VERIFICATION OF TEXT MINING TECHNIQUES ACCURACY WHEN DEALING WITH URBAN BUSES MAINTENANCE DATA

Mateusz MARZEC, Tadeusz UHL, Dariusz MICHALAK

AGH University of Science and Technology, Dept. of Mechanical Engineering and Robotics
Al. Mickiewicza 30, 30-059 Kraków, Polska, e-mail: mamarzec@agh.edu.pl

Summary

Constantly increasing maintenance costs impose optimal maintenance policy planning. One possible way which helps to minimize maintenance costs and prevent bus fleet availability is analysis of historical maintenance records, which contain information about failures and performed repairs. In many cases this data have free text form and their analysis require individual log-by-log examination of their content. In order to automate this process, text mining methods can be applied. But, accuracy of the analysis depends on data quality and employed methods and should be tested before using this approach. This is especially important when the service decisions, which influence safety and maintenance costs, are made on this basis.

The aim of this paper is to determine whether existing and currently used text-mining methods are sufficiently accurate to be used in classification of unstructured urban bus maintenance and repair data. For that purpose the case study and literature review has been conducted.

The study shows great capabilities of proposed classification model. The model has 99% of accuracy and can be applied to support maintenance decisions.

Keywords: text mining, maintenance optimization, urban buses maintenance management

WERYFIKACJA DOKŁADNOŚCI METOD TEXT MINING W PRZYPADKU ANALIZY HISTORYCZNYCH ZBIORÓW DANYCH DOTYCZĄCYCH SERWISU AUTOBUSÓW MIEJSKICH

Streszczenie

Stale rosnące koszty utrzymania taboru autobusowego wymuszają potrzebę kształtowania odpowiedniej polityki serwisowej. Niezbędna w tym zakresie jest analiza danych historycznych, które zawierają informację o zaistniałych awariach i wykonanych naprawach. W wielu przypadkach dane te posiadają formę tekstową, co wymaga ich indywidualnej oceny rekord po rekordzie. W celu zautomatyzowania tego procesu istnieje możliwość zastosowania metod klasy text mining. Aby jednak wyniki analizy text mining mogły zostać wdrożone muszą wykazywać się one odpowiednią dokładnością. Jest to szczególnie istotne w przypadku, gdy na podstawie tych wyników podejmowane są decyzje serwisowe wpływające na bezpieczeństwo i koszty eksploatacyjne.

Celem niniejszego artykułu jest weryfikacja, czy powszechnie stosowane metody text mining są wystarczająco dokładne, aby analizować historyczne dane serwisowe autobusów. W tym celu dokonano przeglądu literaturowego oraz analizy text mining tego konkretnego typu danych.

Przeprowadzone badania wykazały, że dokładność klasyfikatora wynosi 99%. Na tej podstawie można stwierdzić, że są to metody wystarczająco dokładne, aby za ich pośrednictwem podejmować decyzję serwisowe.

Słowa kluczowe: text mining, optymalizacja serwisu, zarzadzanie flotą autobusów miejskich

1. INTRODUCTION

The economic impact push bus operators to decrease cost of operations. One of the most important part of these costs are maintenance and repair costs [4][8][14]. Optimization of the maintenance and repair cost is challenging for many bus operators. Nowadays, bus manufacturers and operators collect a great amount of maintenance and repair data of their buses fleets. These data contain a possible wealth of information that can be used to develop the maintenance procedures of buses, reduce downtimes, and prevent failures. These data could be analyzed using computer assisted methods to achieve information useful for operation cost decreasing. It has been estimated that about 80% of all data, collected by operators are in the form of unstructured text documents or records [1]. Unstructured data format makes impossible to automatically extract important and useful information from them. Therefore the data require individual examination log-by-log of their content, which is very inefficient and time consuming. Consequently there is a necessity to use appropriate methods, which would be able to transfer unstructured bus maintenance data into structured attributes, which can be accomplished with Text Mining techniques. Advanced text analysis techniques show a great potential in extracting of valuable information from the complex, varying, and incomplete entries usually found in these maintenance logs. However, results obtained by the usage of such techniques are always more or less accurate. In many cases, misclassification rate of text mining model is crucial for given applications and should be assess, mainly when dealing with economic/safety - related decisions.

