Foundations of Management, Vol. 2, No. 1 (2010), ISSN 2080-7279
DOI: 10.2478/v10238-012-0023-x

DATA ANALYTICAL PROCESSING IN DATA WAREHOUSES

Katarzyna ROSTEK

Faculty of Management
Warsaw University of Technology, 02-524 Warszawa, Poland
email: k.rostek@wz.pw.edu.pl

Abstract: The article presents issues connected with processing information from data warehouses (the analytical enterprise databases) and two basic types of analytical data processing in data warehouse. The genesis, main definitions, scope of application and real examples from business implementations will be described for each type of analysis. There will be presented copyrighted method of knowledge discovering in databases, together with practical guidelines for its proper and effective use in the enterprise.

Keywords: data, data warehouse, data management, data analytical processing, OLAP, data mining.

1 Introduction

Current record of economical events is a basic data tool for the needs of operational management. This data is necessary for proper functioning of an organization on the level of everyday business processes. Transactional systems are implemented in order to increase the effectiveness of operational data gathering and storing. However such solutions do not allow making the processed information available what can be extremely helpful in management decision making. One of the reasons for such state of things is the aim of creation of transactional systems as well as the character of data stored in them:

- operational data is usually detailed, whereas managers expect aggregated information,
- data model of transactional base is optimized in order to speed up the realization of basic transactional processing operations and not the analytical processing of this data,
- data needs to be periodically deleted from the transactional system in order to maintain the satisfying level of its efficiency, whereas the analysis usually requires the historical data to be complete,
- enterprises usually have many different transactional systems, while the full picture of the analysis can only be reached on the basis of integrated data originating from all its systems,
- analytical processing of data in a transactional system will result in the drop of its efficiency and current transactional operations will not be processed.

Sharing the solutions built on the basis of transactional systems' resources, directed at analytical, not transactional, needs, becomes a necessity. Such requirements are fulfilled by the decision making supporting systems that are based on data warehouses [16]. The main task of data warehouses is to gather, integrate and process operational data in order to gain data that is new, previously unknown and useful from the business point of view, which can support the decision making process. The aim of this article is to present issues connected with processing information from data warehouses – the analytical enterprise databases.

The article presents two basic types of analytical data processing. The genesis, main definitions, scope of application and real examples from business implementations will be presented for each type. Copyrighted method of discovering knowledge from data, together with practical guidelines for its proper and effective use, will be presented.

2 Meaning of analytical data processing for the decision making process in an enterprise

Currently every enterprise can gather a data collection of a practically unlimited size. For example the data resources of eBay equaled to 2 PB in November of 2006 [24]. RapidShare hosting company shared a disk space of few petabytes with its customers [23] in the October of 2007. Yahoo! website announced that it is the owner of the biggest structured database that was ever implemented in a manufacturing environment, equal to 1 PB [2], in May of 2008. The company also stated that the database will grow to reach 10 PB in 2009.

However, rapid growth of databases in enterprises, organizations and institutes led to limitation in the analysis and interpretation of gathered data. For example Google in the December of 2008 announced that it

is able to sort 1 PB of data in 6 hours and 2 minutes but it needs 4000 computers to realize this task [11].

It is a perfect example of the complexity of the free access issue to the data stored in an enterprise. In order to not only possess but also share the data with potential users, it needs to be organized in sorted and usually central databases, where data warehouses are currently the biggest known databases. Currently the biggest data warehouse is set in the Sun SPARC Enterprise M9000 server and administered with Sybase IQ relation database [3]. It is able to perform a single download of 1 PB of independent and unstructured data. Therefore one can assume that there are tools that allow the enterprises not only to gather data but also put it to effective use. However, the tools and techniques for automatic and intelligent analysis of databases, in order to gather processed information, supporting the management decision making processes became a necessity.

The necessity to make decisions is accompanying the activities of managers since the beginning of organization management history. Development and practical implementation of IT techniques in management allowed decision making process supporting with IT systems, known as decision making aiding systems. One of the first decisions aiding systems' definitions states that they are systems supporting business and organizational decision activities [11]. These systems include applications, which are used to analyze, conclude, simulate, model and evaluate models, schedule or forecast [11]. DSS systems (Decision Support System) prove useful in such decision activities, which do not have procedures leading to optimal solutions and their task is to provide exact, processed information for managers, purchasers and analysts. Their task is to increase the effectiveness of decision making process realization due to limitation of the three following latencies [4]:

- data latency – time necessary to gather and prepare the data for analysis,
- analysis latency – time necessary for analysis and transformation of data into a useful managing information,
- decision latency – time necessary for transferring the data for analysis, interpretation and decision making on needed actions.

Meanwhile, the reaction time, from the moment of the event's start until the moment of making a decision, has a key meaning for the business value of decisions made. The bigger the latency is, the greater is the value decrease (see Figure 1) - even the most valuable information that is not up-to-date will have no significant meaning in the enterprise management.

In order to limit the effects of the latencies, DSS solutions need to assure: access to data coming from many different and dispersed sources, data analysis according to the classification, forecast and simulation methods and finally an easy distribution of this data to the high management responsible for decision making.

Therefore, they are based on data warehouses and analytical processing of gathered data, providing the best functionality and support for decision making, which is based on multilevel historical data analysis. Modern DSS solutions allow reaching for external (not originating form the enterprise) and real-time data (operational), in order to increase the value of the analysis.

The heart of such solutions is the data warehouse. W. H. Inmon [7], the creator of the concept, defines the data warehouse as directed, consistent, chronological and unchangeable data collection. On the other hand, S. Kelly [8], emphasizing the business issues, defines data warehouse as a structure independent of the operational, designed for users who need deep economical knowledge – not only IT knowledge. Moreover he claims that the warehouse structure should relate to the structure of the organization, be unchangeable and reflect the status of the organization in time. Summing up, data warehouse is an independent, but operating within an existing system, directed read-only database, which is used as a basis for decision making support. The main idea of data warehouse is to connect the data from different transactional bases in one base.

Thus data warehouses assure access to gathered resources in a form, character and scope optimized towards the realization of complex multilevel analysis. However, the realization of these analyses is a task for analytical systems, which can be divided into two main groups of tools: current analytical processing and data exploration. Detailed characteristics of both types of data will be presented in the following chapters.

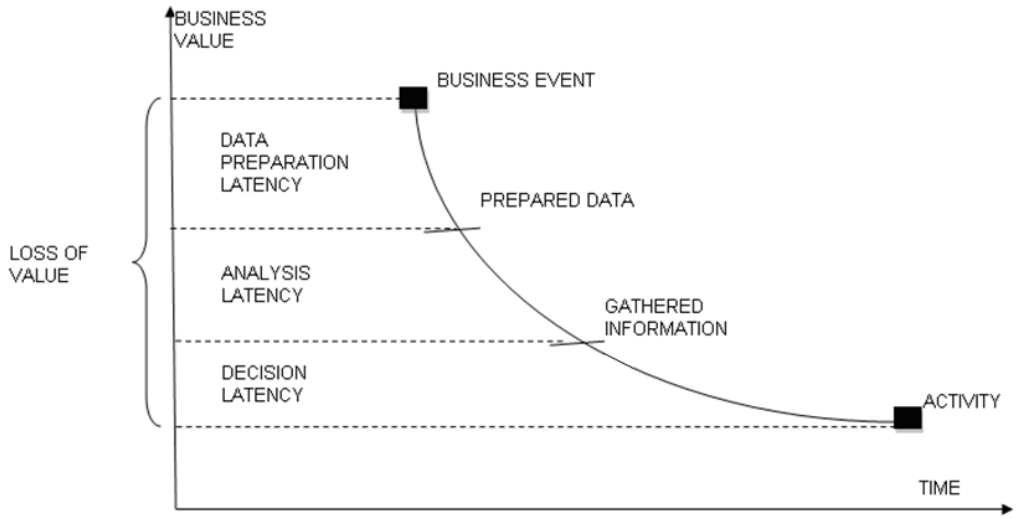


Figure 1. Loss of business value of managing decision
(source: [5])

