:

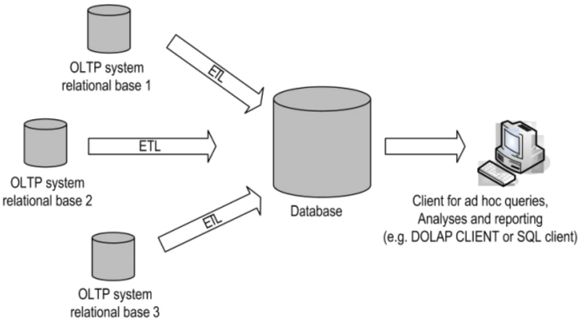

Figure 3 Data load into databases

2.3. OLAP Systems

In 1994, Codd published an article introducing the term OLAP, listing the 12 OLAP rules. The article eventually turned controversial, when it was revealed that it had been financed by an OLAP tool producer. Regardless of this, Codd formalized the concept of dimensional analysis and reporting under the acronym of OLAP. The essence of the whole concept is that data is organized in multidimensional cubes consisting of facts grouped by dimensions. Such organization enables fast access to data, which is a prerequisite for research "at the speed of thought" (Panian & Klepac, 2003, p. 237).

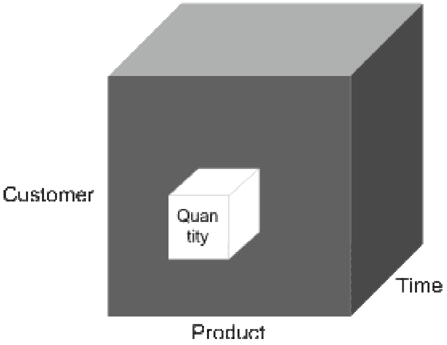

Figure 4 OLAP cube with the dimensions of customer, product, and time, and the ratio of product quantity

Modern OLAP tools enable complex data analyses, and with regard to the data resource, they are divided into multidimensional OLAP tools (MOLAP), using their own dimensional database, (ii) relational OLAP tools (ROLAP), (iii) desktop OLAP tools (DOLAP), i.e. tools designed for personal computers (without the need for dedicated OLAP servers) and (iv) hybrid OLAP tools (HOLAP), which combine the first two concepts.

Figure 5 shows two OLAP system architectures – MOLAP and ROLAP:

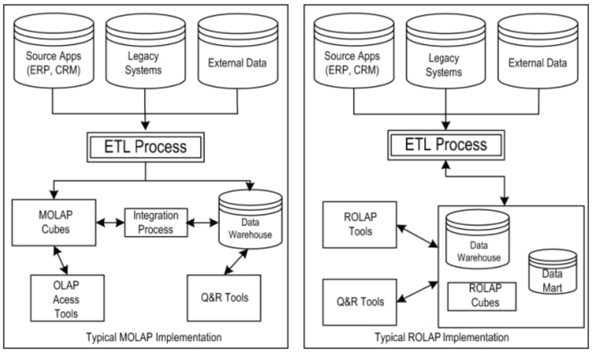

Figure 5 MOLAP and ROLAP system architectures (Moreno & Mancuso, 2002)

OLAP tools enable complex analyses and reports, with dimensional analyses, by way of slice and dice, as well data pivot, then support hierarchical structures which enable summing up and aggregation of detailed data into new data at a higher hierarchical degree, then compiling graphic reports, etc. One must bear in mind that OLAP tools are devised so that they can also be used by people without particular technical skills (e.g.) marketing experts. In this respect, it can be said that one of the roles of OLAP tools is bridging the information gap in a business organisation. (Pendse, 2008)[2]

OLAP tools are the foundation of the business intelligence system, i.e. decision-making support systems, and together with corporate warehouses, they are the backbone of the reporting system in organisations. Dimensional databases can also serve as an excellent data source in the mining process, as these data are filtered and standardised through the ETL process, which is an important prerequisite of successful project. For this reason, OLAP tools must be taken in considerations as one of the ways of data discovery in databases.