The aim of this paper is to determine whether existing and currently used text-mining methods are sufficiently accurate to be used in classification of unstructured urban bus maintenance and repair data. If the methods are enough accurate then they can be applied for supporting maintenance decision. For that purpose the case study and literature review was conducted at first.

The study was carried out for one of the biggest bus manufacturer in Europe. This company collects a great amount of data in semi-structured form derived from warranty accounting software. This data contains structured information concerning bus mileage at the time of service as well as description of performed activities in the free text form.

The aim of presented study was to determine if particular description is related to corrective or preventive maintenance. Corrective maintenance (CM) means all actions performed as a result of failure, to restore an item to a specified condition. Preventive maintenance (PM) means all actions performed in an attempt to retain an item in specified condition by providing systematic inspection, detection, and prevention of incipient failures [18]. If the accuracy of the categorization is high enough, then the method can be basis for decision. In the case study it has been assumed that acceptable accuracy is 95%.

There are many literature works where text mining methods were used to classify unstructured data into categories. Some of them are strictly related to maintenance in automotive industry [5][7][9][12]. Authors in [9] used the text mining to map the problem description to their appropriate diagnostic categories, such as engine, electrical, brake, and transmission. The techniques, such as text document categorization and term weighting schemes, similarity functions, and latent semantic indexing are used to cluster the data and to identify the similarities between a problem description and a diagnostic categorization. Performances of the proposed approach were less than 80%. In [7] author applied text mining techniques to categorize customer feedback on new cars obtained through phone survey results and transcribed phone calls. The system assigned documents into predefined and dynamically created categories respectively. Author claimed a 90% recall and precision. In [12] author proposed a novel ontology-based text mining system to efficiently analyze unstructured textual diagnosis data collected in automotive domain during the warranty period to identify best-practice repairs. The average accuracy of the model was 88%. Issue closest to the one we faced in our case study was presented in [5]. In this paper authors used text data mining techniques to analyze historical data from dam pump station maintenance logs stored as free text form. The goal was to classify the data as scheduled maintenance or unscheduled repair jobs. This case study tested a decision tree trained using term weights and a neural network trained on SVD (Singular Value Decomposition) components of data set. Depending on used classifier misclassification rate varied from 15% to 17%. The best performance was obtained from the decision tree trained to learn the target variable using the term weights as input.

Basing on literature review it can be noticed that accuracy of presented models is always below 95% which is unacceptable in a case of supporting maintenance decisions. It also can be seen that accuracy of text mining techniques depends mostly on complexity of a target, used algorithms, nature of data, their quality and amount. Therefore, it cannot be directly assessed that text-mining methods are sufficiently accurate to be used in case study of bus fleet maintenance data. As a consequence, study of this particular case has to be conducted. In our research, due to similarity of the study conducted in [5] decision tree as an algorithm, which had the best accuracy in the reported study has been used. To decrease misclassification rate, mentioned in [5], authors

used IDF (Inverse Document Frequency) together with SVD algorithm in order to get proper input for the decision tree (terms IDF and SVD were explained in chapter 2).

The paper is structured as follows: in chapter 2 text mining process as well as commonly known methods and algorithms were described. Moreover, typical methodology which can be used to automatically classify maintenance data were presented. In chapter 3 the case study of bus fleet, in which mentioned methodology was applied was shown. At the end, in chapter 4 summaries of obtained results and indications for future research were presented.

2. TEXT MINING PROCESS AND TECHNIQUES

Text Mining can be defined as a technique which is used to extract interesting information or knowledge from the text documents which are usually in the unstructured form [3].

