INFORMATYKA EKONOMICZNA BUSINESS INFORMATICS 2(28) · 2013

ISSN 1507-3858

Marcin Relich
University of Zielona Gora
e-mail: m.relich@wez.uz.zgora.pl

KNOWLEDGE DISCOVERY FROM AN ERP DATABASE IN THE CONTEXT OF NEW PRODUCT DEVELOPMENT

Abstract: This paper is aimed at using an ERP database to identify the variables that have a significant influence on the duration of a project phase. In the paper, some methodologies of the knowledge discovery process are compared and a model of knowledge discovery from an ERP database is proposed. The presented approach is dedicated for the industrial enterprises that use an ERP system to plan and control the development of new products. The example contains four stages of the knowledge discovery process, such as data selection, data transformation, data mining, and the interpretation of patterns. Among data mining techniques, a fuzzy neural system is chosen to seek relationships between data from completed projects and other data stored in an ERP system.

Keywords: knowledge management, project management, data selection, data mining.

1. Introduction

In recent years, the advancement of information technology in business management processes has placed Enterprise Resource Planning (ERP) system as one of the most widely implemented business software in various enterprises. ERP is a system for the seamless integration of all the information flowing through the company such as finance, accounting, human resources, supply chain, and customer information [Davenport 1998]. The goal of an ERP based integrated information system is to make the system effective, efficient and user friendly. The primary task of an integrated system is to maintain the data flow of an organization and to reduce redundancy [Imtiaz, Kibria 2012; May et al. 2013].

The present information and communication technologies have become one of the most important factors, conditions and chances of company development. These technologies enable the collection, presentation, transfer, access and the use of enormous amount of data. The data are a potential source of information that, combined with managerial skills and experience, may influence the choice of the correct decision. ERP systems help to collect, operate, and store data concerning the

daily activities of an enterprise (e.g. client orders), as well as the results of previous projects (development of products).

One of the functionalities of an ERP system concerns project management that a company can use to develop new products. To obtain a project schedule, there is required data specification concerning resources and activities, including their sequence and duration. Project parameters can be specified by the experts or estimated with the use of an ERP database. The former approach is suitable for projects that have a very unique form, e.g. construction projects. In turn, if an enterprise develops new products and a new version of product is connected with the superficial modification of a product specification, then it is possible to acquire knowledge from the ERP database and to use it for the improvement of the estimation quality of project parameters.

The goal of this paper is to present the possibility of the use of the ERP database for seeking the relationships between the ERP attributes (e.g. delay of material delivery by suppliers, number of subcontractors, team members) and the project parameters (e.g. project duration and cost). The sought-after relationships can support the user in the assessment of project parameters, and as a consequence to obtain more relevant estimates. It is unrealistic to expect very accurate estimates of project effort because of the inherent uncertainty in development projects, and the complex and dynamic interaction of factors that influence its development. However, even a small improvement in the estimation quality can influence positively planning and monitoring the project, for instance, in project cost, resource allocation, and schedule arrangement.

2. Model of a knowledge discovery in databases

The knowledge discovery in databases (KDD) is concerned with the development of methods and techniques for making sense of data. KDD is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The basic problem addressed by the KDD process is one of mapping low-level data (which are typically too voluminous to understand and digest easily) into other forms that might be more compact (e.g. a short report), more abstract (e.g. a descriptive approximation or model of the process that generated the data), or more useful (e.g. a predictive model for estimating the value of future cases). At the core of the process is the application of specific data-mining methods for pattern discovery and extraction [Fayyad et al. 1996].

Table 1 compares the steps according to the most used methodologies for developing data mining and knowledge discovery (DM & KD) projects. CRISP-DM (CRoss-Industry Standard Process for Data Mining) states which tasks have to be carried out to complete a data mining project [Cabena et al. 1998]. Cios et al. adapted the CRISP-DM model to the needs of the academic research community, providing a more general, research-oriented description of the steps model [Marban et al. 2009].