3. Data Mining

As a discipline of business intelligence, data mining, found its role with the development of computing technologies in 1990s, not only in the storage segment, but only in developing modern systems for relation database management. An increasing number of organisations base their operations on information transaction systems processing huge amounts of data in real time. Such data are a large information potential that needs to

[2] The OLAP market was worth somewhat under 5 billion dollars in 2005. This figure is forecast to reach the value of almost 6 billion dollars in 2007. The leading produces are Microsoft, Hyperion, Cognosm who took up a 62,3% market share in 2005.

be liberated one way or another, and one of those is data mining. The term was introduced by analogy with excavation of valuable ores from the depths of the earth.

There are several definitions of data mining, e.g.

1. "The process of research and analysis, automatic or semi-automatic, of large amounts of data with the aim of discovering meaning patterns and rules;" (Berry & Linoff, 2000, p. 7)
2. "The process of selecting, exploring and modelling large amounts of data discovering regularities or relationships unfamiliar at first sight, with the aim of obtaining clear and useful results for the database owner;" (Giudici, 2003, p. 2)
3. "The process of discovering various models, summaries and derived values from the given data set". (Kantardzic, 2003, p. 5)

All these definitions share two notions: process and data.

The notion of process somehow suggest that this is not a common procedure that can be completed with out-of-the-box[3] program tools, but a complex one comprising several steps, where an important role is played by the analyst, i.e. a data mining expert, his experience and skills. Berry and Linoff emphasise this, and compare data mining with photography, i.e. evolution of this process through history – from the first *camera obscura* and hand-developed photographs to the most modern and completely automated cameras: regardless of how technology is automated, the photographer's talent and the possibility of adjusting individual parameters are decisive for the final quality. In other words, regardless of how automated the data mining process, i.e. modelling phase is, the analysts' knowledge of the domain of the business problem being solved conditions the quality of results.

The second notion – data – point to the fact that the data are in the centre of the process. The success of a particular data mining project directly depends on their quality. What is explicitly not mentioned in definitions are mining techniques. Actually, there are several techniques gathered under the blanked term of data mining, originating from statistics and machine learning.

Machine learning as a discipline of computer science, or, in other words, as a part of the artificial intelligence domain, deals with the development of techniques and algorithms enabling learning, e.g. distinguishing between samples or developing computer programs. According to Berry and Linoff (2000), corporate use of discovered artificial intelligence techniques was achieved in early 1980s, at the time of sudden decline in investment in artificial intelligence, as the expectations from the 1960s and 1970s (e.g. machine translation) were not met. Some of the techniques used both in machine learning and data mining are decision making trees, neural networks and genetic algorithms.

On the other hand, statistical methods have been used in analysing gathered data even earlier, and one of the better-known statistical techniques is forecasting trends, i.e. regression.

The application of techniques for business purposes in both areas received a common name – data mining.

3.1. The Tasks and Techniques of Data Mining

In view of the two basic goals, the tasks of data mining can be classified into two categories:

- predictive data mining, also referred to in literature as asymmetrical, supervised and directed.
- descriptive data mining, also referred to as symmetrical, unsupervised and undirected.

Each task is completed with one of the several techniques, which are, in turn, implemented with various algorithms.

3.1.1. Predictive Tasks

Predictive data mining encompasses a set of tasks with the common goal of building a model that represents the relationship between several independent input variables and a single dependent input variable. In predictive data mining, the target known the variable is the one whose value is to be related to values of remaining variables. In predictive data mining, the model is built on known data, including the value of the target variable. Applying thus obtained model on a set of data without known value of the target variable is in fact "guessing" this value, but with the knowledge of relations from the data set on which the model was built, and the assumption that these relations also apply to the data set on which the model is applied. The aim is to have a model that will produce the most accurate amount of the target variable's value.

The category of predictive data mining procedure includes the tasks of estimation,

[3] This term denotes a software product that can be immediately used for the purpose for which it was developed, without additional adaptations. An example of this is Microsoft Office and similar final products.

prediction in the narrower sense, and classification. The paradigm of tasks is not different; they only differ by the type of the target variable, i.e. its time context. Actually, the estimate implies modelling a continuous target value, which is not set in the time context, whereas continuous value is modelled in the future in the case of prediction. In the other hand, classification implies modelling categorical, discrete target variable – without time contest.

