USING DATA MINING TOOLS TO SHOW CORRELATIONS BETWEEN FAILURES OCCURRING IN CITY BUSES

Mateusz MARZEC, Tadeusz UHL

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 ; tuhl@agh.edu.pl

Summary

A failure in a bus or other technical device increases its operational costs. Apart from repair costs, a failure might make the work scheduled for the given time impossible, which leads to financial consequences equalling the value of the unaccomplished work or its effects. Because of that it is very important to run research aimed at improving reliability.

The following work presents application of data mining tools that show correlations between failures of various components. Such an approach enables improving the reliability of buses or other technical devices by pointing out design errors and defining control procedures that enable early failure detection.

The basket analysis done in this work is based on the databases in which baskets were created with the use of an innovative programme with a dynamic frame that segregated data.

Keywords: data mining, basket analysis, reliability of buses.

WYKORZYSTANIE NARZĘDZI DATA MINING DO OCENY AWARII AUTOBUSÓW MIEJSKICH

Streszczenie

Awaria autobusu lub innego urządzenia technicznego przyczynia się do zwiększenia kosztów eksploatacji. Poza kosztami napraw, awaria powoduje, że praca zaplanowana w określonym czasie może nie zostać wykonana. Pociąga to za sobą konsekwencje finansowe odzwierciedlające wartość niewykonanej pracy bądź jej efektów. W związku z powyższym bardzo ważne jest, aby wykonywać badania mające na celu zwiększenie niezawodności tych obiektów.

Niniejsza praca proponuję aplikację podejścia z wykorzystaniem narzędzi data mining do wskazania relacji miedzy uszkodzeniami poszczególnych części. Takie praktyki pozwalają na zwiększenie niezawodności autobusów lub innych obiektów technicznych poprzez wskazanie błędów konstrukcyjnych oraz procedur kontrolnych mających na celu wczesne wykrywanie uszkodzeń.

Analiza koszykowa przeprowadzona na potrzeby niniejszego opracowania korzysta ze zbiorów danych, w których koszyki zostały stworzone z wykorzystaniem innowacyjnego programu działającego w oparciu o dynamiczną ramkę dokonującą podziału danych.

Keywords: data mining, basket analysis, reliability of buses.

1. INTRODUCTION

Data mining is an analytical process used for examining large data sets in order to obtain regular patterns and systematic correlations between variables and evaluating the results by implementing the obtained patterns in new data subsets [1]. The data mining process is implemented mostly in fields such as banking, medicine and industry, but it can also be used in all applications in which computerised systems enable gathering empirical data in the form of databases. To obtain adequate patterns in a database, association rules can be used. An association rule is an implication in the form of (1):

$$X \Rightarrow Y \quad (1),$$

where X and Y are arbitrary subsets of elements from the β set and fulfil $X \subset \beta$, $Y \subset \beta$ i $X \cap Y = \phi$ [2].

The investigation of association rules was motivated by the Market Basket Analysis problem, implemented among others in hypermarkets in order to indicate groups of products often bought together – that can be found in the same basket. The gathered information can be used for such distribution of products on shelves that can result in higher sales. In case of city bus failures the association rules can provide information on simultaneous failures and on failures occurring in a specified sequence and on a given time. Thanks to such knowledge availability of buses can be increased by structural modifications or by optimization of maintenance services. The analyses of that type are especially useful in the economic analysis of buses' operation. For example, a failure that is cheap to service and seems insignificant may lead to a serious breakdown and put the vehicle out

of action, which together with the service costs generates considerable financial loss [3].

The importance of association rules is described by variables such as support, confidence and lift, the threshold values of which are assigned by the user while he defines the analysis. Thanks to that the association rules with minimal importance can be ignored.

The **support** of a set is a fraction of records containing that set. In other words, it measures how often a single- or multiple-element failure has occurred. High support of a given element or a set of elements signifies its high failure frequency. In case of databases describing buses' failure frequency the algorithm used to obtain association rules should account for minimum thresholds of support. An example association rule (3):

$$[\text{brake pads}] \Rightarrow [\text{brake pad sensors}] \quad (3)$$

is going to have much greater support than the following rule (4):

$$[\text{air compressor}] \Rightarrow [\text{brake hose}] \quad (4)$$

In sequence (3) replacing brake pads and sensors is included in maintenance services and such a sequence is insignificant from the point of view of analysis goals. However, it may be important that air compressor causes brake hose failures. In case of high support settings such a sequence would have been ignored because of low occurrence of such failures.

