

NEW CONDITIONS FOR APPLICATION OF DATA MINING METHODS IN THE SECTOR OF MID-SIZED ENTERPRISES

Business activity carried out on a wide and wide scale results in arising of huge data sets. The functional relationships between data are very often unclear or in the best case very uncertain. Verification of theoretical models became possible only since the last ten years mainly because of the intense computerization progress. Development of methods of data collecting, processing and analysis (for example Data Mining, OLAP) has increased the possibilities of retrieving of knowledge on the one hand, but on the other hand the problems of acceptance of analysis outcomes has occurred.

NOWE WARUNKI ZASTOSOWANIA METOD DATA MINING SEKTORZE ŚREDNICH PRZEDSIĘBIORSTW

Prowadzenie na coraz szerszą skalę działalności gospodarczej przedsiębiorstw implikuje powstawanie coraz większych zbiorów danych. Funkcjonalne związki pomiędzy nimi są często nierozpoznane lub w najlepszym przypadku niepewne. Adaptując do opisu rzeczywistości gospodarczej, wypracowane w teorii modele ekonomiczne praktycy życia gospodarczego stanęli obecnie przed szansą ich praktycznej weryfikacji. Postępujące procesy informatyzacji spowodowały rozwój elektronicznych metod gromadzenia i przetwarzania dużych zbiorów danych oraz doskonalenie metod ich eksploracji. Do takich metod zaliczane są: metody drążenia danych (ang. Data Mining), wizualizacji (OLAP), które z jednej strony dostarczyły instrumentarium, służące odkrywaniu ukrytej w zbiorach danych wiedzy. Z drugiej zaś - zrodziły problemy akceptacji wyników analiz uzyskiwanych za pomocą skomplikowanych metod teoretycznych.

1. CHANCES AND RISKS OF DATA MINING

Dissemination of integrated informatics solutions for example ERP (Enterprise Resource Planning) in the sector of small and middle sized enterprises gives rise to a lot of discussion about the possibilities of taking advantage of huge data archives which arise in the databases of such systems. Enormous advantages seem to result from the work with an integrated solution on the one hand, on the second hand information requirements of top

and middle management as regards possibilities of preparation of managerial information rise as quickly as the data sets.

First experiences in introducing proALPHA[®] ERP¹ solution on polish market show the first level of data quality referring to three level quality approach has been in the most cases accomplished². Basic reports and statistics (balance sheet, customer's and article's statistics) offered nowadays as "a standard equipment" with most of the ERP systems are able to meet most of basic internal and external information needs. After the system is introduced and the first quality's level– so called "required quality"³ was achieved, the customers of information start to require more advanced analytical tools, which would enable them to achieve the second and the third quality level (satisfactory and delightful quality). The achievement of this goal becomes a starting point in application of more and more advanced method of data explorations, for example Data Mining

In the next part of article some aspects of data quality assurance on the basis of proALPHA[®] Analyzer are discussed⁴.

1.1. Some aspects of data quality assurance

Data quality assurance on satisfactory level expects flexibility and reliability as two basis features, which should characterize the process for preparation of managerial information. When many on situation dependent information needs could be met with a number of quite different ways (charts, tables) and as quickly as possible, then the flexibility and reliability is done. w. In the most cases such needs relate to traditional analysis of performance measurements, for example.:

- turnover and margin of some group of customers,
- payment history of customers,
- guaranteeing.

The following is an example of an analysis that could be prepared as above mentioned. In order to create an ad hoc analysis of influence of seasons on the sales level the data collected in proALPHA[®] system⁵ were exported to proALPHA[®] Analyzer.

An example

There is demanding a measurement of influence of seasons on sales level

Solution:

There are in proALPHA[®] Analyzer only four step, that must be taken to retrieve the information:

1. Selection of function "Analyse" from toolbar in order to create multidimensional

¹ proALPHA[®] – a representative of ERP class systems has emerged polish market since 2004.

² The conception consists of three quality levels: quality required, satisfactory and delightful.

Detail in [3, s. 41 – 74] presented

³ ibid.

⁴ proALPHA[®] Analyzer is an OLAP based application that enable to display and analyze all kinds of data. Analyzed data can be displayed in the form of diagrams and reports. An application is offer together with proALPHA system.