An example of classification problem is classifying a person into one of predefined income brackets (high, medium-sized and low income) based on data on work experience, educational background, profession, gender etc. In the case of estimation, a good example is assessing systolic blood pressure based on data on the observed person (gender, body mass index etc.). A case of client attrition from a telecommunication services supplier is an example of predictive mining task. Future tendency of client attrition is predicted based on data on earlier client attrition trends.

Techniques used in predictive mining include, for instance regression in statistics, or decision trees and neural networks from the machine learning area. Regression as a technique is very popular in the predictive mining domain, so that some author refers to prediction as regression method.

Figure 6 shows the decision tree technique:

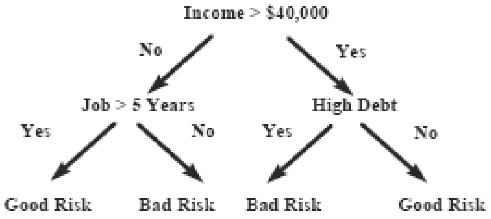

Figure 6 An example of a simple decision-making dree tree of a classification task (Two Crows Corporation, 2005, p. 15)

3.1.2. Descriptive Tasks

Unlike predictive data mining, there is no target variable in a descriptive task; that is to say, one does not build a model describing the target variable as the outcome dependent on other variables. Descriptive mining is applied on data in order to discover unusual or non-trivial relations and samples among them.

The category of descriptive data mining task includes data description and summarisation, clustering and association.

Data description and summarisation is a succinct description of characteristics of data in their basic and aggregated, i.e. summarised form. The users are thus familiarised with data, and visible regularities within the data are established. Moreover, these analyses set hypotheses which are confirmed or refuted. Data description and summarisation is, in most cases, a sub-project within a larger data mining project, which will result in a built model. In other words, the goal of data mining does not always have to be building a model, but also geting to know and describing available data. This contributes to the fact that application of tools for analysis and report, such as OLAP tools, also belongs among knowledge discovery methods.

Clusterisation is the process of identifying natural segments in a data set. More precisely, it is data segmentation, but not in accordance with predetermined criteria. The result is a set of clusters (the number of clusters is determined in advance, as a modelling parameter) where data is grouped according to a shared characteristic. After clustering, ti is necessary to analyse each cluster from the business aspect and identify its characteristics. A sample of clustering is customer behavioural segmentation with service providers. Based on habits in service utilisation, groups of users are determined based on some dominant common characteristic. The most used technique is k-means. However, techniques applied in predictive data mining such as neural networks and decision trees are also applicable in clustering tasks.

Association is a task whose aim is to discover relationships between variables. Association is also often referred to as affinity analysis, or consumer basket analysis. The aim is to discover rule quantifying the relationship between two or more variables. The outcomes are associative rules of the type “If event A, then event B”, together with support and confidence ratios. For instance, a study at an American chain store showed that 200 out of 1000 customers buy diapers on Thursday evenings, and 50 out of 200 buy beer. The associative rule would be “If they buy diapers, then they buy beer”, with the support ratio of 200/1000 = 20%, security ratio of 50/200 = 25%. The techniques used include *a priori* algorithm and GRI algorithm.

Predictive and descriptive data mining tasks, i.e. problem-oriented approaches, can be combined within a single mining project; for instance, one could first apply clustering on the initial data set to determine the client segment that will be used for modelling the client attrition problem by predictive regression technique.

4. CRISP-DM Process Model[4]

The data mining market was in infancy in the mid-1990s; however, there were signs of possible sudden expansion, which could have led to difficulties. More precisely, various data producers, i.e. the industry, developed their own approaches to the application of data mining, which could have resulted in non-standardised processes. This led to the need to define standards, i.e. standard data mining methods that could be introduced into any industrial environment consistently. In addition to cost cutting, the advantage of the existence of a defined method, i.e. data mining as a clearly defined process is optimizing the learning curve of the organisation that has decided to introduce data mining to solve its own business problems.