Confidence is a conditional probability that a set containing element A will also contain element C. In other words, that rule helps to determine what is the probability of failure A in case of failure B. The rules with 100% support might be helpful in preparing repair sets, which should speed up the process of issuing parts from the storehouse. Such information is also useful in determining control procedures to be observed during technical inspections of buses. For example, if one detects a failure of part A and support of a rule 'if A then B' is 100%, the control procedure will have to include an inspection of part B.

Lift is defined on the basis of confidence and support. In a set 'if A then C' with 100% confidence (a failure A is always followed by a failure C) with little support of failure C (failure C happens rarely), the set will have high lift. In other words, high lift of a set 'if A then C' will mean that a failure of element C has been most probably caused by a failure of element A. The information about the lift can be used in both preparing repair sets and control procedures, however, above all it is an information for the designer. In case of sets with high lift he should consider methods of eliminating the dependence of those failures.

In order to find the association rules in the database a module of STATISTICA Data Miner – Sequence, Association and Link Analysis has been used. The described tool utilizes FP-growth algorithm (tree-based pattern) which is more effective than the apriori algorithm in case of low minimal support or dense data sets (sets that contain many big frequent itemsets) [4].

2. BUS DATABASE

In order to obtain output data on bus failures and resultant costs a cooperation with MPK (Municipal Transport Company) in Kraków has been established. The most adequate source of data that enabled analyzing various aspects was a list of components that had been issued from the storehouse during the period of three years. As a component is ordered in a storehouse when a similar one in a bus breaks down, such a list is a great source of information on failures. What is more, the database provides not only information on precise dates of failures, but also on the cost of their repair.

However, it should also be noticed that such a way of reasoning may lead to some mistakes. For example, in case of brake hose abrasion the worn part will be repaired with the use of brake hose connector, so one the basis of the database one could assume that it was the connector that failed. During planned maintenance services the whole bus is checked and that can also lead to mistakes as failures that might not be linked would be discovered at one time. To sum up, the biggest disadvantage of the approach is the lack of information on the forms of failures, causes of failures, types of failures and on the components used to deal with a given problem.

3. DATA PREPARATION

At the beginning of the analysis the data have to be properly prepared. The failures might not only be correlated inside superior systems such as suspension or pneumatic system but also in between systems. For example, during a maintenance service brake disks and a door button can be replaced. In such case the programme will look for the following sequence (5):

$$[\text{brake disk}] \Rightarrow [\text{door button}] \quad (5)$$

It is quite obvious that the brake disk failure has nothing to do with the door button failure, so we should not expect correlations between the braking system and the bus body. However, it is possible that a failure of suspension air bellows might cause a failure of pneumatic system, and the other way round. Because of that, the correlations between the given parts and their superior systems have been investigated. Those correlations are described by the block diagrams in figure 1 which enabled determining the group of objects for further analysis. Thanks to such an approach mistakes caused by data from maintenance services could be eliminated.

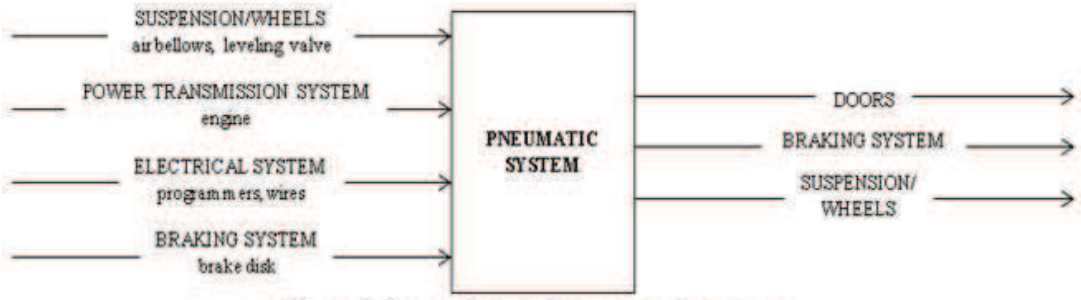

Figure 1. Interactions in the pneumatic system.

