PROPOSAL OF APPLICATION DATAWAREHOUSES INTO CONTROL PROCESS

Author: ***Andrej TRNKA***
Workplace: ***Faculty of Natural Sciences, University of ss. Cyril and Methodius in Trnava***
Email: ***andrej.trnka@ucm.sk***

Abstract:

This paper describes possibilities of datawarehouse implementation and Data mining methods into control process. Aim of my doctoral thesis is to propose knowledge discovery methodology from control process by Data mining methods. In first part is described datawarehouse design, next part describes subject of Knowledge data in databases and Data mining methodology. In conclusion is described my next research in Data mining area.

DATAWAREHOUSING

A datawarehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions. The datawarehouse contains granular corporate data. Data in the data warehouse is able to be used for many different purposes, including sitting and waiting for future requirements which are unknown today. (1)

Subject oriented means that the data is organized around subjects rather than operational applications.
Nonvolatile means that the data, once placed in the warehouse, is not usually subject to change.
Integrated means the data is consistent. Integration is a process through which the data passes after it leaves the application database and before it enters the warehouse database.
Time variant means that historical data is recorded. (2)
Figure 1 shows architecture of datawarehouse.

Within the context of datawarehouse is used term ETL. ETL stands for Extract, Transform, and Load. It is the process of retrieving and transforming data from the source system and putting it into the datawarehouse.

The source systems are the OLTP systems that contain the data want to load into the datawarehouse. Online Transaction Processing (OLTP) is a system whose main purpose is to capture and store the transactions. ETL system brings data from various source systems into a staging area. ETL is a system that has the capability to connect to the source systems, read the data, transform the data, and load it into a target system (the target system doesn't have to be a data warehouse). The ETL system integrates, transforms, and loads the data into a dimensional data store (DDS). A DDS is a database that stores the data warehouse data