It is for this reason that Cross Industry Standard Process – Data Mining (CRISP-DM) was initiated in 1996. It was launched by three veterans in the field of this young, but propulsive discipline of business intelligence. Three companies – SPAS, DaimlerChrysler and NCR, have a long-standing experience in data mining. SPSS launched Clementine, the first commercial data mining tool, on the market in 1994; DaimlerChrysler represented industry as a leading data mining user, whereas NCR was an experienced market player in the data warehousing domain. A year later, the “trio” established a consortium and received financial assistance from the European Commission, and started implementing their idea. CRISP-DM was to be an industrial tool- and application-independent, standardised process.

To achieve successful implementation, they had to gather as much experience from data mining users as possible. The consortium therefore established an interest group, CRISP-DM Special Interest Group (SIG), which was to unify all interested parties in the process of defining a standardised process model. The first conference in Amsterdam gathered a surprising multitude of participants who supported the idea of defining the process and presented their experiences in projects of applying data mining for business purposes. It turned out that there were many points of convergence among the participants – from professional jargon to process phases that they applied in their business environments.

What followed was a period of developing CRISP-DM through projects and workshops, in cooperation with over 200 SIG members. The first framework process model was published in mid-1999, and, having been assessed through numerous projects, it was released as Version CRISP-DM 1.0. Time has shown that the standard was soon adopted and adapted by numerous firms. It is important to point out that the standard was based on experiences of users from various industries and various data mining projects.

4.1. CRISP-DM Methodology

The methodology of CRISP-DM process describes mapping a generalised, generic process and tasks comprising the process into a specific case, i.e. data mining process on a specific problem. The CRISP-DM process model describes the main phases of a project, and tasks that must be performed to have the project implemented. The model represents a generalised pattern applicable to any data mining process, and does not describe any a problem-specific application.

The CRISP-DM process can be viewed through four hierarchical levels describing the model at four levels of details, from general to specific (Figure 7).

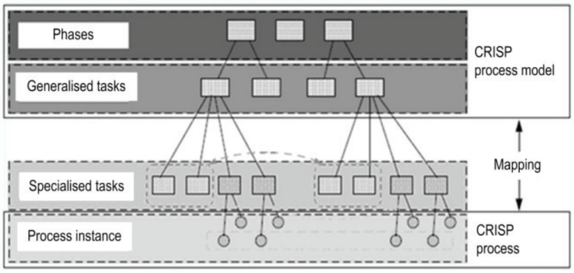

Figure 7 Four abstraction levels of the CRISP-DM process model (Chapman et.al, 2000)

The first level defines the basic phases of the process, i.e. the data mining project. Each specific project passes through the phases at the first level; the first level is, at the same time, the most abstract.

At the subsequent level, each stage is broken down into generalised, generic tasks. They are generalised, as they cover all the possible scenarios in the data mining process, depending on the phase the project is in.

The third level defines particular, specialised tasks. They describe how an individual generalised task from the second level is executed in the specific case. For instance, if the second level defines a generic task of data filtering, then the third level describes how this task is executed depending on whether it is a categorical or continuous variable.

Finally, the fourth level contains the specific instance of the data mining process, with a range of

[4] Chapman et.al, 2000

actions, decisions and outcomes of the actual knowledge discovery process.

Vertically, CRISP-DM methodology is mapping from the general CRISP-DM process into a process with a specific application, while horizontally, it gives an overview of the reference process, i.e. its phases consisting of tasks.

When mapping from the abstract into a specific process, one must view the context of the knowledge discovery. The context has four dimensions (Chapman et.al, 2000, p.11):

- Domain of use: specific area where the knowledge discovery project is conducted, i.e. forecasting client attrition of a telecommunication service provider;
- Type of problem: descriptive or predictive knowledge discovery, clustering, forecasting, classification, etc.;
- Technical aspect: possible technical problems, e.g. data quality or non-existent values, etc.
- Tools and techniques: SPSS Clementine, SAS Enterprise Miner, regression, decision trees, neural networks etc.