Table 1. Comparison of DM & KD process models and methodologies

Model	Fayyad et al.	Cabena et al.	Anand & Buchner	CRISP-DM	Cios et al.
Steps	Developing and Understanding of the Application Domain	Business Objectives Determination	Human Resource Identification	Business Understanding	Understanding the Data
			Problem Specification		
	Creating a Target Data Set	Data Preparation	Data Prospecting	Data Understanding	Understanding the Data
	Data Cleaning and Pre-processing		Domain Knowledge Elicitation		
	Data Reduction and Projection		Methodology Identification	Data Preparation	Preparation of the data
	Choosing the DM Task		Data Preprocessing		
	Choosing the DM Algorithm				
	DM	DM	Pattern Discovery	Modeling	DM
	Interpreting Mined Patterns	Domain Knowledge Elicitation	Knowledge Postprocessing	Evaluation	Evaluation of the Discovered Knowledge
	Consolidating Discovered Knowledge	Assimilation of Knowledge		Deployment	Using the Discovered Knowledge

Source: [Marban et al. 2009].

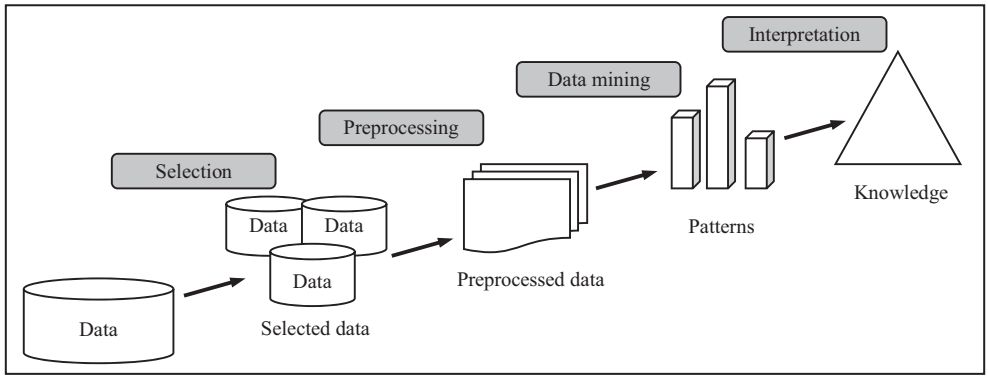

Figure 1. Model of knowledge discovery

Source: self study base on [Fayyad et al. 1996].

The steps of the KDD process in the above-presented models can generally be grouped into four main tasks [*Generowanie wiedzy…* 2004]:
– data selection,
– data transformation,

– data mining,
– interpretation/evaluation of patterns.

Figure 1 presents the model of KDD process that is further considered in the context of knowledge acquisition from an ERP database.

The knowledge acquisition from an ERP system requires using the KDD methods that are suitable for the characteristics of an ERP database. The proposed method for seeking relationships in the order of project parameter estimation is presented in the next section.

3. Method for project parameter estimation in the context of ERP systems

The presented method is dedicated for industrial enterprises that also use an ERP system to support the development of new products. The phases of product development depend on the characteristics of the product and company in which it is designed. However, some common phases can be distinguished, for example [*Advanced product…* 1994]:

1) plan and define program,
2) product design and development verification,
3) process design and development verification,
4) product and process validation,
5) production.

These phases can also be considered in the context of concept initiation, program approval, prototype, pilot, and launch. Each phase requires the specification of duration and cost. In each of these phases, the critical factors (parameters of an ERP database) that significantly influence new product development are sought. The estimation of these parameters is especially desired in the medium and large enterprises that develop a few new products simultaneously. In the case of a significant variance of a project parameter, the use of the average or time series models can result in inaccurate estimates. Thus, a search for conditional rules using an ERP database is proposed. The sought-after relationships can improve the quality of estimates that are an input into an ERP system, in a project management module.