A crucial part in preparing data for analysis is determining the 'baskets'. In case of the described database each day may be such basket while all the parts collected from the storehouse to repair a given bus on that given day would be the items. However, such an approach might not always be correct as some repairs might last for two or three days, or even longer if a given part is not available in the storehouse. What is more, consequences of some failures might be visible only a few thousand kilometres later, so the parts that should be correlated would be placed in different baskets. The STSTISTICA data miner module mentioned before enables both sequential and non-sequential analysis. Sequential analysis enables creating sets (baskets) with the use of time parameter, so that the number of days in one basket can be defined. However, that solution is not correct in all cases. For example, if thanks to time parameter the programme created three-day-long baskets, a failure happening on the third day of basket *n* would not be linked to a failure that happened on the first day of the *n+1* basket. Because of that a programme in the Visual Basic environment had been written that enabled creating baskets with the use of a dynamically moving frame. An example in figure 2 explains functioning of the programme. It is assumed that revealing the consequences of each failure might take up to three days. If STATISTICA had been used together with the three-day-long time parameter sequential analysis, a failure from 3rd January would not be linked to the failure from 4th or 5th, because they would end in separate baskets. In case of the programme mentioned above that uses the dynamic frame, baskets can be defined in such a way that they contain a sequence of events. Input data in such a form enables using non-sequential analysis to finally solve the problem.

However, it is important to notice that such an approach might distort the real values describing the association rules. For example, the part 7364 appears in three baskets, while in reality its failure only happened once. Still, if we concentrate on the aim of the analysis, which is showing the correlations between failures, that distortion does not seem so significant

4. RESULTS

The results of the analysis were presented on network diagrams and rule diagrams on which the interpretation could be based. For example, figure 4 presents a network diagram of a failure in door system, in which the following correlations characterised by high lift can be distinguished:

1. bottom arm tip (4) ⇔ door hinge with a potentiometer (3),
2. bottom door track (1) ⇔ bottom guide (2),
3. left bottom arm tip ⇔ right bottom arm tip,
4. bottom seal ⇔ rotary post,
5. bottom seal ⇔ top door arm,
6. rotary post ⇔ top door arm.

Chosen elements of the door system are presented in figure 3.

DATA			Normal Frame		Dynamic Frame			
Bus no.	Date	Part code	Part code	ID trans.	Part code	ID trans.	Part code	ID trans.
DC501	1-01-2009	1356	1356	1	1356	1	7364	3
DC501	2-01-2009	2485	2485	1	2485	1	1124	3
DC501	3-01-2009	7364	7364	1	7364	1	3845	3
DC501	4-01-2009	1124	1124	2	2485	2	1124	4
DC501	5-01-2009	3845	3845	2	7364	2	3845	4
DC501	6-01-2009	4445	4445	2	1124	2	4445	4

Figure 2. An example of creating baskets in the programme written in the Visual Basic environment

Figure 3. Components of the door system: 1-bottom track, 2-bottom guide, 3-door hinge with a potentiometer, 4-door arm tip

In sequence 1 the bottom arm tip (4) wears quickly because of the difficult operating conditions and as a consequence doors operate with higher resistance and the door hinge with a potentiometer (3) also wears quickly. It should be noticed that repairs of the latter are the most costly repairs in the analysed bus. High lift in the sequence (2) is caused by the fact that its elements work together and wear one another. Such information might be used to point out errors in design and to establish adequate control procedures that would enable early failure detection.

Some of the presented relations might be obvious and have little importance in the context of requirements of the analysis. The sequence (3) does not mean that bottom arm tips influence each other's failure frequency. The relation is based on the fact that they are replaced at the same time because they work and wear alike. Similar is the relation with the bottom seal, which is replaced while the other elements are repaired. Such relations can lead to mistakes and are the result of complete lack of information on the forms of failures, causes of failures, types of failures and on the components used to repair a given failure.

5. SUMMARY

The described analysis with the use of data mining algorithm enabled pointing out a number of relations between failures of components in city buses. Such information can be used to increase the availability of buses or other technical devices.

On the other hand, a critical look on the form of the database used has to be presented. Above all, the database does not give information of how the failures happened, which makes the unequivocal assessment of association rules impossible. What is more, the credibility of the analysis is closely connected with the amount of analyzed data. If there had been access to more data (compared to the amount analyzed) it is very probable that more association rules would have been observed. Creating a huge database, for example in cooperation with other carrier companies, would also improve the analysis as each company's bus operation policy would be different and the condition of road infrastructure in various cities would influence the buses differently. However, the computerised systems that register the issued parts are a new introduction in many transport companies and differ one from another. Because of that the data collection process would have to be time consuming and would require individual approach to each company.