The specific data mining context is a combination of these four dimensions, i.e. the problem of forecasting client attrition in telecommunication by applying the decision tree technique, and using the SAS Enterprise Miner tool.

4.2. CRISP-DM Reference Process Model

The phases of a CRISP-DM process, i.e. a single data mining project, are described in a reference model. The model shows phases, tasks within individual phases, and relations between them. In essence, the process model describes the lifecycle of the data mining process, comprising six basic steps (Figure 8). The order of the phases is not fixed; in other words, the next step is decided depending on the outcome of each individual phase. If the stage outcome is unsatisfactory, or some knowledge was gained that can be useful in one of the previous phases, it is desirable to return to the previous phase.

Data mining projects are iterative; once a goal is reached, new knowledge and insights are discovered, and their application on resolving the same problem yields better results. It can be said that the knowledge discovery process is evolutional. Once the end of the process is reached, a new cycle begins, supported by information from the previous one.

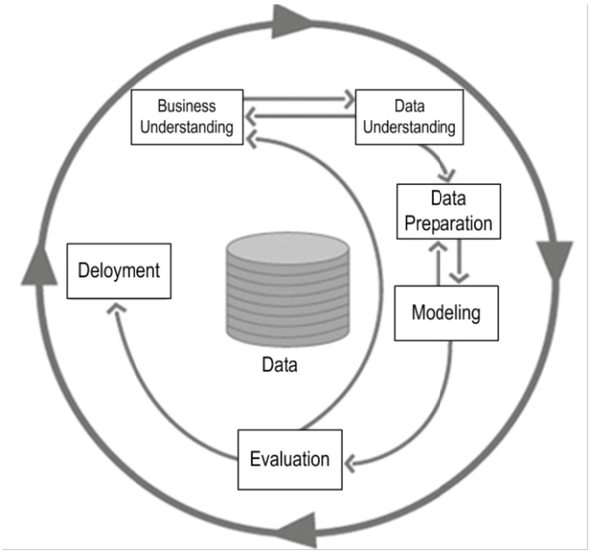

Figure 8 Phases of the CRISP-DM reference process model (Chapman et.al, 2000)

The CRISP-DM reference model defines the following phases:

1. Business understanding;
2. Data understanding;
3. Data preparation;
4. Modelling;
5. Evaluation;
6. Deployment.

4.2.1. Business Understanding

This is a phase of defining goals and demands according to the data mining project from the business aspect. This is accomplished in cooperation with clients whose requirements are determined and shaped into a data mining problem. The demands of business users, i.e. clients, are often conflicting, so that the analyst's task is to extract the key fact from the demands, so that the project will not go in the wrong direction at later stages. In addition, the problem domain is determined (marketing, user support or something similar), and the organisation's business units involved in the project. Furthermore, resources required for this project are identified, such as hardware and tools for implementation, as well as the human resources, especially the domain-specific experts. This is an extremely important phase of the project, so that it must be approached carefully, so as to complete the other project phases with equal success.

A project plan must be developed at the end of the first phase, with a list of phases, with tasks and activities, as well as the effort estimation. Resources for tasks, their interdependencies, inputs and outputs are also defined. A special attention in project planning must be paid to risk assessment and management.

Formally, a phase comprises four tasks: determining business objectives, assessing situation, determining data mining goals, and project plan production.

4.2.2. Data Understanding

Once the project goals have been determined, and the data mining problem determined, the next phase is getting familiar with the data available in the client organisation (the data sources are defined in the previous step). The data is obtained from defined sources, and the criterion must be chosen with the specific business problem in mind. Tables are defined, and if the data source is a relational database, or data warehouse, variations of the tables to be used are also defined.

When the data is obtained from source, the next step is analysing the basic characteristics of the data, such as quantity and types (e.g. categorical or continuous), then analysis of the correlations between variables, distribution and intervals of values, as well as other simple statistical functions, using specialised statistical analysis tools if necessary. It is important to establish the meaning for every variable, especially from the business aspect, and relevance to the specific data mining problem.