⁵ Real data come from a company, which has been used proALPHA system since couple of years

data cube⁶ with attribute "Amount" as aggregate and attributes "Document type", "Article", and "Month" as dimensions (figure 1)

2. Inserting of new analysis group into infospace and preparing the report chart components in order to achieve a better visualization effect (figure 2).

Innovation aspect within the confines of the third quality level underlines the possibilities of application of modern analytical tools (visualization tools, predefined reports created in the technology „button on"). Combination of these methods and tools could contribute to achieving the new interesting analysis results provided that, they are acceptable and understandable for the front users.

Unfortunately data quality assurance on the satisfactory and delightful level encounters the limitations. One of them is difficulty in assurance of data quality on the required level. Registration of thousand and thousand master and transactional data in the ERP system is the cause of many data mistakes. Some of typically data failures could be:

- wrong or incomplete master data (one of the most common reason of problems with data quality in an enterprise),
- mistakes committed during registration of transactional data,
- correcting operations (for example delete).

Assurance of required quality level is also of great importance as regards others areas of applications of modern computer aided methods, for example Advanced Planning and Scheduling (APS). Using of those methods and techniques can help one to reach important rationalization effects as regards the possibilities of processing of existing or quite new information. But a necessary condition for obtaining such effects is that the rationalization effects with regard to the first basic quality level have to be assured.

Application of Data Mining in a sector of small and mid-sized enterprises often comes across an important obstacle to get through. This is an attitude of management (or just „boss") toward collecting and processing huge data sets (acceptance problem). Internal controlling systems often take the simplified form and the data archives are diffused, what makes it difficult to explore them.

2. NEW POSSIBILITIES OF APPLICATION OF MATHEMATICAL MODELS IN THE MANAGEMENT OF SMALL AND MID-SIZED ENTERPRISES

From the application point of view have always Operations Research dealt with the two complex subject:

- poor quality and missing data;
- acceptance problems (as a result of difficulties with economical interpretation of logic of many algorithms. Impossibilities of assessment of benefits before applications of models is also the point⁷).

⁶ proALPHA[®] Analyzer supports OLAP function.

⁷ In addition to efforts of data entry and delivery of quality there is often a need to buy special software or special modules.

The wide possibilities of delivery of dates were in the first chapter introduced. The second one tries to focus on idea how to overcome (partially at least) the acceptance problems relating to condition in first chapter presented.

Actually the commerce praxis have no reservation about application of mathematic models on the one hand, on the other hand these models have been used only in a simplified manner. For example :

- performance indicators, which used to rely upon four basic calculations
- separation of fix and variable costs with appliance of (mostly linear) regression analysis.

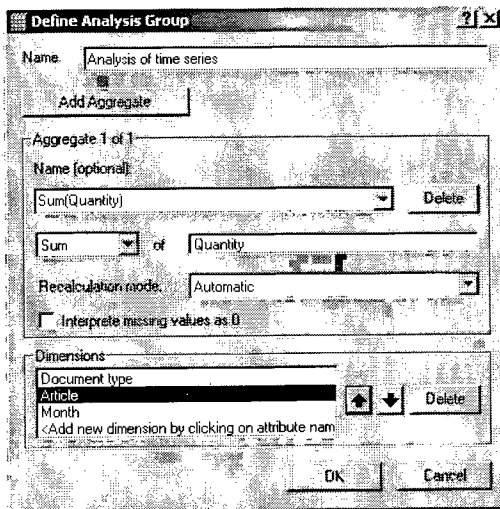


Figure 1. Creating of multidimensional cube

This is the reason why the application of these modeling structures for integrated planning (mostly required by the credit institutions) can come up against some acceptance problems. Within an interesting (from point of view of OR models) logistic area it applies in particular to⁸:

- logistics costs planning (for example in transport), for the moment by means of linear functions together with the simplified planning of possible cost items on the basis of process costing approach (for example a number of purchase orders or a number of customer orders);
- planning of inventories and inventory costs on the basis of more acceptable⁹ optimizations within an purchasing area ;

⁹ It relates to classical order quantity models and to the rules of dimensioning of minimum inventory levels as well.

- taking into consideration of quantitative results of „throughput planning dilemma“ by means of imprecise assessments of throughput times of operations.