Text mining is expansion of data mining into the zone of tasks and applications connected with text documents. Most importantly, these documents do not need precisely determined structure. *Text mining* process has several stages. In the frame of first stage it is necessary to define objectives and scope of the analysis. Knowledge of the purpose of analysis enables suitable preparation of data, choice of correct techniques of analysis and pattern of representation of results.

Initial data processing is transformation of documents into text form (e.g. through removal of HTML tags when documents come from a website), unification of the way of coding of characters characteristic for particular language (e.g. ó to o or ä to a) as well as spell check and check of abbreviations.

A very important step of text mining analysis is determination of a method of storage of information from processing of collection of documents in a computer system. Currently, the most popular method is presentation based on a list of words occurring in a particular document. This method assumes, that the structure which represents documents is a vector whose individual elements inform about the number of occurrences of individual words. Representation of this type requires pre-processing of a document through removing punctuation marks and words unessential for analysis (by using so-called stop list) as well as conversion of words occurring in a document into its basic form (stemming).

Tab. 1. Construction of incidence matrix

F1*: "change batteries measuring charging, charging good but poor battery capacity"*
F2: *"replacing of 1 unit start batteries"*
F3: *"replacing the batteries. Poor capacity of both batteries"*

	F1	F2	F3	F_j
change	1	0	0	...
batteries	1	1	1	...
measuring	1	0	0	...
charging	2	0	0	...
replacing	0	1	1	...
unit	0	1	0	...
start	0	1	0	...
...	...	...	...	...
i	...	...	...	x_{ij}

Words that come from separate documents are combined into one, common list. Then frequency of every word in every document is counted. Collected data create a incidence matrix A. As an example incidence matrix for documents F1, F2 and F3 is presented in the table 1. Columns of the matrix are documents considered in the data set, while rows represent consecutive words occurring in the documents. The Matrix elements x_{ij} determine the number of occurrences of i^{th}word in j^{th} document. An essential stage of preparation of the incidence matrix is removing words, which occur too often or too seldom because they do not indicate similarity with considered documents. Basing on the matrix it can be inferred that separate documents are similar. It can be conducted by comparison of matrix rows and columns.

Determined frequencies of words occurrence can be transformed to emphasize common features between documents, which is very helpful in such analysis. In many practical cases similarity of documents could be better prove based on occurrences of the same words than precise number of their occurrences. To the most often used methods of transformation of given words frequencies are [1] [10]:

- **Binary frequencies** method assumes that registered is only the fact of occurrence of i^{th} word in j^{th} document. The effect of a binary representation is a binary matrix, where 1 means occurrence of the word and 0 its non-existence.

- **Log Frequencies** method consists in replacing all non-zero elements of incidence matrix into values *(1)*.

$$x_{ij}^{'} = (1 + \log(x_{ij})) \qquad (1)$$

  Log frequencies method is similar to binary frequencies method, except that binary representation considers the number of word occurrences to a lesser extent.

- **IDF (Inverse Document Frequencies)** basis for definition of the method was observation, that if a word occurs in every considered document, then the method does not enable to distinguish and group texts, while in terms of possibilities of classification particularly useful are words occurring in relatively few documents. This kind of approach reflects in the application of formula [11]:

$$idf(i,j) = \begin{cases} 0, & if\, x_{ij} = 0 \\ (1 + \log(x_{ij})) \log \frac{N}{df_i}, & if\, x_{ij} \geq 1 \end{cases} \quad (2)$$

where:
N - the total number of documents,
df_i - is the document frequency for i'th word.