3 OLAP – OnLine Analytical Processing

The basic type of analysis performed in data warehouses is the current analytical processing OLAP (On-Line Analytical Processing). This concept was introduced in 1993 by E.F. Codd. He claimed that data originating from transactional systems are not sufficient to provide complete answers for the managers. He based his theory of online analytical processing on these observations. The basic idea of OLAP is to allow the users to manipulate data models in many dimensions in order for them to understand the changes taking place in the data. E.F. Codd elaborated a set of 12 rules that distinguish analytical processing OLAP ([1], pp. 12-17):

- multidimensionality – the system needs to assure a multidimensional data view with the possibility to create complex dimension hierarchies system and building data sections simultaneously in relation to many dimensions,
- understandability and usability for the user – users do not need to possess advanced IT knowledge to use OLAP tools; they use commonly known tools to gather data necessary for proper business related decisions,
- availability – OLAP is a middle-ground between operational data sources and the analytical area; used tools should allow finding a proper data source adjusted to a particular analytical case and next performing necessary data conversion,
- collective availability - OLAP tools should support teamwork as well as analysis and idea exchange between the users,
- client-server architecture – OLAP tools should be able to work in the client-server environment; server should be intelligent enough to allow connecting and integration of many different clients with minimal memory usage and amount of programming,
- automatic adjustment of the physical level – OLAP system should adjust its physical scheme automatically in order to adapt itself to the type of model, data volume and the level of data dilution,
- generality of dimensioning – all dimensions must be equal both in the structural and operational possibilities manner; basic data structures, report and form formatting should be equally easy to perform in all dimensions,
- intuitive data manipulation – possibility of direct activities on the data through direct dragging of the objects, realized with a computer mouse, without the need to use complex menu systems or performance of other complex operations,
- unlimited number of dimensions and aggregations – technically it is an unreachable (impossible) feature due to limitations of the hardware that is realizing the processing; in most cases no more than ten dimensions and few hierarchical levels are used; acceding to Codd the maximal number of dimensions should not exceed twenty,
- homogenous efficiency of reporting - reporting efficiency cannot be considerably weakened with

the increasing number of dimensions or the changing of the database size,
- flexible reporting – user should gain every required data view and present it in a way that suits him or her.

Listed specification indicates that fulfilling all of the Codd rules is rather impossible, even though many producers of analytical data processing tools claim the compliance of their products with the OLAP standard. In most cases these systems fulfill only some of the requirements. Some of the manufacturers even modify the general OLAP requirements what is weakening their original meaning. It is relatively easier to gain compliance with other OLAP definition – FASMI from 1995. FASMI stands for Fast Analysis of Shared Multidimensional Information. All of the definition concepts can be explained in the following way ([15], p. 12):
- fast – most questions are performed in few seconds and the basic analysis take less than a second,
- analysis – users can define their own calculations,
- shared – system implements all kinds of data protection, even the ones for single cells,
- multidimensional – the system needs to provide multidimensional, conceptual data view with the possibility to define dimension hierarchy and creation of different data sections,
- information – system provides all necessary information, gathered from data resources of any size.

Summing up, OLAP is a software category, which enables the managers and analysts a free access to information through fast, interactive and broad selection of views that includes properly transformed data and analysis results.

These views reflect the multilevel nature of an enterprise in a way that it is perceived by the user. Users have the possibility to define their own analysis based on available data, creation or modification of analysis dimensions as well as the creation of the aggregates with the use of statistical functions analysis. Presented results can be additionally formatted with the use of: upward and downward drilling, rotation, slicing and dicing, dimension modification, adding own calculating formulas, graphical data presentation, exporting of data to external bases in order to perform detailed analysis (e.g. data exploration). The idea of OLAP analysis is presented below in a simple example.

Example 1

OLAP cubicle, which was created on the basis of a star schema, consists of one fact table and three connected dimension tables (see Figure 2). In this case OLAP's main task is to analyze the data concerning project analysis in a designer-servicing company.

Fact table consists of data: CzasProj (ProjTime), UdzialProc (PercShare), on the basis of which the analytical measures are calculated LiczbaProjektow (ProjectNo), CzasProjektow (ProjTime), and UdzialPracownikow (EmployeeShare). Dimension tables include data, e.g. MiejsceZat (PlaceOfEmpl), Stanowisko (JobPos, RokZak (FiscYear), MiesiacZak (FiscMonth), which will be the analytical sections. The structure defined this way allows to determine, which factors influence the number of realized projects, time scale and the level of employee engagement.

Created OLAP cubicle allows elaborating analytical perspectives (see Figure 3), where the cubicle dimensions stand for analytical sections, e.g. RokZak (FiscYear), MiesiacZak (FiscMonth), DzienZak (FiscDay), MiejsceZat (PlaceOfEmpl), Stanowisko (JobPos), Staz (Seniority) and the analyzed parameters are its measures, e.g. LiczbaProjektow (ProjectNo). Perspective presented in the Figure 3 allows determining the relations between the number of realized projects and the project finish data (with the division for year, month and day of the finish) as well as the project team (with the division for parent department, job position and job tenure).

Obtained pivot table can be also analyzed with the following methods:
- pivoting – determining measure and defining dimensions in which the selected measure will be presented,
- drilling down – is based on diving in the hierarchy of a particular dimension in order to perform a more detailed data analysis,
- drilling up – is based on navigating upwards particular dimension's hierarchy in order to perform an analysis on a higher level of dimensions hierarchy,
- rotating – allows to present the data in different layouts,
- slicing and dicing – allows narrowing analyzed data to selected dimensions and in terms of selected dimension – narrowing the analysis to specific values,
- ranking – allows data ordering in particular dimension according to the values of selected measures.

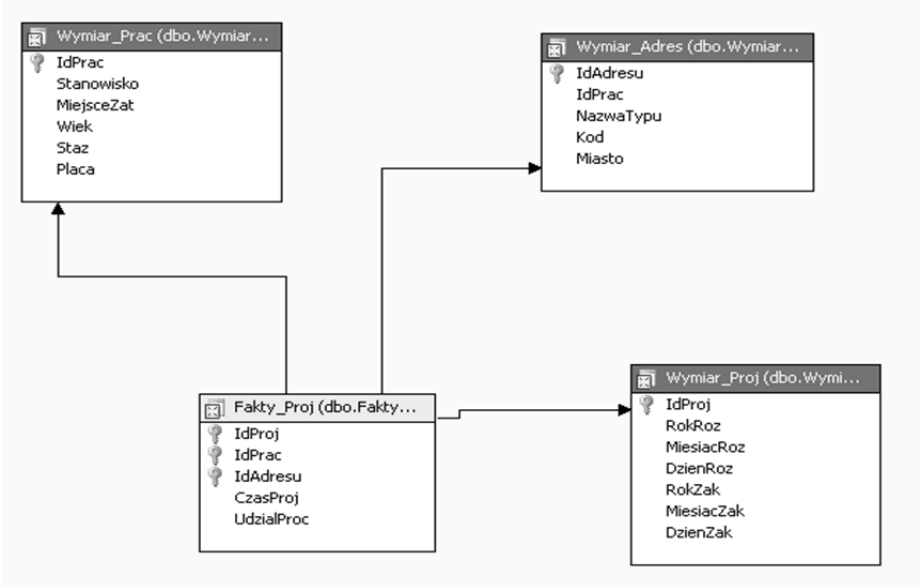


Figure 2. OLAP cubicle
(source: self study)