What follows is a more complex analysis of the data set, possibly by the use one of OLAP or similar visualisation tools. This analysis directly addresses questions related to the data mining project itself, such as shaping hypotheses and transforming them into a mining problem space. In addition, project goals are determined more precisely. The end of this phase is the moment of determining the quality of the data set, such as completeness and accuracy of data, and frequency of discrepancies or nil-values. It is important to bear in mind that, apart from appropriate problem formulation, data is the most important key to success.

One can generally say that, at this stage, the analysts familiarise themselves through exploratory data analysis, which includes simple statistical characteristics and more complex analyses, i.e. setting certain hypotheses on the business problem.

Formally, the phase comprises these tasks: collecting initial data, describing data, exploring data and verifying data quality.

4.2.3. Data Preparation

Having analysed the data available to analysts, it is necessary to prepare data for the mining process. It includes choosing the initial data set on which modelling is to begin. Such a set is referred to model set. This is a very demanding process phase.

When defining the set for the subsequent modelling step, one take into account, among other things, elimination of individual variables based on the results of statistical tests of correlation and significance, i.e. values of individual variables. Taking these into account, the number of variables for the subsequence modelling iteration is reduced, with the aim of obtaining an optimum model. Besides this, this is the phase when the sampling (i.e. reducing the size of the initial data set) technique is decided on.

This is also the phase when the issue of data quality is addressed, as well as the manner in which non-existent values will be managed, and which strategy will be applied when handling particular values. Furthermore, new variables are derived, the values of the existing ones are transformed, and values from different tables are combined in order to obtain new variables, i.e. values. Finally, individual variables are syntactically adjusted to individual modelling tools, without changing their meaning.

The phase comprises the following tasks: data selection, data cleaning, data construction, data integration and data formatting.

It is important to note that the data preparation phase during the data mining project is performed recurrently, as findings from the subsequent modelling phase sometimes require redefining the data set for the next modelling step. A specific example is reducing the number of variables by eliminating low-impact correlations and variables based on criteria obtained as the outcome for the modelling phase.

4.2.4. Modelling

After the data set on which some of the mining techniques will be applied has been prepared, the subsequent step is choosing the technique itself. The choice of the tool, i.e. technique to be applied depends on the nature of the problem. Actually, various techniques can always be applied on the same type of problem, but there is always a technique or tool yielding the best results for a specific problem. It is sometimes necessary to model several techniques and algorithms, and then opt for the one yielding the best results. In other words, several models are built in a single iteration of the phase, and the best one is selected. This is where the analyst's experience is manifested. One thing, however, is clear: it must be established at the very beginning whether it is an issue for predictive or descriptive mining.

Before modelling starts, the data (model) set from the previous phase must be divided into subsets for training, testing and evaluation. The evaluation subset is used for assessing the model's efficiency on unfamiliar data, whereas the test subset is used for achieving model generality, i.e. avoiding the overfitting on the training subset.

Division of the data subset is followed by model building. The effectiveness of the obtained model is assessed on the evaluation subset. In the case of predictive models, applying the obtained model on the evaluation subset produces data for the cumulative gains chart showing how well the model predicts on an unfamiliar set. Based on the obtained graph, the model quality ratio (surface below the graph) and other ratios – such as significance and value factors for individual variables, and correlations between them – parameters for the subsequent modelling step are determined. If necessary, the developer returns to the previous phase to eliminate noise variables from the data, i.e. model set.

If several models were built in this phase (even if it was done using different techniques), then models are compared, and the best ones are selected for the next iteration of the modelling phase. Each obtained model is interpreted in the business context, as much it is possible in the current phase iteration. The developers then assess the possibility of model deployment, result reliability, and whether the set goals are met from the business and analytic aspect. The modelling phase is repeated until the best, i.e. satisfactory model is obtained.

Tasks foreseen in the modelling phase are: generating test design, building the model, and assessing the model.

