Data analysis of information system TURAWA in the aspect of data mining methods

B. SZAFRAŃSKI[*], M. ZIEJA[**], J. WÓJCIK[*], K. MURAWSKI[*]

boleslaw.szafranski@wat.edu.pl
miroslaw.zieja@itwl.pl
jaroslaw.wojcik@wat.edu.pl
krzysztof.murawski@wat.edu.pl

[*]Military University of Technology, Faculty of Cybernetics
W. Urbanowicza 2, 00-908 Warsaw, Poland

[**]Air Force Institute of Technology
Księcia Bolesława 6, 01-494 Warsaw, Poland

The article is devoted to the analysis of data coming from the operation process and collected in computer system TURAWA, which focuses on supporting the management of flight safety in the Polish Air Force. The Armed Forces are equipped with a system, which collects and processes data concerning the whole air crew, all performed flights and all aircraft. The increasing opportunities in obtaining data and the continuous development of data mining methods allow to extract information never been known before, which, together with conclusions obtained from the data analysis, will help to improve the level of flight safety.

Keywords: flight safety, data mining, TURAWA.

DOI: 10.5604/01.3001.0012.2004

1. Introduction

Air Transport is becoming more and more popular mean of transport, what is indicated by the increasing tendency of the number of flights in the world [1]. In spite of this fact, the number of aircraft accidents in recent years remains stable, what is depicted in Figure1 [1–3].

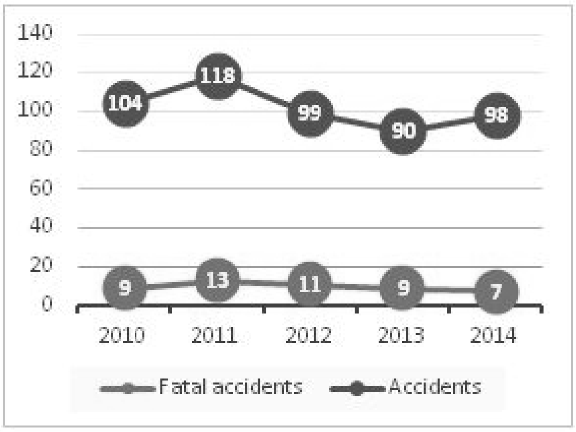

Fig. 1. The annual number of accidents and fatal accidents on the world in civil aviation in the years 2010–2014 [2]

Nowadays, it is not possible to improve the level of flight safety at a rate comparable to previous decades. Every activity aimed at improving the safety of air transport is associated with huge financial expenditure. However, a continuous increase of air operations results in searching for new, innovative solutions in this area, which will allow to maintain the acceptable level of flight safety. It is also worth mentioning the constantly increasing unit costs of aircraft accidents resulting from the application of modern technology.[1] In the pioneer years of aviation, a falling aircraft could only lead to the break of rafters and the destruction of the house's roof. Currently, the aircraft (a/c) crashing on the ground, equipped with tonnes of flammable fuel, might result in a disaster with unpredictable consequences [3]. Nowadays, to improve the level of flight safety mostly actions taken are based on reports compiled after the aircraft incidents, i.e. following a situation negatively affecting the level of safety of the performed flight [4]. Expert opinion suggests that the final reports, unfortunately, do not bring the desired improvement in the field of flight safety.

[1] Social costs are omitted here, because the value of human life, the loss of close friends or family members are hard to estimate.

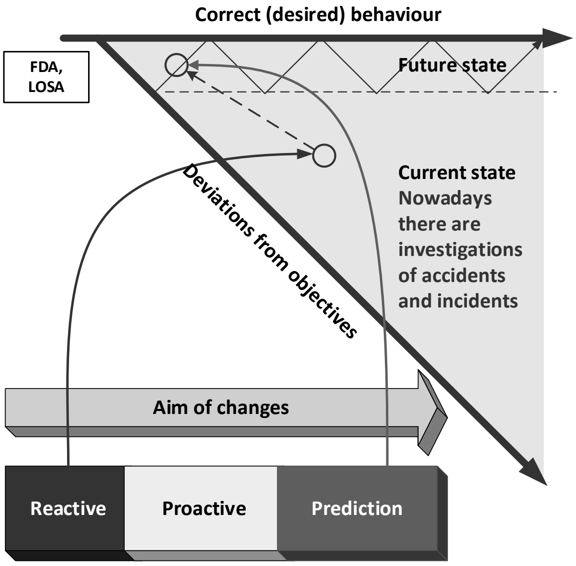

Fig. 2. Aim of changes – reaction to small deviation [3]