	Rok Zak	Miesiac Zak	Dzien Zak	
	⊞ 1996	⊞ 1997	⊞ 1998	Suma końcowa
Miejsce Zat / Stanowisko / Staz	Liczba Projektow	Liczba Projektow	Liczba Projektow	Liczba Projektow
⊞ 100	170	7	457	634
⊞ 110	30		328	358
⊞ 120	191	4	467	662
Suma końcowa	391	11	1252	1654

Figure 3. OLAP cubicle analytic perspective
(source: self study)

			Rok Zak	Miesiac Zak	Dzien Zak	
			⊞ 1996	⊞ 1997	⊞ 1998	Suma końcowa
Miejsce Zat	Stanowisko	Staz	Liczba Projektow	Liczba Projektow	Liczba Projektow	Liczba Projektow
⊟ 100	⊞ analityk		60		181	241
	⊞ kierownik		18	1	31	50
	⊞ projektant		68	6	144	218
	⊞ sekretarka		24		101	125
	Suma		170	7	457	634
⊟ 110	⊞ analityk				69	69
	⊞ kierownik				64	64
	⊞ programista				101	101
	⊞ projektant				84	84
	⊞ sekretarka		30		10	40
	Suma		30		328	358
⊟ 120	⊟ analityk	5			141	141
		11	18	1	9	28
		16	19		37	56
		Suma	37	1	187	225
	⊞ grafik				52	52
	⊞ kierownik		8		4	12
	⊞ programista		130	3	137	270
	⊞ projektant		16		8	24
	⊞ sekretarka				79	79
	Suma		191	4	467	662
Suma końcowa			391	11	1252	1654

Figure 4. Data drilling view
(source: self study)

Figure 4 presents an example of data drilling reached with dimension particularization for: MiejsceZat and Stanowisko.

OLAP analysis results can be directly used in the enterprise management process. Real life implementation of decision making supporting system, based on data warehouse and the OLAP tool, is presented below.

Example 2

Carbon S.A., company that belongs to SGL Carbon Group, is a leading manufacturer of carbon and graphite products. In the end of the nineties a problem of lacking data emerged, which included the whole production process [13]. This production process is extremely long (lasts from 1 to 3 months) and the production includes few hundreds of different products for many different customers. Tracking such amount of production cycles in a transactional base is very difficult. Therefore aggregated quantitative information about particular products was essential for effective management of production processes. Proposed IT solution, which was to improve the production process data access, was the MEDIA Management Information System, which was supposed to function on the basis of the data warehouse resources oriented towards production process and its technological parameters (see Figure 5).

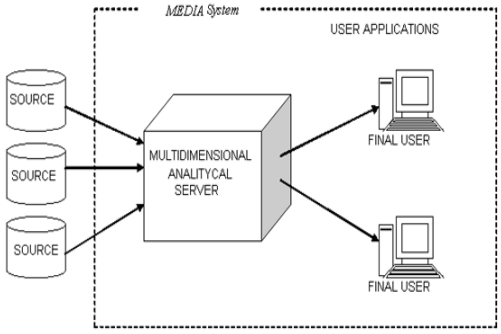


Figure 5. OLAP system in SGL Carbon
(source: self study on the basis of [13])

PMS Labs together with SGL Carbon employees tried to integrate quantitative information with selected technological parameters originating from production objects, product groups and particular products. This allowed to process management improvement through easier access to production data, dispersed in transactional bases and automatic production planning systems. Data warehouse was designed in MOLAP technology (Multidimentional OLAP), which uses multidimensional tables to store data. Such tables include different data from many sources.

Creation of proper relation structure makes it easier for the users to select interesting information e.g. product from certain batch, product from given operation in the production process.

Currently the system is gathering data from many transactional systems, aggregating it and enabling the tracking of a production process. Inbuilt tools, which enable the creation of hierarchical structures, allow performing the data drilling operations and the dynamic changes result data presentation form as well as creation of individual inquiries. Due to these functions the user is able to quickly analyze the production process flow for a single product or the groups of products. In the last level of detail, in order to further drill information, there is a possibility to engage transactional programs with source data originating from production process directly from the MEDIA system. SGL Carbon enterprise gained the following benefits due to the OLAM system implementation [13]:

- easier access to the production data,
- graphic environment of analysis and reporting,
- possibility to perform analysis for groups of products,
- online access to indicators that were previously inaccessible or needed to be calculated every time,
- working in a uniformed software environment.

Summing up, OLAP system is implemented in a mode designed for multiple users. It offers quick answers for inquiries and questions, independent of their size and the complexity of the database. It helps the analytical system user to synthesize the information about the enterprise via a comparative insight into its resources as well as through the analysis of historical, current and forecasted data in "what if?" analysis. One of OLAP's characteristic features, being the system's drawback at the same time, is that its activity is based on the verification of hypothesis constructed by the analyst. Therefore the effectiveness and usability of this method is, to high extent, dependent on the imagination and creativity of the people performing the analysis. OLAP technology assumes that the user has an absolute knowledge about the object of the analysis and can control this process. That is why the analyst is forming the questions and performs the analysis of the data stored in the warehouse. The second group of analytical data processing – data mining – is independent of this constraint.

4 Data mining

Data mining enables the data analysis for such problems, which were difficult to solve for humans (due to its volume) or the ones difficult to solve because of lacking knowledge (that is to be mined from the data). Data mining methods are based on effective revealing of new relations and connections between pieces of data. Automatic data mining brings new possibilities in the scope of user-warehouse data system interactions, enabling formulation of questions and inquiries on a more abstract level than the OLAP technology would allow. Data mining is treated as searching process leading to revealing the interesting regularities, patterns or anomalies in the data from the point of view of the user. These objects include information that is particularly interesting for the user and should aid the creation of e.g. OLAP cubicles. Such knowledge is new and unknown to the user. Business value of this knowledge is one of the evaluation criteria for the KDD - Knowledge Discovery in Databases process.

Knowledge discovering is one of the interdisciplinary sciences, which combines the elements of statistics, econometrics, computing, artificial intelligence methods and the database theory. Knowledge discovering concept first emerged in the literature in the late eighties and since that time it has many different definitions. One of the most popular definitions is the one published in 1991 by Frawley and Piatetsky-Shapiro [16], which describes the knowledge discovering process as: significant process of identification of important, new potentially useful and understandable data patterns. Knowledge discovering is a process of searching for new relations, tendencies and regularities, which can be observed due to detailed data analysis, gathered over a period of time. The greater the data collection and the longer the research time period is, the analysis possibilities are more significant as well as the results are of better quality. At the same time, the knowledge discovering process is an intensive process of cooperation between a human user and an IT system. It is the user who defines interesting and useful phenomena discovered during the research. On the other hand, the IT system is performing a preliminary information selection in order to allow the user to get to the most interesting pieces of information, presenting them in the form of a list e.g. according to the greatest degree of credibility.

It is important to highlight that data mining is only a stage of the analysis in the knowledge discovering process. Few definitions of this concept will be presented. David J. Hand [6] defines it as: science dealing with data gathering from large data collections or databases. M. J. Berry and G. S. Linoff [1] claim that this concept can be defined as: mining and analysis of large amounts of data in order to find significant relations, schematics, patterns and rules. According to A. Sokołowski [19] data mining is a process of finding interesting patterns, relationships, anomalies, hidden structures in large data collections gathered in data warehouses (without a determined research aim) what is typical for data mining is the entity of unsorted data and he necessity to use computers.

One of the crucial features of data mining is the possibility to realize it in two ways:

- main aim of the mining research is known from the beginning e.g. searching for factors collection that are responsible for loosing an employee to the competition,
- main aim of the mining research is not known from the beginning e.g. search for new, interesting and useful relations and patterns hidden in a data collection designated for the research.

Data mining, other than regular statistical analysis (usually one-dimensional), includes the simultaneous influence of many factors on researched phenomenon. This enables to distinguish feature sequences, usually connected with mutual hierarchies, which presence in the population increases or decreases the possibility of certain phenomenon's occurrence.