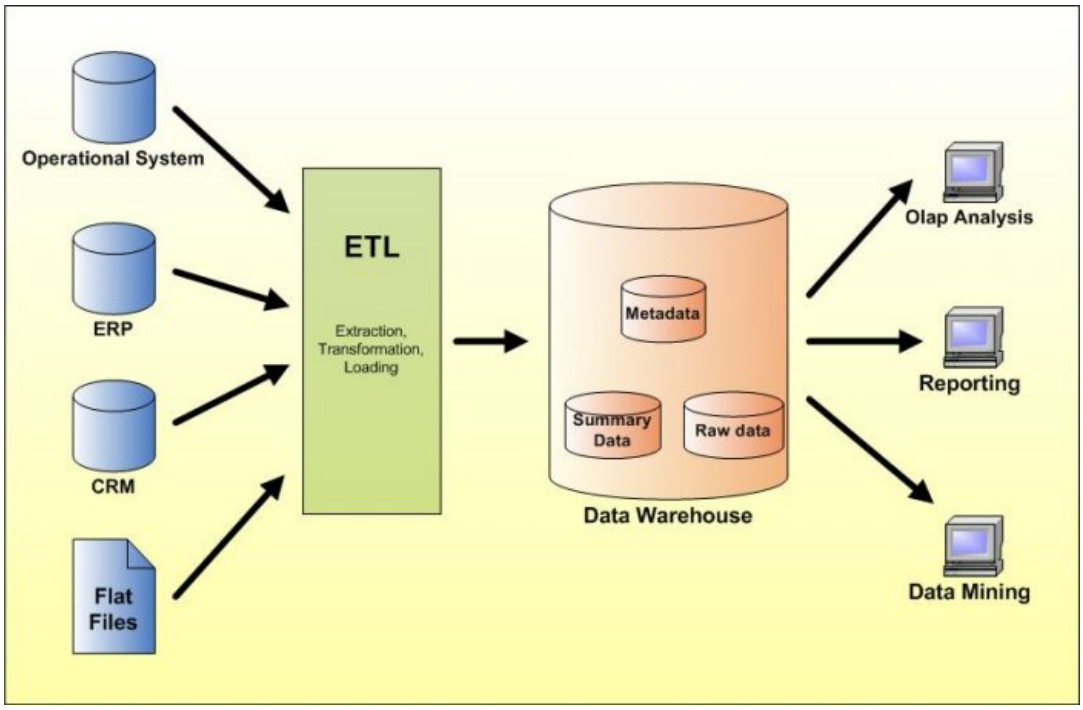

Figure 1 *Datawarehouse architecture*

in a different format than OLTP. The reason for getting the data from the source system into the DDS and then querying the DDS instead of querying the source system directly is that in a DDS the data is arranged in a dimensional format that is more suitable for analysis. The second reason is because a DDS contains integrated data from several source systems. (3), (4).

KNOWLEDGE DISCOVERY IN DATABASES

Knowledge discovery in databases (KDD) is a fast-growing field of research. Its popularity is caused by an ever increasing demand for tools that help in revealing and comprehending information hidden in huge amounts of data. This explosion came about through the increasing use of information technologies. Rich sources of data, stored in databases, datawarehouses, and other data repositories, are readily available but not easily analyzable.

In my doctoral thesis I try to find methodology for extracting the knowledge hidden in the data, too. Figure 1 shows KDD process.

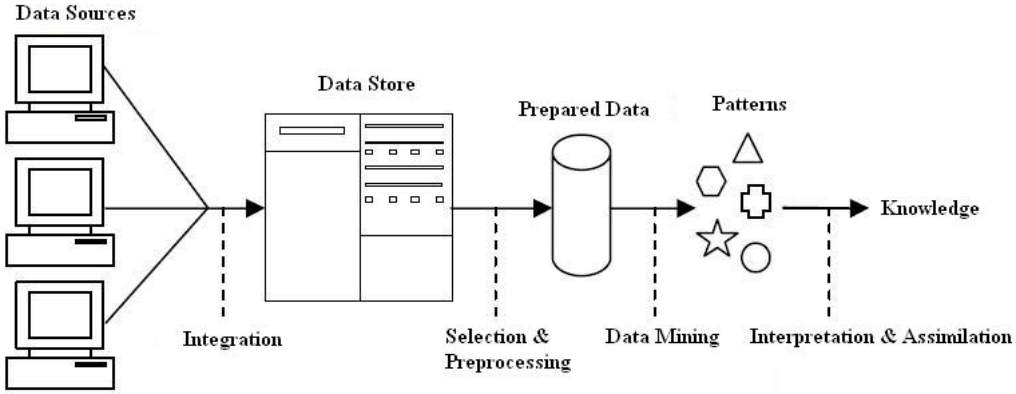

Figure 2 *KDD process*

Data comes in, possibly from many sources. It is integrated and placed in some common data store (datawarehouses). Part of it is then taken and pre-processed into a standard format. This data is then passed to a data mining algorithm which produces an output in the form of rules or some other kind of patterns. These are then interpreted to give new and potentially useful knowledge. (5), (6)
Data Mining is only a part of KDD process. The most widely used method of KDD is CRISP-DM.

CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)

CRISP-DM is the industry standard methodology for data mining and predictive analytics. The current process model for data mining provides an overview of the life cycle of a data mining project. It contains the corresponding phases of a project, their respective tasks, and relationships between these tasks. At this description level, it is not possible to identify all relationships. There possibly exist relationships between all data mining tasks depending on goals, background and interest of the user, and most importantly depending on the data.