Proper representation of information included in text documents gives possibility of application of suitable text mining methods, which enable to realize the aim of the analysis. Mostly used technique of text mining is **Singular Value Decomposition SVD**. In a case of an analysis of big sets of documents including big amount of words, appear difficulties connected with big size of the incidence matrix. SVD is a method, which enables to reduce a matrix size [13]. The method bases on application of SVD in relation to the frequency matrix. The diagram of frequency matrix according to singular values presents the formula (3):

$$A = U\Sigma V^T \quad (3)$$

where,
A - incidence matrix
U - matrix of words in space determined by elements,
V - matrix of documents in space determined by elements,
Σ - diagonal matrix, meaning of consecutive elements.

The aim of the calculations is to define the space, in which an analysis of sets of words occurring in documents (matrix U) and an analysis of a set of documents (matrix V) would be possible. Every document is represented by one row of particular matrix, whereby particular coordinates are in descending order according to its information values. It enables to reduce the space through consideration of some number of initial elements of the vector.

Another technique used in Text Mining could be a **C&RT (Classification and Regression Trees)** method. The method enables both construction of models, which can be used to solve regressive problems (where the dependent variable is quantitative attribute) as well as classification problems solution (qualitative dependent variable). Classification trees are used to determine affiliation of objects (documents) to classes of qualitative dependent variable (preventive or post-accident replacement) based on measurements of one or more explanatory variables (predictors - words) [15]. Classic C&RT algorithm was presented in [2]. The result of functioning of the algorithm is represented as a decision tree or as rules.

Detailed description of commonly used methods and algorithms of text mining was presented in [6]. Example of usage of those methods is presented in next section of the paper.

3. CASE STUDY

The main goal of current research is to optimize maintenance policy for bus fleet, which would reduce downtimes and cost of maintenance. For the optimization purpose it is necessary to define the characteristics of buses failures as cumulative distribution functions. To estimate distribution parameters, distance between failures has to be known. The company collects a great amount of data in semi-structured form derived from warranty supporting software. Each report in the database comprises information including ID of replaced part, present bus mileage as well as description of performed activity in free text form. The description can be related to corrective or preventive (e.g. product recall) activities. The main problem is defining if particular text is related to failure or not, because only on that basis it is possible to calculate distance between failures. Manual classification of hundreds in set of thousands system logs can be extremely inefficient. In this analysis use of textual features obtained from unstructured data is transformed automatically to classified records. If the accuracy of the classification is high enough, then decisions can be based on that data.

A case study was undertaken based on more than two hundred thousand reports prepared in polish language. At the beginning, with usage of Microsoft Excel, misspellings, abbreviations and characters were checked. Those actions were to get better text coherence. Next fifteen thousand randomly selected reports were expertly classified as related to preventive activities or not. As a result, spreadsheet contained two columns: description of performed maintenance activity in free text form and the second column, which specifies type of performed activity (CM or PM) in binary form. Data prepared in this way enabled to build and test data mining model.

The text mining analysis was conducted with usage of *Statistica Data Miner* Software. Firstly, it is necessary to represent the text data in computer memory as a vector of words. For that purpose words from unstructured text data are indexed. Many of these indexed words recur in documents very frequently but are essentially meaningless for

the analysis i.e. "bus", "cracow", "unit" or "driver". These words are omitted by the usage of stop-list. Then the words are stemmed i.e. reduced to their basic form. From almost 3000 words indexed by the software only 206 occurred more than once. Experts examined those words and chose only these, which could really influence dependent variables (performed activity). Finally 76 words were selected. Te słowa zostały wykorzystane do zbudowania macierzy częstości. Liczba wystąpień każdego słowa w macierzy została przedstawiona z wykorzystaniem algorytmu IDF. By employing of SVD method concepts were extracted [17]. Then a Chi-square statistic and p-value for each word and concepts, to find these, which have strongest relation with dependent variable were computed. The plot presented in figure 1 shows 15 words and concepts with the strongest relationship to dependent variable. Using this set of words and concepts classification and regression tree (fig. 2) can be obtained. Final model is quite simple and uses 7 variables to split the data.