4.2.5. Evaluation

Unlike evaluating the model in the previous phase, related to the model's technical characteristics (efficiency and generality), this phase assesses the final model, i.e. how much it meets the goals set in the first phase of the process. If the information gained at this point affects the quality of the entire project, this means returning to the first phase and re-initiating the cycle, but now taking the new facts into account. If he model meets all the business goals and is considered satisfactory for application, this step is followed by the review of the entire data mining process, in order to secure the quality of the entire process, i.e. check whether there have been any oversights in its execution. Basically, it is a general quality management and provision, within the domain of the project management discipline.

The final step is defining the next steps, including decision on moving to the subsequent phase, the phase of model deployment, or repeating the entire process for improvement.

Tasks comprising the evaluating phase are: evaluating results, reviewing the process and determining next steps.

4.2.6. Model Deployment

Once the model has been evaluated, it is deployed in business, taking into account the way of measuring the model's benefits and its supervision. Actually, it is of extreme importance to define a plan for model supervision, as, due to the nature and dynamics of the business process, it is necessary to repeat the modelling process from time to time so as to keep it up-to-date. Acting according to the results of a model that has not been updated may lead to unwanted circumstances. In other words, it is necessary to establish the dynamics of changing the model and choose the technique of verifying model quality over time.

The application of the model in strategic decision making of a business organisation can be used for direct measurement of the benefits of the obtained model, and gather new knowledge for the subsequent modelling cycle. Berry and Linoff therefore refer to data mining as *virtuous circle of knowledge*.

The end of this phase also completes a particular data mining project. Final reports and presentations are made. The project is concluded by overall review, i.e. analysis of its strengths and weaknesses. Documentation with experiences usable in possible future projects is also compiled. Project contributors and beneficiaries are interviewed.

Formally, this phase includes the following tasks: plan deployment, plan monitoring and maintenance, final report production and project review.

5. Conclusion

This article has clearly, systematically and analytically described the procedures of knowledge discovery in databases, and reached at least two conclusions. On the one hand, it is beyond doubt that knowledge discovery implies a process where data mining, i.e. acquisition is the key step, and includes certain techniques and tools. On the other, data mining has also imposed itself as a synonym for data discovery and acquisition, and features as an important factor of business intelligence as a young and relatively autonomous social discipline emerging within computing technology.

References

Berry, M. J., & Linoff, G. S. (2000). *Mastering Data Mining – The Art and Science of Customer Relationship Management.* New York: John Wiley & Sons.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., et al. (2000). *CRISP-DM 1.0 Step-by-step data mining guide.* Retrieved May 14, 2011, from SPSS: http://www.spss.ch/upload/1107356429_CrispDM1.0.pdf

Database normalization. (n.d.). Retrieved May 14, 2011, from Wikipedia: http://en.wikipedia.org/wiki/Database_normalization

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From Data Mining to Knowledge Discovery in Databases.* Retrieved June 15, 2011, from KDnuggets: http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

Giudici, P. (2003). *Applied Data Mining - Statistical Methods for Business and Industry.* Chichester: John Wiley & Sons.

Kantardzic, M. (2003). *Data Mining – Concepts, Models, Methods and Algorithms.* New Jersey: IEEE Press.

Kimball, R. R. (2002). *The Dana Warehouse Toolkit – The Complete Guide to Dimensional Modeling.* New York: John Wiley & Sons.

Moreno, A., & Mancuso, G. (2002). *The Role of OLAP in the Corporate Information Factory.* Retrieved May 14, 2011, from Information Management and SourceMedia: http://www.information-management.com/issues/20021101/5990-1.html

Panian, Ž., & Klepac, G. (2003). *Poslovna inteligencija.* Zagreb: Masmedia.

Pendse, N. (2008, July 3). *OLAP market share analysis.* Retrieved May 14, 2011, from BI Verdict: http://www.bi-verdict.com/fileadmin/dl_temp/f8d840e599f87665172dc20ac5059324/market.htm

Two Crows Corporation. (2005). *Introduction to Data Mining and Knowledge Discovery* (3th ed.). Maryland: Two Crows Corporation.

Boris Brešić
Faculty of Organization and Informatics, Varaždin
Paulinska 2
42000 Varaždin
Croatia
Email: boris.bresic@gmail.com