Among the KDD steps, two steps seem to be especially important in the context of knowledge acquisition from an ERP database, i.e. data selection and data mining. A large number of independent variables in a large data set can present two major problems. Firstly, too many variables result in long training times when the model is built. Secondly, a large number of observations and variables tend to retain redundant information through multicollinearity leading to unreliable models. Some of the variables present in historical data are needed for some problems and some variables for others. Often, different variables may carry the same information [Colmenares, Perez 1998].

An ERP database contains hundreds of attributes that can be irrelevant to the mining task or even redundant. To reduce the data set size, an attribute (feature) subset selection can be used that removes irrelevant or redundant attributes (or dimensions). The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. Mining on a reduced set of attributes has an additional benefit. It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand [Han, Kamber 2006]. One of the variable reduction methods is principal component analysis, that reduces the dimension for linearly mapping high dimensional data onto a lower dimension with minimal loss of information.

Knowledge acquisition requires some data mining techniques that cope with the description of relationships among data and that solve the problems connected with e.g. classification, regression, and clustering. These techniques include neural networks, fuzzy sets, rough sets, time series analysis, Bayesian networks, decision trees, evolutionary programming and genetic algorithms, Markov modeling, etc. Data mining should be connected with matching the goals of the KDD process from the user's viewpoint to particular methods, for example, summarization, classification, regression, clustering. The blind application of data mining methods can lead to the discovery of meaningless and invalid patterns [Fayyad et al. 1996].

The database of an ERP system comprises an enormous number of parameters that can be considered as potential variables to identify the project parameters. One of the data mining techniques is the fuzzy neural system that can take into account the imprecise character of data, cope with enormous amount of data, and identify the relationships among data. Fuzzy logic and artificial neural networks are complementary technologies and powerful design techniques that can be used in the identification of patterns from among a large database and noisy data.

The fuzzy neural system has the advantages of both neural networks (e.g. learning abilities, optimization abilities and connectionist structures) and fuzzy systems (simplicity of incorporating expert knowledge). As a result, it is possible to bring the low-level learning and computational power of neural networks into fuzzy systems and also high-level human like IF-THEN thinking and reasoning of fuzzy systems into neural networks. The fuzzy neural method is rather a way to create a fuzzy model from data by some kind of learning method that is motivated by learning procedures used in neural networks. This substantially reduces development time and cost while improving the accuracy of the resulting fuzzy model. Being able to utilize a neural learning algorithm implies that a fuzzy system with linguistic information in its rule base can be updated or adapted using numerical information to gain an even greater advantage over a neural network that cannot make use of linguistic information and behaves as a black box [Azar 2010].

The behaviour of a fuzzy neural system can be represented by a set of humanly understandable rules or by a combination of localized basis functions associated

with local models, making them an ideal framework to perform nonlinear predictive modelling. Nevertheless, one important consequence of this hybridization between the representational aspect of fuzzy models and the learning mechanism of neural networks is the contrast between the readability and performance of the resulting model [Azar 2010]. The combination of fuzzy systems and neural networks has recently become a popular approach in engineering fields for solving problems in control, identification, prediction, pattern recognition, etc. [Cheng et al. 2010; Chien et al. 2010; Zeng 2007]. One well-known structure is the adaptive neuro-fuzzy inference system (ANFIS) that is a universal approximator and enables non-linear modelling and forecasting [Relich 2010].

4. Illustrative example

The following example refers to four steps of knowledge discovery from an ERP database in the context of new product development. The output variable is the duration (in days) of the *j*-th phase in project *i*. In turn, the input variables of the *j*-th phase in project *i* are presented in Table 2. These variables are derived from the ERP system modules that are connected with new product development (project management). The development of a product prototype requires the purchase of materials from the suppliers, storage of materials, and usage of materials in production.