Data mining is a natural extension and supplement of data warehouses creation, which means organization of large, multidimensional data collections that aid analytical data gathering process.

Data mining techniques can operate on any kind of unprocessed data as well as they can be used to browse and compile data generated through OLAP inquiries and in this case can provide much more detailed and deeper multi-aspect knowledge. Data mining is an analytical approach that is the expansion of the OLAP techniques.

Mining analysis methods can be divided into eight basic classes [14]:

- association revealing – the most broad of all classes that is bases on revealing of new, unknown and interesting relations or correlations between data,

defined as associations; association rules collections and sequence patterns are the result of the association methods,

- classification – includes model revealing techniques (classifiers) or the functions describing relations between set object classification and their characteristics,
- prediction – using well-known and recognized classifiers to describe new objects with unknown classification,
- grouping (data clustering) - includes data analysis methods and finding of finite sets of object classes with similar features,
- revealing of singular points – includes the techniques of singular point finding, which differ from the general data model (known from the classification and prediction methods) or class models (cluster analysis); usually the methods of revealing of singular points are the integral part of other data mining methods e.g. grouping methods,
- time tracking analysis – includes time tracking analysis methods used to find: trends, similarities, anomalies and cycles; revealed descriptions can have the form of characterizing or discriminating rules,
- trends and deviations analysis – includes the variable data analysis changes over a time interval in order to find differences between current and expected data values,
- exploration of selected data types – includes the following methods: spatial data mining, multimedia, text and websites.

Data mining is using all kinds of methods that allow creating knowledge from data (relationships, patterns, trends) and therefore it does not have determined standard techniques. However, there is a considerable collection of algorithms that are often used in mining and exploration research. This collection includes: association algorithms, regression, decision trees, neural networks, time series or Bayes networks. Proper selection of a data drilling techniques depends on the type of the problem. Some algorithms dedicated to specific tasks, can be highlighted in Table 1.

Mining research scenarios can also be determined, which would allow to gain certain knowledge concerning objects or events (see Table 2).

The idea of mining analysis is presented below on the basis of a simple example.

Table 1. Selection of data mining techniques for given analytical tasks
(source: self study on the basis of [22])

Analytical task	Mining algorithm
Discrete variables forecasting, e.g. does a certain customer use promotions tailored in marketing campaigns	- decision trees - clustering - naive Bayes classifier - neural networks
Constant variables forecasting, e.g. certain product sales forecast in the upcoming year	- decision trees - time series
Sequences revealing, e.g. click-stream in one session, specific for a certain website	- sequential clustering
Association revealing, e.g. stock transaction or purchasing portfolio analysis	- decision trees - association rules
Grouping, e.g. finding homogenous customer groups	- clustering - sequential clustering
Discrimination analysis, e.g. establishing characteristic product features of leading products in relation to poor performing products	- decision trees - regression

Table 2. Data mining scenarios
(source: self study on the basis of [21])

Detailed task	Mining algorithm
Main task: maximization of profits generated by customers	
Basic customer classification	- clustering
Establishing characteristic product features of high-profit customers	- decision trees
Customer preference recognition	- association rules
Customer behavior analysis	- sequential clustering
Potential customers profitability forecast	- neural networks
Main task: Construction of successful marketing campaigns	
Customer segmentation	- clustering - decision trees
Research on customer reactions on marketing campaigns	- decision tree - Bayes naive classifier - clustering - neural networks
Selection of sub-optimal campaign variant	- cluster analysis
Campaign receivers prediction	- neural networks
Strategy evaluation in relation to marketing campaign reactions	- listed models updating
Main task: Fraud detection and prevention	
Transaction sequence research	- sequential clustering
Unusual event detection	- neural networks - decision trees - clustering
Model update according to real-time data	- listed model access from the level of metamodels

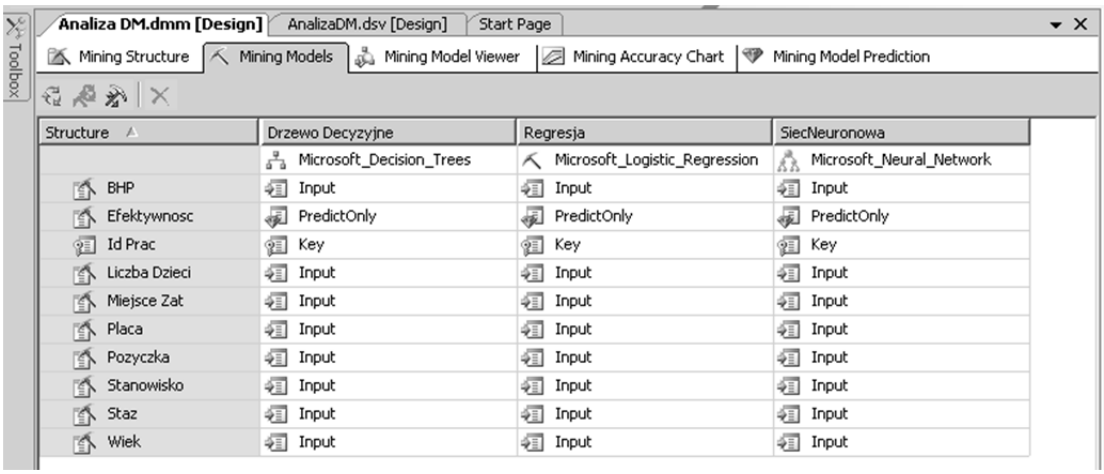
Example 3

OLAP analysis data, in relation to example 1, turned out to be interesting enough to perform a detailed analysis with a mining model. Only 60 of 450 employees of a company take active part in the realization of projects. Board of the enterprise tries to activate its employees (or trigger changes in the employment structure) as well as wants to find out what kind of features of current and future employees should be welcome and expected. Therefore an additional variable, Effectiveness, was assigned to every employee in the data warehouse.

Employees, who participated in more than 10 projects and were engaged by more than 30% in at least one of them, will have the variable value equal to 1. Remaining employees will have a 0 mark for the Effectiveness variable. Such defined variable allows to perform classification analysis, which aims at description of an employee that is effective and committed in the project activities of the enterprise.

Three data mining techniques were used to perform this analysis: decision trees, logistic regression models and neural networks. Variables, which will be used in the model, need to be determined before the start of the analysis (see Figure 6). Effectiveness, the explained variable, in the models is marked as PredictOnly. Explanatory variables are marked as Input in the model. Exploration analysis reveals such values of explanatory variables that have the biggest influence on the value (1 or 0) of the explained variable.

Establishing and starting of the model will trigger the presentation of the results, which can have the graphical form (see Figure 7) or the descriptive form (see Figure 8).