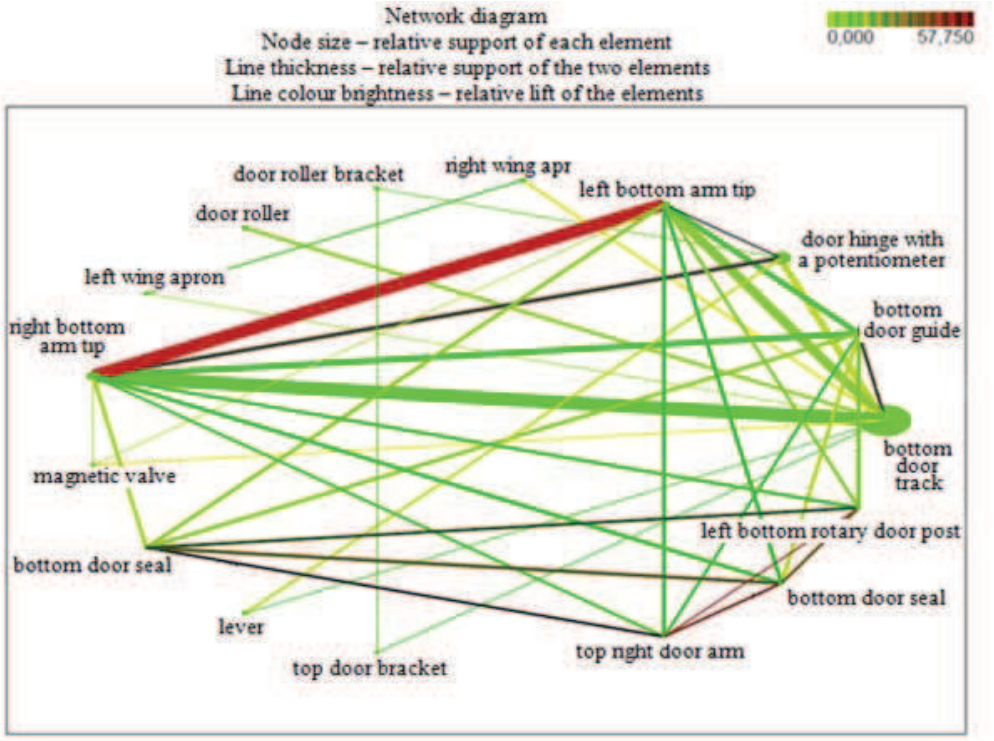

Figure 4. Network diagram of components of the door system

The described arguments show a huge need for a unified, computerised system that would gather empirical data on failures and service (repair anticipation time, repair time, number of servicemen etc.) and would be used to increase the reliability of city buses and to optimize maintenance services. The system would function on a website accessible for servicemen who would have their individual logins and passwords. Thanks to a user-friendly interface based on drop-down lists and predefined information base (with information such as vehicle types and designations used by a given transport company) data gathering would be very convenient for the users of the system. The amount of information expected from each user would be optimised and adjusted to the expectations from the system. In case of a system oriented on increasing bus' and its components' reliability, apart from basic information it would require data on form and causes of failure and in case of a system oriented on maintenance services optimization, data on repair anticipation time, reasons for anticipation, repair time and number of servicemen would be required. Such information could be also used for ranking the failures on the basis of repair time, number of servicemen or the cost of components used. An advantage of such a system would be the fact that it would contain ready data (without the need to process them first) that could be easily analyzed with the data mining tools.

The authors would like to express their thanks to the Innovative Economy Programme, for the financial help in the project 'Mechatroniczne stanowisko testowe typu END LINE', project number POIG-01.03.01-12-035/08-00.

BIBLIOGRAPHY

1. Ogiela L., Tadeusiewicz R., Ogiela M.: "*Graph-Based Structural Data Mining in Cognitive Pattern Interpretation*", Lecture Notes in Computer Science, Springer-Verlag Berlin Heidelberg , 2006.
2. Klosgen W., Żytkow M.: "*Handbook of data mining and knowledge discovery*.", Oxford University Press, 2002.
3. Marzec M.: "*Analiza dostępności obiektów mechatronicznych*", master thesis, Akademia Górniczo-Hutnicza, 2012.
4. Zheng, Z., Kohavi, R. i Mason, L.: "*Real world performance of association rule algorithms*." Conference on Know ledge Discovery and Data Mining, San Francisco, USA, 2001.
5. Gołąbek A.: "*Niezawodność Autobusów*.", Politechnika Wrocławska, 1993.
6. Klosgen W., Żytkow M.: "*Handbook of data mining and knowledge discovery*.", Oxford University Press, 2002.

Mateusz MARZEC M. Eng. – graduate of the Faculty of Mechanical Engineering and Robotics at the AGH University of Science and Technology in Kraków

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.