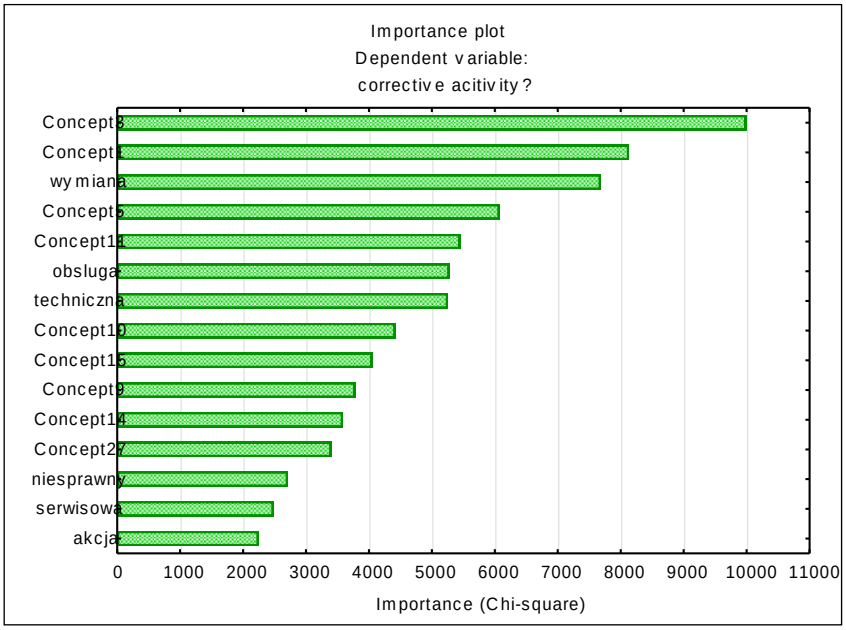

Fig. 1 Variables with the strongest relationship to dependent variable.