Structure	Drzewo Decyzyjne	Regresja	SiecNeuronowa
	Microsoft_Decision_Trees	Microsoft_Logistic_Regression	Microsoft_Neural_Network
BHP	Input	Input	Input
Efektywnosc	PredictOnly	PredictOnly	PredictOnly
Id Prac	Key	Key	Key
Liczba Dzieci	Input	Input	Input
Miejsce Zat	Input	Input	Input
Placa	Input	Input	Input
Pozyczka	Input	Input	Input
Stanowisko	Input	Input	Input
Staz	Input	Input	Input
Wiek	Input	Input	Input

Figure 6. Mining models definition
(source: self study)

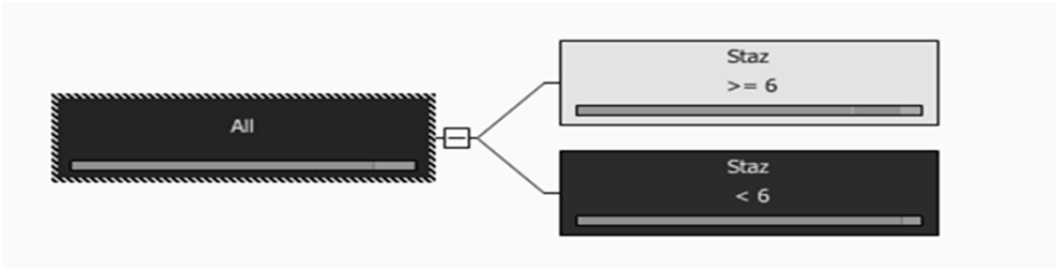
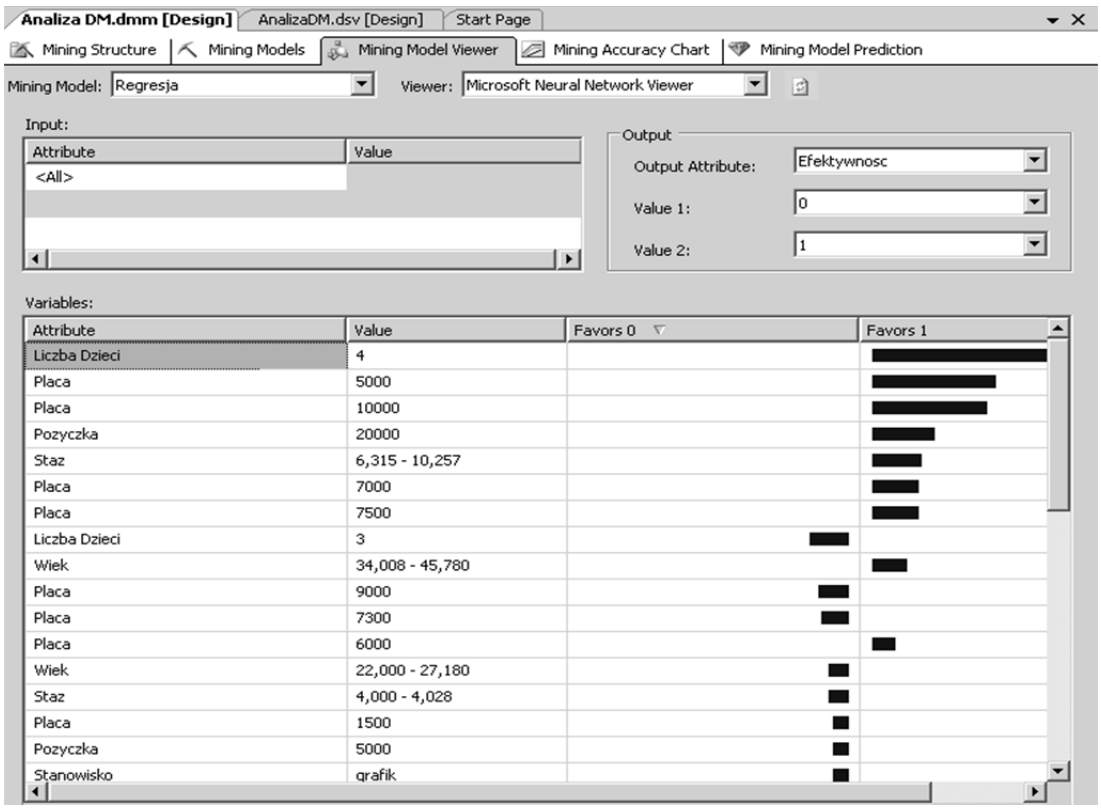


Figure 7. Decision trees model results
(source: self study)



Attribute	Value	Favors 0	Favors 1
Liczba Dzieci	4		
Placa	5000		
Placa	10000		
Pozyczka	20000		
Staz	6,315 - 10,257		
Placa	7000		
Placa	7500		
Liczba Dzieci	3		
Wiek	34,008 - 45,780		
Placa	9000		
Placa	7300		
Placa	6000		
Wiek	22,000 - 27,180		
Staz	4,000 - 4,028		
Placa	1500		
Pozyczka	5000		
Stanowisko	grafik		

Figure 8. Logistic regression model results
(source: self study)

When analyzing the model results one can prepare a profile of an effective employee: it is a middle-aged person with numerous family and working experience, considerable salary and commitment to the company due to a relatively high loan, and the ineffective employee: young employee, low salary, low working experience – one of such examples is a graphics designer. This interpretation can lead to another analysis, which will try to find an answer to the question: why graphics designers do not willingly participate in enterprise projects. Maybe the work style does not allow the designers to be active as well as relatively low salary forces them to simultaneously work outside in another company to gain satisfactory salary level what results in spare-time limits and low commitment of this group. In this way data mining explores new, previously unknown, information that broaden the knowledge about the enterprise, its activities, customers, employees, threats, competition etc.

Example of a data mining system implemented in an insurance company is presented below.

Example 4

This example is a result of research performed by J. Kowalska and B. Trawiński [10] in a real time sales data of a manufacturing enterprise. The data included:

- product sales invoices issued in the first nine months of 2005 – products were characterized with the following attributes: brand, type, category, model, function, usage, size, color, offer etc.,
- product receivers – warehouses, shopping centers,
- sales managers.

Data from the relative structure were transformed into a flat structure, built upon invoice positions, including the quantity, values and dates of sales of particular products. The structure also included the product and receivers attribute values. In this stage of preparation the incoherent and zero sales data was rejected. About 65 thousand records were used to research customers and sales. The analysis was performed with the SAS Enterprise Miner tools.

Data mining included the factors influencing the sales value, sales prediction, sales volume increase, sales managers' evaluation and customer grouping. The following analysis techniques were used in this research: decision trees, regression, association rules, graphical visualization and cluster analysis.

Sales value increase factors analysis was performed with decision trees. Products were divided into product groups with similar average value and sales volume. High-runners (leading products) and low-runners (poor performing products), which need to be included in joined sales or sales promotions, were distinguished. Desired production volume was also determined for every of researched products.

Regression was used as supplementary to decision trees analysis. It allowed predicting the sales transaction values scope and not only the average value (determined with decision trees). This led to distinguishing of product groups with low sales volume and preparation of corrective actions.

Association rule analysis allowed determining which products can be included in joined sales and sales promotions. The aim of this analysis was to increase the sales of the products with low demand.

Graphs and figures were used to present the changes of volume and value of sales in particular provinces and for particular sales managers. This allowed the high management to evaluate the effectiveness of their employees. Managers were selected to service and support key customers as well as managers for less important customers and sales parts of lower importance.

In order to differentiate customers cluster analysis was used. Four customer groups were distinguished as a result of the algorithm activity:

- group 1 – customers who buy many products with medium or low margin; this group includes big number of transactions ant the customers generate large profits – this group does not require any corrective actions,
- group 2 - customers who buy few products with high price and medium margin; this group is especially valuable for the company but the volume of sales in this group should be increased e.g. through adding low-cost products as a bonus to the regular products, without any changes in product's price,
- group 3 – customers who buy very little products with low price and low margin; this group generate low profits – this group requires actions to increase the sales volume e.g. through researching the most popular products and offering sales promotion in this segment,
- group 4 - customers who buy little products with medium price and low margin; corrective actions for this group include actions similar to the ones' for the third group of customers.

Performed mining research led to the formulation of conclusions helpful for sales and marketing departments. It also led to the formulation of new research questions and issues that were the subject of following mining research. Therefore the effectiveness of production, marketing and sales departments increases with the maintenance or decrease of the functioning costs.

How is it possible to effectively organize and realize this type of research will be presented in the following chapters of this article.

5 Knowledge discovering methodology

Many IT tools that support the knowledge gathering processes were created. However even the best programs cannot solve all problems connected with the realization of the process and cannot assure the success of the endeavor. Effective and efficient methodology is necessary (see Figure 9).