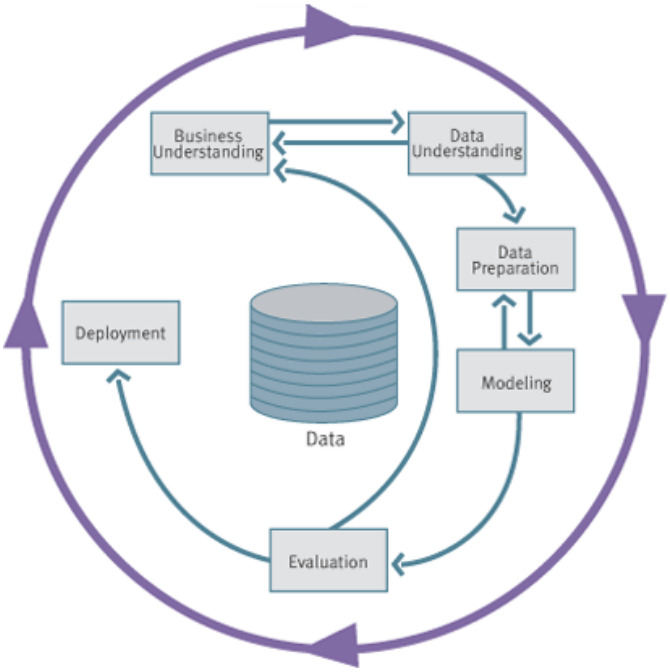

Figure 3 *Phases of the CRISP-DM Process Model*

The life cycle of a data mining project consists of six phases. The sequence of the phases is not strict. Moving back and forth between different phases is always required. It depends on the outcome of each phase which phase, or which particular task of a phase, that has to be performed next. The arrows indicate the most important and frequent dependencies between phases.

The outer circle in the Figure 3 symbolizes the cyclic nature of data mining itself. A data mining process continues after a solution has been deployed. The lessons learned during the process can trigger new, often more focused business questions. Subsequent data mining processes will benefit from the experiences of previous ones. (7)

PHASES OF CRISP-DM

Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.

Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

Data Preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

Evaluation

At this stage in the project you have built a model (or a model) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions will need to be carried out in order to actually make use of the created models. (7)

DATA MINING (DM) TASKS

The main definition of DM is from Gartner Group: Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts

of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.
The following list shows the most common data mining tasks:

- Description – find ways to describe patterns and trends lying within data.
- Estimation – similar to classification except that the target variable is numerical rather than categorical.
- Prediction – similar to classification and estimation, except that for prediction, the results lie in the future.
- Classification – there is a target categorical variable, which, could be partitioned into classes or categories.
- Clustering – refers to the grouping of records, observations, or cases into classes of similar objects.
- Association – finds which attributes “go together”. (8)

I my opinion, one of possible task used in control process can be prediction. The best way to solve prediction in control process is using Neural Network Algorithms. For solving DM tasks I use DM tools Clementine from SPSS.