Table 2. Input variables for product prototype phase

Purchasing	Materials Management	Production	Project Management
Value of material purchase	Number of materials in warehouses	Productive capacity (actual/maximal)	Number of standard tasks in the project phase
Number of suppliers selling required materials	Number of supplier withdrawal notices	Number of resource overloads	Number of unique tasks in the project phase
Number of delivery reminder documents	Number of warehouse transfers	Time of machine inspection	Number of changes in the project phase specification
Delivery duration		Number of machines	Number of subcontractors
Delay of delivery		Number of work orders	Number of team members
Changes of price list			Number of materials needed in the project phase

Source: self study.

The first step of the knowledge discovery process concerns data selection and it can be divided into two approaches: the expert chooses the data according to his/her experience and the use of one of the variable reduction methods. The input variables presented in Table 2 have been chosen according to the expert approach. All these

variables have a numerical format, but different units, for example, purchasing is in monetary unit, delivery duration in days, and productive capacity in percent. Therefore, the data requires transformation before one of the variable reduction methods is used. The principal component analysis was chosen as the variable reduction method. This analysis transforms the input data so that the elements of the input vectors will be uncorrelated. In addition, the size of the input vectors may be reduced by retaining only those components that contribute more than a specified fraction of the total variation in the data set. After the use of principal component analysis, the data set was reduced from 20 input variables to 6 (Delivery duration − DD, Delay of delivery − DoD, Number of work orders − NoWO, Number of standard tasks in the project phase – NoST, Number of unique tasks in the project phase – NoUT, Number of team members − NoTM). Moreover, the data set was normalized so that it has a zero mean. Data preprocessing (transformation) is the second step of the knowledge discovery process and helps a fuzzy neural system obtain more accurate results.

The third step of the knowledge discovery process regards the use of data mining techniques/tools. In order to seek relationships, the adaptive neuro-fuzzy inference system (ANFIS), which is the tool of Matlab® software, was applied. Figure 2 presents the structure of ANFIS for the duration of the project prototype phase.

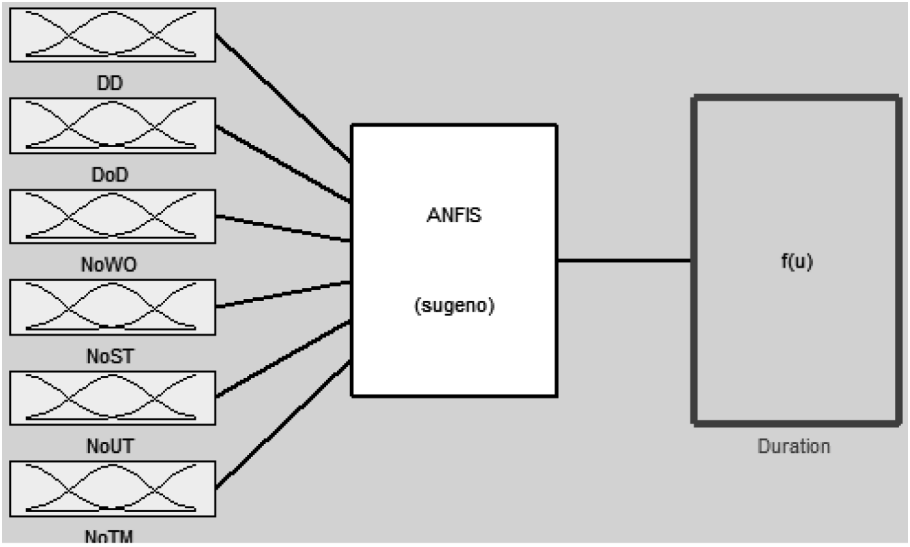

Figure 2. Structure of adaptive neuro-fuzzy inference system

Source: self study.