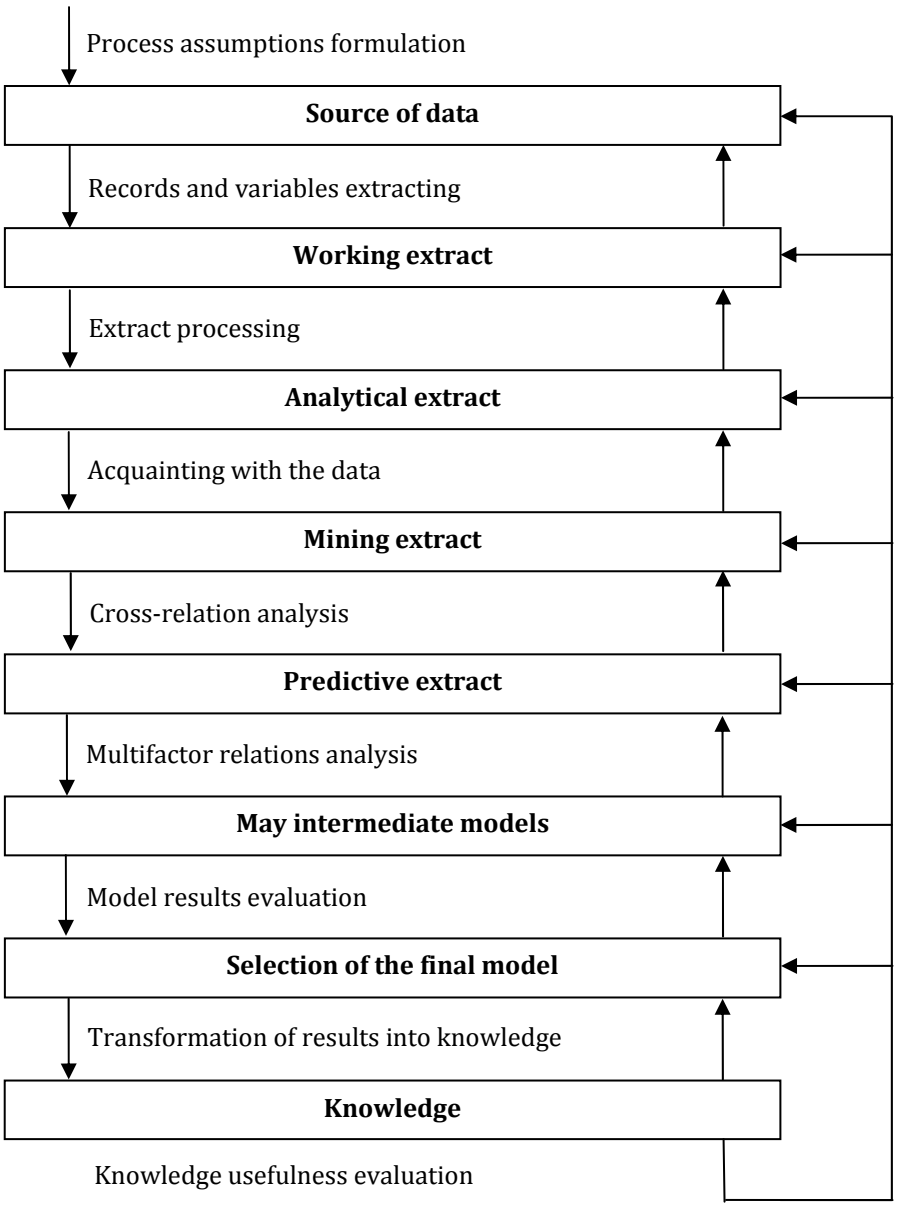


Figure 9. Knowledge discovering process methodology
(source: self study)

5.1 Project assumptions formulation

Knowledge discovering process assumptions need to ambiguously define user expectations in relation to the results of the whole process. Elements that should be included in the assumptions formulation are as following:

- aim of the research – defines the direction of the research and describes the modeled event,
- scope of the research – timescale and research scope determination,
- logistic parameters of the process – define the methodology of the process realization, process realization methods, place of realization, project coordination, research team, process consultants team,
- technical parameters of the process – define the software, network resources, programming and other tools necessary for the realization of the process,
- process realization schedule – presents successive process stages with the detailed information on: stage start date, stage realization characteristics and final stage result presentation, stage finishing date,
- process realization cost calculation – needs to include the following elements: cost of machine tools, cost of research team training, cost of process data preparation, and cost of consultancy and data warehouse usage, server and personal computer capacity.

Detail and accuracy of the process assumptions formulation influence the usefulness of process results and discover knowledge as well as if the process realization will have to be repeated.

5.2 Records and variables extracting

Creation of draft extract is based on the selection of proper records and variables, necessary for the knowledge discovering process realization and placing them in a single table. Time period of this scale directly depends on the quantity and structure of source data, which are the basis for extract data. If we speak of a single centralized data source, where identification of full history of observed extracts does not involve many connecting and searching operations, this stage will be relatively easy to realize. However, this matter complicates when instead of one data source we have a dispersed data source, which has the characteristics and history of observations collected in many tables or data sets.

Before the start of data extracting it is crucial to ambiguously identify proper objects in data collections in order to reach full consistency of the data extract. With a complex data source it is necessary to elaborate programming tools, which in the following stage will automatically extract the records to the extract. It is a guideline for the enterprises, which plan to implement the data warehouses. Easy and quick extract generation possibility should be included in the design phase of the data warehouse structure.

5.3 Extract processing

Extract input data can have: diversified format (e.g. date of birth in short or long format), inconsistent values (e.g. sex of particular person can be determined as male and female at the same time) and can be "rough" (e.g. include city name in the field destined for car type). During the realization of this stage it is necessary to unify data formats, define allowable values (dictionary) for the variables and eliminate inconsistency in variable fields values. Many of such activities will be repeated in following extracts, therefore it is necessary to prepare IT tools to optimize the extract processing. Another of problems present at this stage is the lack of data. Variables with 20% of missing data are not relevant for the analysis. If the missing data exceeds 50% it should not be included in the analytical extract, because it would determine the process results. The simplest solution in case of missing data is the elimination of the variable from the extract. However, in many cases one cannot allow to loose a variable, which can have a significant influence on the researched event. In such case it is usually necessary to complete the missing data. There are many methods and algorithms in such cases and every tool designated to discovering knowledge from data has its own solutions. Moreover one can use own implemented data completion methods. Nevertheless it is important to remain cautious. Improper data completion can result in the analysis of the lost data, instead of real data, in next stages of the process.

Preparation and processing of a draft extract, in order to create an analytical extract, can become the most time-consuming part of the process and take up to 80% of the time designated for the realization. Therefore it is crucial to bear in mind that the stages will be process efficiently if the data source is prepared properly.

As a result it is necessary to predict realization of such processes during the construction of data warehouses.

5.4 Acquainting with the data

Main task of this stage is a detailed analysis of the variables that are a part of the data extract: values, distribution and description of the variables with the use of descriptive statistics values. Variables determined at this stage are usually: allowable values collection, mean average, minimum and maximum value in the set of observations, standard deviation, and median.

After acquaintance with the distribution of each variable, usually extreme observations that do not fit the entity are eliminated, if they can influence and disturb the analysis results (e.g. changing the value of mean average and median). Sometimes such extraordinary observations can be side effects of the process and become the input data for further analysis. All observations with monovalent and quasi-monovalent distributions are eliminated from the extract (they do not influence the analysis).

If regression models are to be used in following stages of the process, it is justified to transform qualitative variables into zero-one values in order to simplify the analysis and improve the quality of the predictive model.

This stage must end with an absolute knowledge of all variables, their values, distribution and influence on the researched event, because without this knowledge it is not possible to realize further stages of the process.