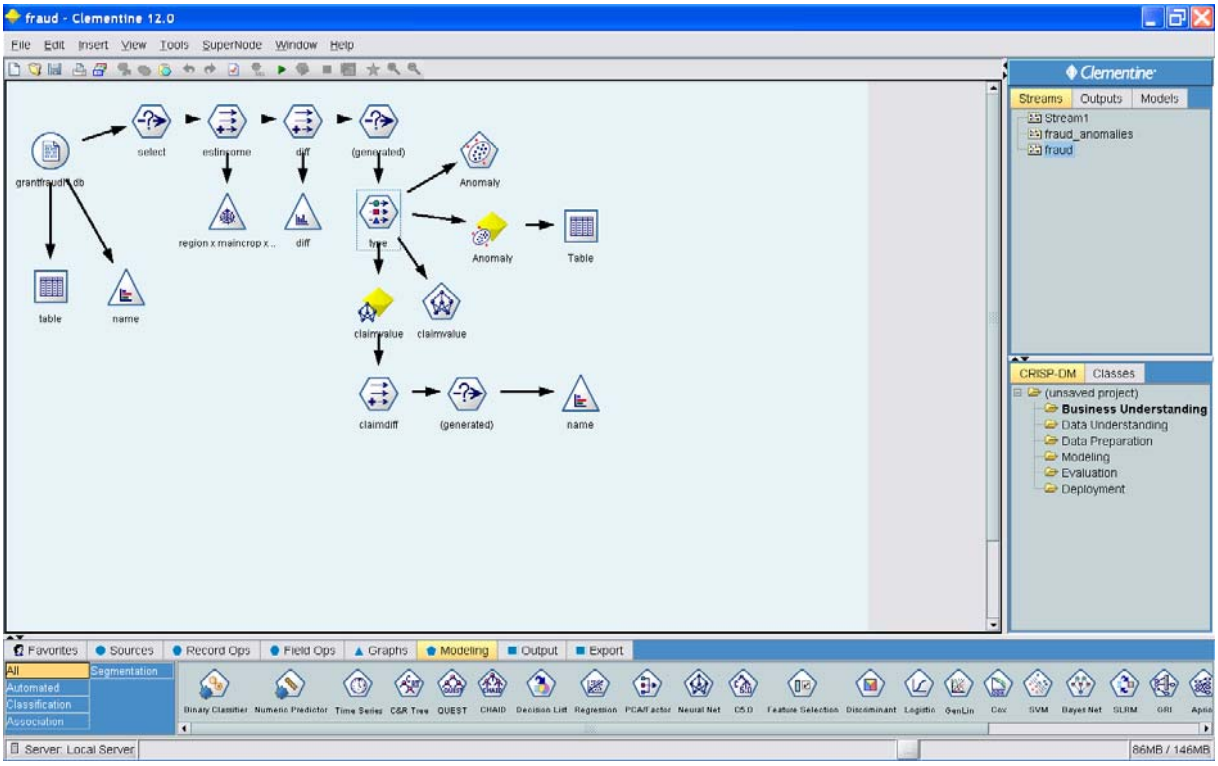

Figure 4 *SPSS Clementine (with Fraud detection example)*

NEURAL NETWORK ALGORITHMS

The basic element of a neural network is a neuron. This is a simple virtual device that accepts many inputs, sums them, applies a transfer function and generates the result, either as a model prediction or as input to other neurons.
A neural network is a structure of many such neurons connected in a systematic way. In Clementine, the neural networks used are feed-forward neural networks, also known as multilayer perceptrons. The neurons in such networks (sometimes called units) are arranged in layers. Typically, there is one layer for input neurons (the input layer), one or more layers of

internal processing units (the hidden layers), and one layer for output neurons (the output layer). Each layer is fully interconnected to the preceding layer and the following layer. For example, in a network with an input layer, a single hidden layer, and an output layer, each neuron in the input layer is connected to every neuron in the hidden layer, and each neuron in the hidden layer is connected to every neuron in the output layer.

The connections between neurons have weights associated with them, which determine the strength of influence one neuron has on another. Information flows from the input layer through the processing layer(s) to the output layer to generate predictions. By adjusting the connection weights during training to match predictions to target values for specific records, the network „learns“ to generate better and better predictions. (9)

Directions for further research

- Define concrete problem in control process.
- Analyze this problem.
- Suggest the methodology to solve this problem.
- Compare used DM methodologies for solving problem.
- Validate the results.

References

[1] INMON, W. *Building The Data Warehouse*. Wiley Publishing, 2005, 543 p., ISBN 0-7645-9944-5

[2] TODMAN, CH. *Designing a Data Warehouse: Supporting Customer Relationship Management.* Prentice Hall, 2000, 360 p., ISBN 0-13-089712-4

[3] RAINARDI, V. *Building a Data Warehouse: With Examples in SQL Server.* Apress, 2008, 523 p., ISBN 1-59059-931-4

[4] Data Warehouse. [on-line] http://www.datawarehouse4u.info/

[5] CIOS, K., KURGAN, A. Trends in Data Mining and Knowledge Discovery. In *Advanced Techniques in Knowledge Discovery and Data Mining.* editor: Nikhil P., Lakhmi J., Springer, 2005, 254 p., ISBN 1-85233-867-9

[6] BRAMER, M. *Principles of Data Mining*. Springer, 2007, 343 p., ISBN 1-84628-765-0

[7] CRISP-DM: Process model. [on-line] http://www.crisp-dm.org

[8] LAROSE, D. *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley, 2005, 222 p., ISBN 0-471-66657-2

[9] Clementine 12 Algorithms Guide, SPSS, 2007, 273 p.

This paper was supported by VEGA agency (project No. 1/4078/07)