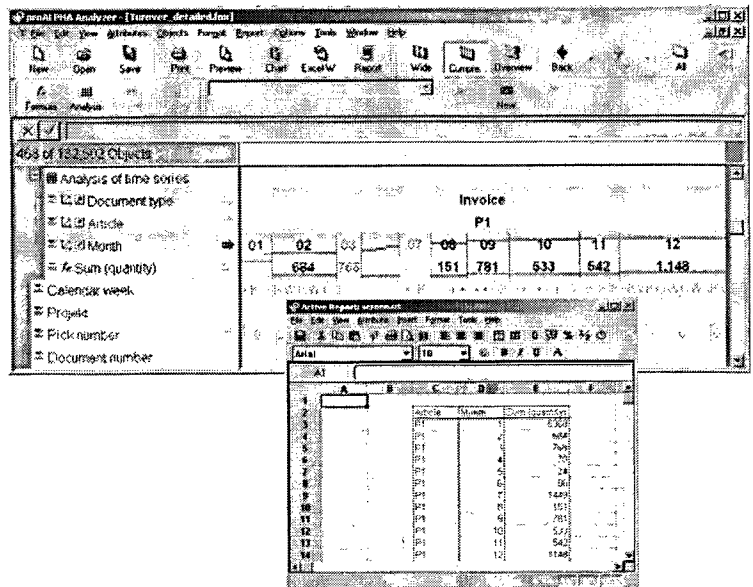


Figure 2 The possibilities of time series visualization.

Performance indicators are the results of such integrated planning and the analysis of them (plan/real deviation analysis at least) is often required by the credit institutions. The reasons of deviations could be explained much better by iterative application of real values of planned parameters into the planning model. Collection and preparation of real data is taking into account the usability of ERP systems and OLAP analysis tools only the organizational problem, compare chapter 1.

The analysis of plan/real deviation is crucial for an approach of increasing acceptance. From this point of view noticeable are:

- Aggregation of effects of influence of many parameters on deviation

Theoretically this is possible to determine the effect of deviation for each parameter separately. But it could also lead to unclarity because of variety of parameters. Areas of activities (for example cost and profit center) and the areas of overall problems (logistic or liquidity) seems to be the most reasonable criteria for parameter aggregation. Demo planning program consists of five hypothetically fields of activities. As follows:

- „Sales“ (S) with all parameters as regards sales prices and sales quantities (according to groups of products and delivery dates) as well as client (according to terms of delivery);
- „Purchasing“ (Z) with all parameters as regards purchase prices and purchase quantities (according to groups of products and terms of delivery);

- „Personnel“ (P) with all parameters as regards salary and wage payment (includes the components dependent in logistic);
 - „Technology“ (T) with parameters as regards lot sizes and standstill periods;
 - „Overhead costs“ (KO) with fix and variable cost rates.
- b) Sequence of determination of deviation effects

For theoretical purposes have the sequence significant influence on quantitative characteristic of deviation effects. This attitude may result in case of many parameter groups in increase of preparatory efforts (see the demo example). In practice there is no need to analyze the smallest deviations because of expected information benefits („ABC-way of thinking“) and quality of data. Factual reasons for determination of sequence will therefore dominate. As regards logistics problems area the above mentioned sequence is chosen. The reasons are:

- Marketing establish always basis conditions for basis operations of enterprise. This is the reason that marketing has to be in the fore.
- „Material“ and „stuff“ are the most important production factors.
- Combination of production factors is determined by (organization) and technology.
- All above mentioned areas have the effect on „overhead costs“.

The results of evaluation deviation process are on the figure 3 presented. Given that the parameters are grouped this is reasonable to analyze firstly the groups of effects. But it is recommended to revoke the grouping and carry out an additional analyze, if the group effects become too large. Interpretation of grouping effects on the basis of presented example is:

- the biggest effects result from parameters of „sales“ area (with unfavorably influence on RoI). This is the reason that sales area must be analyzed in detail. A more detailed examination shows that the decrease in planned turnover is mainly attributed to cut in prices and some „freebie“ activities in logistic area. As a consequence the sales volume has increase but not enough to balance the additional costs. Thus, the Data Mining analysis of influence of marketing mix actions on profitability could be helpful and reasonable as soon as could lead to improvement of control of marketing area.
- the great importance of quantity deviation: Approximately 50 % of them are not explainable, even if they are examined separately (taking into account the applied models and the real data)