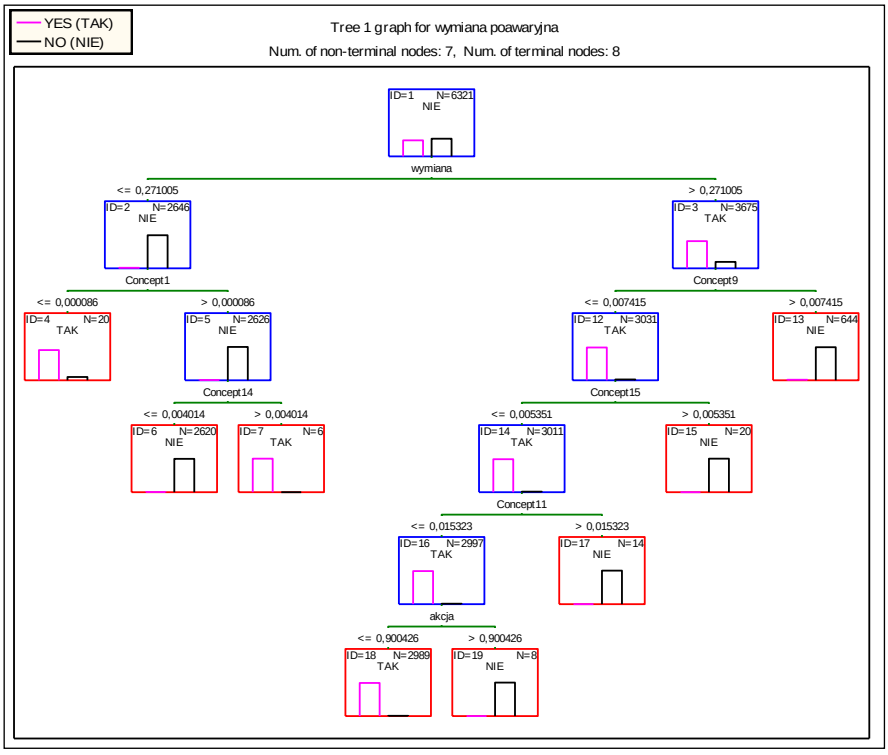

Fig. 2 Classification and regression tree

Tab. 2 Model accuracy data

	Observed	Predicted YES	Predicted NO	Row total
Number	YES	2981	25	3006
Column Percentage		98.87%	0.76%	
Row Percentage		99.17%	0.83%	
Total Percentage		47.16%	0.40%	47.56%
Number	NO	34	3281	3315
Column Percentage		1.13%	99.24%	
Row Percentage		1.03%	98.97%	
Total Percentage		0.54%	51.91%	52.44%
Count	All Groups	3015	3306	6321
Total Percent		47.70%	52.30%	

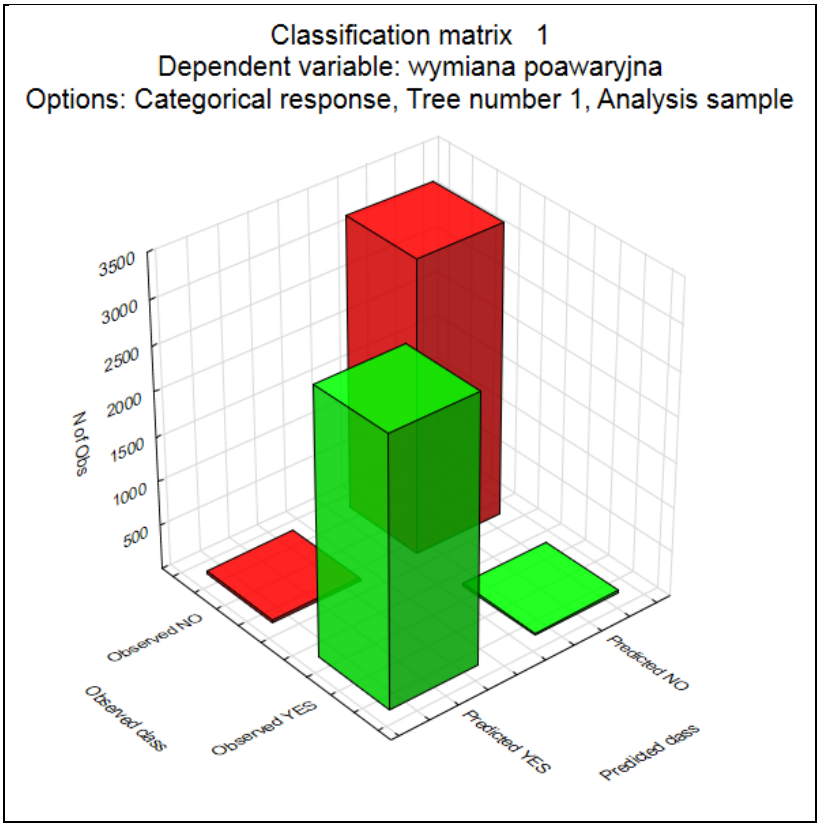

Fig. 3 Model accuracy plot

To check an accuracy of the model the data (50/50) can be randomly split for training and testing purpose. The accuracy of the model is shown in the table 2. Histogram shown in fig.3 is a graphical view of the information. These results show that the model has 99% of accuracy. Then, the model can be deployed to automatic classification of more than two hundred thousand records.

4. CONCLUSIONS

The aim of this paper was to determine whether commonly used text mining methods are suitable for analysis of bus maintenance data. For that purpose case study was conducted. Approach used in the study bases on IDF representation, SVD method as well as classification and regression tree. The study showed great capabilities of the classification model. The model has 99% of accuracy which means that decision can be made based on those methods.

In the future more sophisticated data mining model will be made - it will be able to indicate detailed failure cause i.e. crack, leak or spill, on the basis of its description. If the accuracy of such a model is sufficient, then it will be possible to use it to automate the processes related to account the warranty repairs.