In order to eliminate the overtraining of ANFIS (too strict function adjustment to data) and to increase the estimation quality, the data set is divided into learning

(P1-P8) and testing set (P9-P11). The learning phase requires the declaration of a membership function type of fuzzy sets (e.g. triangular, Gaussian function), defuzzification method, method of weights optimisation, and stop criterion (e.g. error value or the number of iteration). The ANFIS tool proposes two methods concerning the identification of the shape of membership functions, grid partition and the subtractive clustering method. According to the shape of membership functions, the rules are generated. An example of fuzzy rules for the duration is presented in Figure 3.

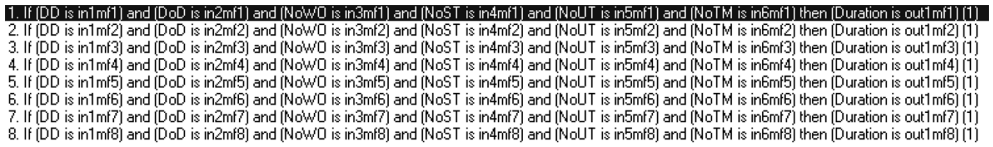

Figure 3. Fuzzy rules for duration assessment

Source: self study.

After the learning phase, the testing data are led to input of system to compare the error between different models. Root mean square errors (RMSE) for various models are presented in Table 3. The least error in testing set for the duration of the project prototype phase has been generated with the use of ANFIS with the subtractive clustering method. It is noteworthy that the error generated with the use of average is smaller than the error of the linear model. A comparison of different models is especially recommended in the case of low level of variance for a dependant variable (in the considered case for the duration of a project phase).

Table 3. Comparison of RMSE for different models

Model	RMSE
Average	24.89
Linear model	53.02
ANFIS – grid partition	10.83
ANFIS – subtractive clustering	5.85

Source: self study.

The membership functions and rules are a basis to evaluate the duration of an actual project. Let us assume that for the actual project we considered the following values: average delivery duration – 24.5 days, average delay of delivery – 10.5 days, number of work orders – 55, number of standard tasks in the project phase – 10, number of unique tasks in the project phase – 3, and number of team members – 5.

Figure 4 presents the membership functions for 8 rules that determine the planned duration of the project phase at 90 days.

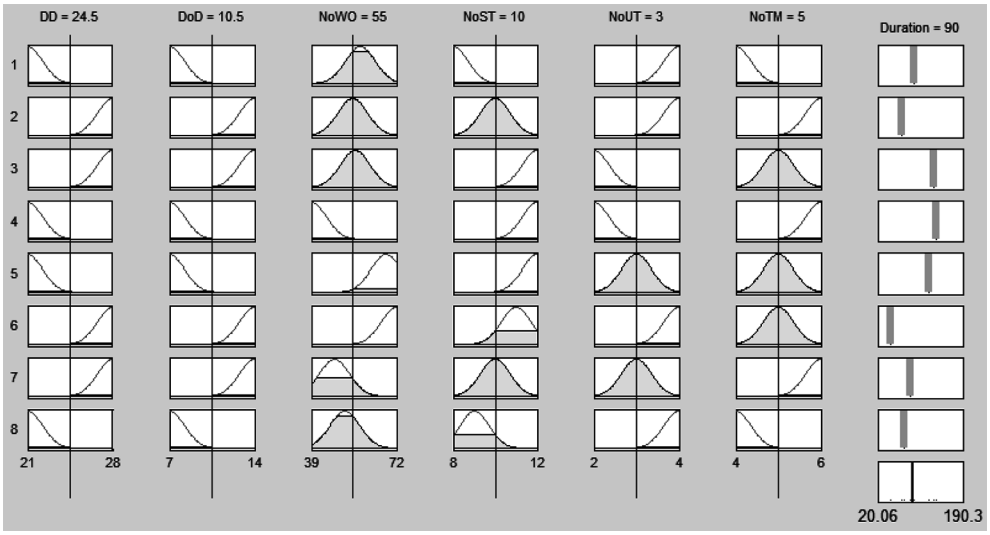

Figure 4. Estimation of project duration

Source: self study.