5.5 Cross-relation analysis

This analysis is based on the use of all possible techniques and statistical methods, which allow researching distributions of all explanatory variables in the particular values of the explained variable and determine the influence of individual variables on the research event.

Explained variable is the one which determine the aim of the modeling. For example in the research of the population of customers who break up insurance policies, Policy Value will be the explained variable and take two states “Broke the policy” or “Did not break the policy”. Every mining analysis must have defined at least one (and usually one) explained variable. Remaining variables in the set will be explanatory. Some of the used methods are:

- cross-relation tables,
- distributions of variables broken into the values of explained variable,
- values of descriptive statistics according to the value of the explained variable,
- correlation matrix,
- statistical cross analysis e.g., Kaplan-Meier analysis.

One of the simplest methods to research elations between explained and explanatory variables is the creation of cross-relation table (Table 3). It allows to research both the numeric and percentage data value distribution.

Table 3. Cross-relation table of explained and explanatory variable
(source: self study)

Cross-relation table of explained/explanatory variable			
Explanatory variable	Explained variable		Sum
	0	1	
0	117907 61%	40842 21%	158749 82%
1	32 0,5%	18 0,5%	50 1%
5	884 0,5%	270 0,5%	1154 1%
10	23544 12%	9100 4%	32644 16%
Sum	142367 74%	50230 26%	192597 100%

Another form of analysis at this stage is the research of the distribution in graphical form. Numeric and percentage distributions presented both in tables and graphically should always be completed with descriptive statistics values.

If on the basis of cross-relation tables and distribution graphs of constant variables one will discover their non-linear influence on the explained variable, classification of these variables in to a finite number of groups – classes needs to be performed (it is best to analyze variables grouped in 2 to 6 classes).

Preparation of mutual variable correlation matrix is the following task of this stage (Table 4). Only one variable is left in the mining extract from the group of correlated variables (correlation index of these variables is higher than the determined value e.g. 0,5) on the basis of the matrix results. Selection of one variable from every group of the correlated variables is a task for the analyst. One of the common causes of the drop of effectiveness and quality of the predictive model is the presence of few correlated variables in the model.

Table 4. Table of mutual correlation of variables
(source: self study)

	va1	va2	va3	va4	va5
va1	1,0000	-0,2953	-0,0195	-0,0098	0,0975
va2	-0,2953	1,0000	-0,0451	-0,0288	0,0338
va3	-0,0195	-0,0451	1,0000	0,8048	-0,0824
va4	-0,0098	-0,0288	0,8048	1,0000	-0,0465
va5	0,0975	0,0338	-0,0824	-0,0465	1,0000

It is possible to observe strong single-dimensional influences of the explanatory variables on the explained variable at this stage. Interesting issues and production anomalies in the population can be identified and distinguished from the mining extract. Such case can influence the direction of further studies as well as the method and techniques selection of the predictive models. However, one should not expect that the cross-relation analysis will bring unexpected results (even if the results seem extraordinary it is necessary to verify them with a predictive model). At the end of this stage variables that create predictive extract are selected. These variables always include the explained variable and a set of explanatory variables, which can differ between different predictive models.

5.6 Multifactor relations analysis

Multifactor relations analysis, in other words data mining, focuses on building of a sub-optimal mining model. Mining models should be constructed on the basis of a certain part of the extract (usually about 40%), namely the training set. Remaining part of the extract is divided between the validating and testing sets (usually about 30% each). These sets are used to teach the models and evaluate the effectiveness of its successive variants. Another important issue of this stage is the selection of proper analytical methods and techniques, used to build mining models.

Each analytical model brings result information. However, it is necessary to evaluate the usefulness of such information and the possibility to use it in particular business activities. For example, discovered relation that the customer will not break the 20 year insurance policy during upcoming 10 years with the probability of 99,9% is not very useful for an insurance company representative. It does not mean that the information is useless. It can be applied elsewhere e.g. to elaborate the tariff for the insurance fee payment for such policies. This example indicates that the selection of the final model and research results evaluation is connected not only with statistical evaluation of the correctness and effectiveness of the model, but also with the subjective evaluation of the customer – user of the knowledge.

5.7 Model result evaluation

Every mining tool must enable an unambiguous evaluation of correctness of model's results. Such evaluation consists of two elements:

- statistical evaluation of reached results,
- model use effectiveness evaluation.

Statistical evaluation is based on the determination of the statistical error, which is accompanying model's results (Table 5). Collects the results indicate the strength and direction of the influence of population's particular features on the researched event. Value of the estimator is determining the influence. However, every value of the estimator has a certain statistical error, which determines in how many real cases determined value was estimated wrongly. Allowable error of the estimator is between 5-7%.

Table 5. Statistical evaluation of model's results
(source: self study)

Feature	Estimator value	Estimator error
Education = Higher	0,3077	0,0456
Education = Secondary	0,2567	0,0022
Education = Vocational	-0,6780 ↓	0,0034
Sex = Female	0,4567 ↑	0,0076
Sex = Male	0,1234	0,0123
District = Mazowieckie	-0,2222	0,0099
District = Małopolskie	0,8677 ↑	0,0233
District = Pomorskie	0,0900	0,0001

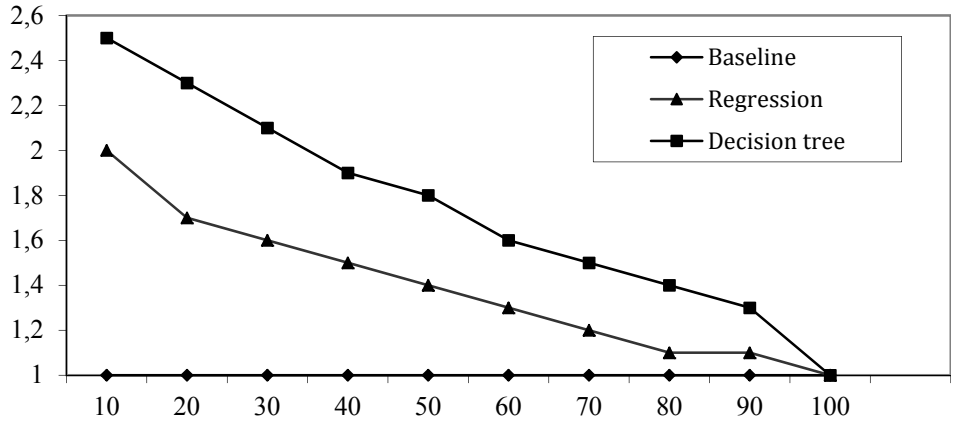


Figure 10. Raising graph
(source: self study)

Table collects the results that indicate the strength and direction of the influence of population's particular features on the researched event. Value of the estimator is determining the influence. However, every value of the estimator has a certain statistical error, which determines in how many real cases determined value was estimated wrongly. Allowable error of the estimator is between 5-7%.

Second element of the model evaluation is the determination of its effectiveness. Such evaluation consists of two main elements:

- number of times the analytical model is better than a random sample,
- evaluation of number of times that the model result was confirmed and he number it was wrong in real life.

In order to evaluate the effectiveness of a model in relation to the random sample one needs to use the raising graph (see Figure 10).

This graph should be interpreted as follows:

- if the 10% trail of the whole researched population is casted or the 10% trial of the whole population is selected with a use of a model, model trial will have 100% higher prediction than the random trial for the regression model (regression forecasting will be 2 times more accurate that the random trial) and with 150% higher prediction that the random trial for the decision trees (decision trees predictions will be 2,5 times more accurate than the predictions of a random trial),
- if one would cat or select a 50% trial of the population, regression model will be better by 50% than the random trial and the decision trees model will be better by 90%.

Usually models selected for implementation are characterized with 2-times higher efficiency than the random trial for the 10% of casted or selected population.

Whereas for the evaluation of event classification accuracy with the use of a model an event classification graph is used (see Figure 11).