	1	2	3	4	5	6	7	8	
		S - Plan	Z - S	P - Z	T - P	KO - T	Q - ty deviation	Sum	
47	Tools	29,43	0,00	0,00	0,00	-57,93	0,00	-28,45	
48	Travel costs	0,00	0,00	79,30	0,00	-183,92	0,00	-104,62	
49	Post costs	28,31	-0,10	5,02	0,00	12,62	-0,13	45,72	
51	Transport costs (sales)	453,24	0,00	0,00	0,00	-486,65	0,00	-33,40	
52	Transport costs (purchasing)	2,48	0,77	0,00	0,00	-15,72	-4,91	-17,38	
53	External service	5,74	0,00	0,00	0,00	14,13	0,00	19,87	
54	Maintenance	19,14	0,00	0,00	0,00	53,98	0,00	73,12	
55	Other	5,15	-0,04	6,02	0,00	5,37	-0,07	16,44	
56	Interest on loan(s)	0,00	0,00	0,00	0,00	0,00	29,84	29,84	
57	EBIT	-592,02	90,27	-251,66	-3,67	663,50	-106,97	-200,75	
58	Inventory	574,33	0,75	-7,79	-0,56	-50,59	67,30	603,43	
59	Amount due	-13,71	0,00	0,00	0,00	0,00	-639,07	-652,78	
60	Other tangible assets	177,58	-11,22	22,16	0,25	-71,52	-254,50	-137,24	
61	Current liabilities	-323,30	21,29	-42,06	-0,47	135,73	1065,50	856,68	
62	Working capital	414,90	10,82	-27,59	-0,79	13,62	259,23	670,09	
64	Rol		-6,74%	0,96%	-2,68%	-0,04%	7,09%	-1,40%	2,82%

Figure 3 Determination of individual deviations.

This was the reason that the quantity deviation were separately analyzed. In doing so it was also controlled why did quantity deviations only in some positions appear: The possible interpretations are:

- Deviations in „sales revenue“ are (except for inventory changes) completely with “sales” parameter explainable. This is a result of planning of revenues on the basis of relation: sales quantity * netto-sales price and determination of real sales price with quotient: revenue/sales quantity. However the quantity deviations are evidence of data mistake.

Following the Data Mining demand was determined. The results of quantity deviation analysis bring some new explanations of them:

- deviations for model output quantities; amount dues, other tangible assets and liabilities are the consequence of simplified planning of liquidity. Accordingly there is no need to carry out the Data Mining activities. Advanced modeling is sufficient on this place.
- however, all the other quantity deviations are as results of faulty of data or logistics models interpretable. Es regards particularly the cost types, which are connected (directly or indirectly) with technology and logistic. If it was possible to exclude the data mistakes (with utilization of proALPHA® Analyzer) then the real Data Mining and modeling demand would be achieved.

On the basis of this approach, there are in ZCIE the following activities with utilization of ERP system and modern analytical tools realized:

- determination of Data Mining demand, that is on enterprise and situation dependent;
- control of possibilities of application and acceptance of Data Mining tools;
- gradually involvement of farther sophisticated models of planning and execution of business processes¹⁰.

¹⁰ E.g. the APS optimization that relate to logistics and production controls

All these activities bring about the development of new conception of data quality management in small and medium sized enterprises. The basis principle are in [3] discussed.

Literature

- [1] Ciechan Kujawa M.: *Rachunek kosztów jakości*, Oficyna Ekonomiczna, Kraków 2005.
- [2] Kluge P.D., Kuźdowicz P.: *Zarządzanie jakością danych przy zastosowaniu rozwiązania klasy ERP w małym i średnim przedsiębiorstwie produkcyjnym*, Warszawa 2004.
- [3] Kluge P.D.: *Zapewnienie jakości danych jako aktualne zadanie kadry kierowniczej*, Wydaw. Akademii Podlaskiej w Siedlcach, Siedlce 2005, s. 41 – 74.
- [4] Kluge P.D., Kuźdowicz D., Kuźdowicz P., Orzeszko P, Ostrowska-Huba A.: *Technologia komputerowo wspomaganego controllingu w MSP w warunkach nowej bazylejskiej umowy kapitałowej – Basel II, Uwarunkowania sukcesu przedsiębiorstwa w gospodarce opartej na wiedzy - SUCCESS 2004 : materiały z konferencji naukowej*. Kazimierz Dolny, Polska, 2004 .- Lublin : Uniwersytet Marii Curie Skłodowskiej, 2004 .