Table 4. Project duration for two criteria (in days)

Number of work orders	Number of team members	
	5	6
50	79.3	78.7
51	80.7	79.8
52	82.4	80.9
53	84.3	81.9
54	86.8	82.9
55	90	83.9
56	94.2	84.9
57	99.4	85.8
58	105	86.7
59	111	87.6
60	116	88.4

Source: self study.

The presented analysis can be broadened into multidimensional analysis to support the decision-maker in determining optimal values of some parameters. For instance, the decision-maker would like to know the number of team members (from 5 to 6) and the number of work orders (from 50 to 60), for which the planned duration of project phase is the least. Table 4 presents the results for the above-mentioned constraints.

The results indicate that the shortest duration is for 6 members of the project team and 50 the number of work orders. Moreover, the increase of project duration with 5 team members is proportionally larger than with 6 members. For instance, the reduction of work orders from 60 to 55 for 5 team members decreases project duration by 26 days (116 days – 90 days), whereas for 6 team members, the decline equals 4.5 days (88.4 days – 83.9 days). In the case of extensive search space, the time of the obtained solution can be significantly reduced, e.g. using constraints programming languages [Bagiński, Rostański 2011; Relich 2011] The above-presented analysis is conducted for each phase of project and the obtained estimates can be used for further evaluation, e.g. in the planning of cash flow, working capital, and financial reserves.

5. Conclusions

Intense competition on the market forces enterprises to develop simultaneously a few new products. Moreover, the enterprise usually acts in the framework of constraints concerning time and resources (financial, human, logistic), and it has to choose an optimal set of new products. This set can be determined with the use of time-cost analysis of new products that requires the estimates of project parameters. Inappropriate choice of a set of new products can lead to decreasing market share, profitability and liquidity, and as a consequence to the bankruptcy of the enterprise. In addition, the failed projects are costly, not only from a financial point of view, but also they can hurt the spirits of team members, damage an organization's reputation, and there is also the opportunity cost of delaying work on other (potentially much more successful) projects. This is the motivation to take into account the past experiences that are stored in the ERP database for improving the estimation quality of project parameters.

The knowledge discovery process in the context of an ERP database can be considered as four steps: data selection, data transformation, data mining, and pattern interpretation. An enormous number of data and attributes in an ERP database require paying attention to the proper choice of variable reduction and data mining methods. In the paper, the principal component analysis was chosen as the variable reduction method and the fuzzy neural system as the data mining technique. The fuzzy neural system has the advantages of both neural networks (e.g. learning abilities, optimization abilities and connectionist structures) and fuzzy systems (simplicity of incorporating expert knowledge). As a result, it is possible to bring the learning and

computational power of neural networks into fuzzy systems and also human like if-then thinking and reasoning of fuzzy systems into neural networks.

The advantages of the proposed approach include the search of conditional rules into an ERP database and using them for the project parameter estimation. This is especially important in the case of a significant variance of a project parameter, for which the average and time series models result in inaccurate estimates. The more exact identification of project duration and cost enables more precision of project planning and control, as well as the improvement of cash flow planning. This approach is especially valuable for an enterprise that has a database of past projects, because there is the possibility to gather additional information in the form of conditional rules. The application of the proposed approach encounters some difficulties, among other things, by collecting enough amounts of data of past similar projects. Moreover, the lack of uniform rules that concern the development of fuzzy neural systems may cause an acceptance problem for the decision-makers. However, the presented approach seems to have promising properties for acquiring information from an ERP system.

Further research focuses on the development of the proposed method for project parameter estimation in the context of ERP systems, for instance, towards the choice of an optimal set of new products. Moreover, future research can be aimed at adjusting the proposed approach in the aspect of risk management in the project, as well as the verification of the practicality of the proposed approach in a real world setting.

References

Advanced product quality planning and control plan. Reference Model, Carwin Continuous Ltd., Essex 1994.