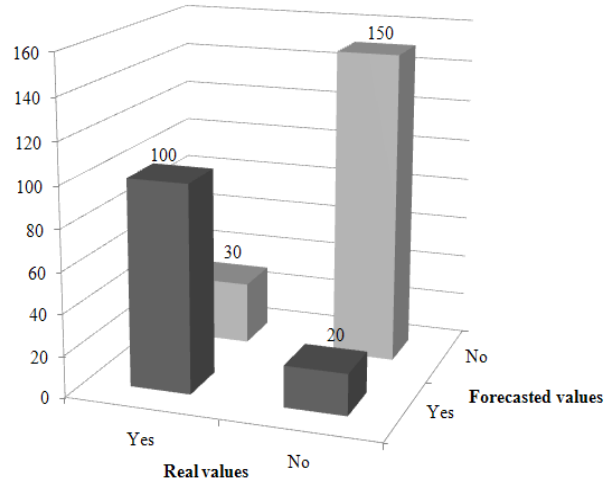


Figure 11. Even classification graph
(source: self study)

Figure 11 shows, for 130 real cases with the YES value, that the model has properly classified 100 cases (real value = value prognosis) and wrongly classified 30 cases (real value ≠ value prognosis). In 170 real cases with the NO value, the model has properly classified 150 cases and wrongly classified 20 cases. It is necessary for the model to classify properly the majority of the cases before its implementation.

5.8 Transformation of results into knowledge

Accepted results of the final model need to be transformed into particular knowledge that concerns researched event. Estimator values (e.g. the example from Table 5) should be changed into the profile description of the clients who break and maintain the insurance policy:

- customer with the highest probability of maintaining the policy is a female from małopolskie district,
- customer with the highest probability of breaking the policy is a person with vocational education.

However the interpretation of model results and transferring them into practical knowledge is not always that simple. For example, if the model result will be determined by one feature e.g. number of paid insurance fees since the start of the insurance agreement. In this case it is not enough to claim that if the customer will pay 3 fees, the probability of maintaining the policy till its end is 99%. There needs to be an explanation provided for the fact that most of the insurance policies are broken in first months. Usually in such cases the experience and knowledge of people who know, better than the research team, the factors influencing the researched event.

5.9 Knowledge usefulness evaluation

This is the final stage of knowledge discovering, although it does not have to be the end of the process. At this stage user evaluates the usefulness of the gathered data for the realization of determined business aims. If the goals were not reached, it usually means returning to one of the previous stages or even restarting of the whole process. It is crucial to emphasize that such problem emerges relatively often (20% - 30% of all performed analysis) and it does not have to mean a negative impact on the process. Complexity and diversity of both the analyzed data and the analytical tasks usually leads to a deviation between user expectations and results reached in knowledge discovering process. It does not indicate that the process failed. Such event gives a signal that thee are corrections which need to be implemented and that the process has to be restarted. It is important to point out that the most time consuming and difficult are the stages which lead to the elaboration of the mining extract. If this stage was discussed with the users and turned out to be success, repeating of following stages, even multiple, will never be as costly and complicated.

Accepted, implemented and effectively used data mining model should be periodically updated in order to assure that the data is correct and up-to-date, which is dependent on parameters that change in time (e.g. seasonality, economical situation of a market sector, new trends in activities of the competition).

6 Summary

Data warehouse and analytical data processing tools are the basic elements of the system aiding the management decision making processes. Traditional warehouses are directed at processing of text and numerical data with statistical and visualization methods, based on current analytical processing technology. At the same time, many modern solutions, dealing with spatial and timely data analytical processing, data with complex structure (semi-structural, objective and multimedia) and the analysis of constant data, are developing.

There is a number of issues, mostly scientific and technical, to be solved for this type of data.

One of the most significant types of analytical tools, based on data warehouse resources, are the applications that allow data mining. There are many software packages in the market, which offer mining of text files or analytical databases. Due to the fact that the process is extremely complex, time consuming and simultaneously the quality of taken decisions is dependent on its efficiency, the proper methodology of the process is a crucial success factor.

The article presented issues connected with analytical data processing that was divided into two groups: current analytical processing and data mining. Methodology of knowledge discovering process, which included the use of data warehouses and both types of analystical data processing tools, was proposed.

7 References

[1] Berry M.J., Linoff G.S. - *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management*. Wiley Publishing Inc., New Jersey, 2004.

[2] Codd E.F., Codd S.B., Salley C.T - *Providing OLAP to User-Analysts: An IT Mandate. Hyperion*, California, Manchester, Singapore 1993.

[3] Golański A. - *Yahoo! ma rekordową, petabajtową bazę danych*, Webhosting.pl, 18-06-2010.

[4] Gustaffson T. - *Sybase i Sun - 1 Petabajt (PB) danych* [in] Komputer w firmie. Czasopismo Nowych Technologii, 28-05-2008.

[5] Hackathorn R. - *Real-Time to Real-Value* [in] DM Review Magazine. DM Review, January 2004.

[6] Hand D.J., Mannila H., Smyth P. - *Principles of Data Mining (Adaptive Computation and Machine Learning)*. The MIT Press, Cambridge 2001.

[7] Inmon W.H. - *Building the Data Warehouse.* John Wiley and Sons, New York 1992.

[8] Kelly S. - *Data Warehousing - the route to mass customization.* John Wiley and Sons, New York 1996.

[9] Kimball R. - *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses.* John Wiley and Sons, New York 1996.

[10] Kowalska J., Trawiński B. - *Zastosowanie metod eksploracji danych do badania sprzedaży w przedsiębiorstwie produkcyjnym* [in] Bazy danych. Struktury, algorytmy, metody. Tom 2: Wybrane technologie i zastosowania. Wydawnictwa Komunikacji i Łączności, Warszawa 2006, pp. 241-250.

[11] Koziński M. - *Systemy wspomagania decyzji – Zamiast kryształowej kuli* [in] PCKurier, No. 5, 2003.

[12] Lai E. - *Google claims MapReduce sets data-sorting record, topping Yahoo, conventional databases* [in] Computerworld News, 24-11-2008.

[13] Marcinek T. - *Produkcja pod lupą* [in] Computer-World Polska, April 2001.

[14] Morzy T. - *Eksploracja danych: problemy i rozwiązania* [at] V Konferencja PLOUG, Zakopane 1999.

[15] Parades J. - *The Multidimensional Data Modeling Toolkit*. Olap World Press, New York 2009.

[16] Piatetsky-Shapiro G., Frawley W.J. - *Knowledge Discovery in Databases* [in] AAAI Press/MIT Press, Menlo Park 1991.

[17] Rostek K. - *Informatyczne systemy i rozwiązania wspierające podejmowanie decyzji zarządczych* [in] Przyszłość systemów informatycznych w zarządzaniu przedsiębiorstwem (ed. K. Sitarski), PTZP, Warszawa 2008, pp. 140-149.

[18] Sobieszczyk T. - *Analiza rentowności w branży ubezpieczeniowej z wykorzystaniem narzędzi Business Intelligence* [in] Controlling i rachunkowość zarządcza, No. 8, 2004.

[19] Sokołowski A. - *Wprowadzenie do zastosowań metod statystycznych i technik data mining w badaniach naukowych*. StatSoft, 2002.

[20] Stokalski B. - *Hurtownie danych – szanse i wyzwania* [in] Computerworld Polska, November 1997.

[21] Tang Z., MacLennan J. - *Data Mining with SQL Server 2005*. Wiley Publishing Inc., Indianapolis 2005.

[22] Weichbroth P. - *Algorytmy eksploracji danych z baz danych*. Microsoft TechNet, April 2008.

[23] http://rapidshare.com/wiruberuns.html, 18-06-2010.

[24] http://www.statemaster.com/encyclopedia/ Petabyte, 18-06-2010.