ANKNOWLEGMENT
We would like to thank StatSoft Polska for providing the Statistica software and technical support during analysis process.

5. BIBLIOGRAPHY

[1] Lula P. *Text mining jako narzędzie pozyskiwania informacji z dokumentów tekstowych.* Data Mining: poznaj siebie i swoich klientów, StatSoft Polska 2005
[2] Breiman L., Friedman J., Stone C. J., Olshen R.A. *Classification and Regression Trees.* Chapman and Hall, 1984
[3] Divya N. *Text Mining Techniques - A Survey.* International Journal of Advanced Research in Computer Science and Software Engineering, Vol.2 (4), 2012
[4] Department of Defence. *DOD Guide for achieving Reliability, Availability, and Maintainability.* Department of Defence, USA, 2005
[5] Edwards B., Zatorsky M., Nayak R. *Clustering and Classification of Maintenance Logs using Text Data Mining.* Australasian Data Mining Conference 2008, Australia, November 2008
[6] Ghosh S., Roy S., Bandyopadhyay S. *A tutorial review on Text Mining Algorithms.* International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1 (4), 2012 7
[7] Grivel L. *Customer feedbacks and opinion surveys analysis in the automotive industry*. Text Mining and its applications to intelligence, CRM and Knowledge Management., WITpress, 2006
[8] Higgins L., Mobley K. *Maintenance Engineering Handbook.* McGraw Hill, 2002
[9] L. Huang, Y.L. Murphey. *Text mining with application to engineering diagnostics.* Proceedings of IEA/AIE'2006, vol. 4031, Lecture Notes in Computer Science, 2006
[10] Manning C., Schutze H. *Foundations of Statistical Natural Languange Processing.* MIT Press, Cambridge, 1999
[11] Rajaraman, A., Ullman, J. D. *Data Mining*. Mining of Massive Datasets. pp. 1–17, 2011
[12] Rajpathak D. G. *An ontology based text mining system for knowledge discovery from the diagnosis data in the automotive domain.* Computers in Industry (64), p. 565–580, 2013
[13] Russ A.: *Taming Text with the SVD.* SAS Institute Inc., Cary, NC, 2004
[14] Smith R., Hawkins R. *Lean maintenance: reduce costs, improve quality, and increase market share.* Elsevier, 2004
[15] StatSoft (2006). Elektroniczny Podręcznik Statystyki PL, Krakow, WEB: http://www.statsoft.pl/textbook/stathome.html.
[16] StatSoft, Inc. (2011). STATISTICA (data analysis software system), version 10. www.statsoft.com.
[17] Sumathy K.L., Chidambaram M, Text Mining. *Concepts, Applications, Tools and Issues – An Overview*. International Journal of Computer Applications, Vol. 80 (4), 2013
[18] Wang H. *A survey of maintenance policies of deteriorating systems.* European Journal of Operational Research (139), p.469–489, 2002

Mateusz MARZEC M.Eng. – graduate of the Faculty of Mechanical Engineering and Robotics at the AGH University of Science and Technology in Kraków. The main areas of his interests are reliability engineering and assets management.

Prof. **Tadeusz UHL** Ph.D– head of the Department of Robotics and Mechatronics at the AGH University of Science and Technology in Kraków. In his works he explores issues of structural dynamics, especially modal analysis and model based diagnostics. He is also interested in broadly understood mechatronics.

PhD **Dariusz MICHALAK** is the Solaris Bus & Coach Deputy CEO. He graduated of the Faculty of Machines and Transportation at the Poznań University of Technology. Michalak joined Solaris in 1998 as body design engineer. Since then, he has been involved in developing the Solaris product range and has been responsible for design projects including hybrid and electric buses, trolleybuses and trams. Since 2006 he is Director of Research and Development, being also responsible for the cooperation with European technology partners. In 2012, he was appointed to the Solaris management board.