Azar A.T., *Adaptive neuro-fuzzy systems*, [in:] *Fuzzy systems*, InTech 2010.

Bagiński J., Rostański M., *The modeling of business impact analysis for the loss of integrity, confidentiality and availability in business processes and data*, "Theoretical and Applied Informatics" 2011, vol. 23, no. 1, pp. 73-82.

Cabena P., Hadjinian P., Stadler R., Verhees J., Zanasi A., *Discovering Data Mining: From Concepts to Implementation*, Prentice Hall, 1998.

Cheng M.Y., Tsai H.C., Sudjono E., *Evolutionary fuzzy hybrid neural network for project cash flow control*, "Engineering Applications of Artificial Intelligence" 2010, vol. 23, pp. 604-613.

Chien S.C., Wang T.Y., Lin S.L., *Application of neuro-fuzzy networks to forecast innovation performance*, "Expert Systems with Applications" 2010, vol. 37, pp. 1086-1095.

Colmenares G.L., Perez R., *A data reduction method to train, test, and validate neural networks*, Proceedings of IEEE, Southeastcon 1998, pp. 277-280.

Davenport T., *Putting the enterprise into the enterprise system*, "Harvard Business Review", July-August 1998, pp. 121-131.

Fayyad U., Piatetsky-Shapiro G., Smith P., *From data mining to knowledge discovery in databases*, "American Association for Artificial Intelligence", Fall 1996, pp. 37-54.

Generowanie wiedzy dla przedsiębiorstwa: metody i techniki, red. M. Nycz, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław 2004.

Han J., Kamber M., *Data Mining. Concepts and Techniques*, 2nd edition, Morgan Kaufmann Publishers, San Francisco 2006.

Imtiaz A., Kibria M.G., *Modules to optimize the performance of an ERP based integrated information system*, IEEE International Conference on Informatics, Electronics & Vision, 2012, pp. 598-601.

Marban O., Mariscal G., Segovia J., *A data mining & knowledge discovery process model*, [in:] *Data Mining and Knowledge Discovery in Real Life Applications*, I-Tech, Vienna 2009.

May J., Dhillon G., Caldeira M., *Defining value-based objectives for ERP systems planning*, "Decision Support Systems" 2013, vol. 55, pp. 98-109.

Relich M., *A decision support system for alternative project choice based on fuzzy neural networks*, "Management and Production Engineering Review" 2010, vol. 1, no. 4, pp. 46-54.

Relich M., *Project prototyping with application of CP-based approach*, "Management", 2011, vol. 15, no. 2, pp. 364-377.

Zeng J., An M., Smith N.J., *Application of a fuzzy based decision making methodology to construction project risk assessment*, "International Journal of Project Management" 2007, vol. 25, pp. 589-600.

POZYSKIWANIE WIEDZY Z BAZY DANYCH SYSTEMU ERP W ASPEKCIE OPRACOWANIA NOWEGO PRODUKTU

Streszczenie: Celem artykułu jest przedstawienie możliwości wykorzystania bazy danych systemu ERP do identyfikacji zmiennych, które istotnie wpływają na czas realizacji poszczególnych faz projektu. W tekście przedstawiono wybrane metodyki dotyczące procesu odkrywania wiedzy z baz danych oraz model pozyskiwania wiedzy z baz danych systemu ERP. Zaprezentowane podejście jest dedykowane dla przedsiębiorstw produkcyjnych, które wykorzystują system ERP do monitorowania etapów opracowania nowego produktu. Przykład obejmuje cztery fazy procesu odkrywania wiedzy, tj. wybór danych, transformację danych, drążenie danych i interpretację wyników. Spośród technik drążenia danych został wybrany system rozmyto neuronowy, którego zadaniem jest szukanie relacji dotyczących zakończonych projektów i innych danych zgromadzonych w systemie ERP.

Słowa kluczowe: zarządzanie wiedzą, zarządzanie projektem, wybór danych, drążenie danych.