The reason for this may be the variety of accidents, in which losses occurred, a serious threat of life or A/C, what makes it difficult to apply the conclusions drawn from the analysed events. It is also worth mentioning that if a given aircraft incident did not result in serious consequences, it may only be caused by a fortunate coincidence. Thus, experts indicate that potential risks have to be diagnosed before the incident occurs [3]. The above view is reflected in the currently valid Aircraft Safety Instructions of the Polish Armed Forces, which imposes an obligation to identify threats having an impact on the flight safety to all services participating in the organization of flights. Additionally, Flight Safety Services shall evaluate the hazard of aviation risks and develop preventive measures [4].

The increasing possibilities of data collection and ever-increasing computing power of computers have increased the popularity of data mining techniques. In many articles, one tried to use them to explore "hidden" (i.e. so far unknown) information in analytical reports compiled after the aircraft accidents [5]–[10]. The Polish Air Force also has the IT system (IS), which contains a wide range of information describing (documenting) the operation process of A/C [11]. Apart from the information regarding aircraft incidents there are also data on aircraft personnel, performed flights or undertaken preventive measures. So in information resources of IS there are additional data, not included in reports from aircraft events. Applying suitable data mining methods to data analysis collected in the IS might support the identification of the so far "hidden" relationships between information data, very crucial from the point of view of the proper risk assessment of aviation threats.

2. Motivation

The article consist of the results of the analysis of the structure and information content of the database IS TURAWA, the application of which has a significant meaning for the process of maintaining a desired safety level of flights, within the widely understood operation of military A/C.

3. Structure and scope of IS TURAWA data

IS TURAWA supports the activity of the Polish Armed Forces in terms of flight safety since 2011. From that moment on, the data regarding every flight or aircraft incident are recorded in information resources of this system. The same regards the air crew, thus the system consists of the record containing, apart from personal data, i.a. data regarding the course of service, conducted aviation trainings, health condition.

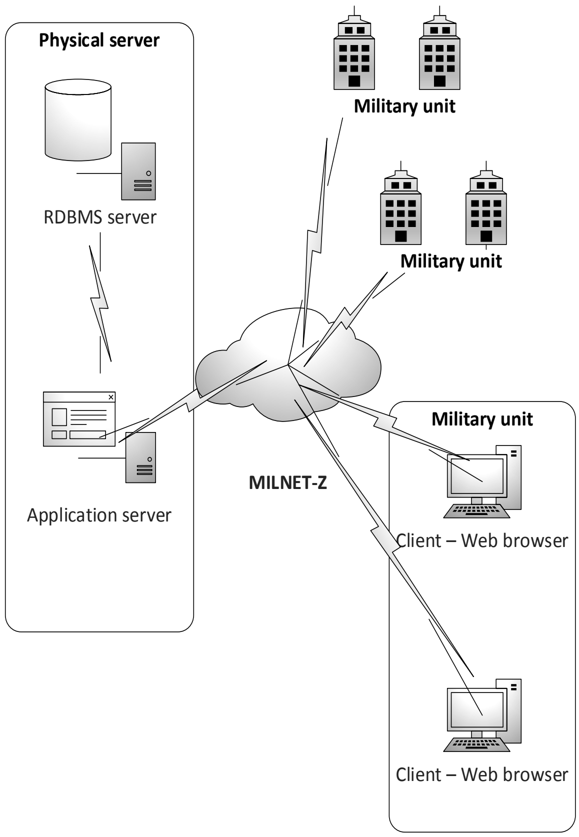

Fig. 3. Sketch of system architecture

The system was developed according to the classic three-tier architecture, which consists of the following elements:

- relational database system (data layer);
- application software in Java Platform technology (application layer);
- Internet browser interpreting generated HTML documents as well as sending requests generated by the user (presentation layer).

In IS TURAWA data are divided into 5 basic groups [9]:

1. Flying personnel.
2. Flights.
3. Aircraft incidents.
4. Aircraft.
5. Prevention.

Flying personnel. Every person being a part of flying personnel and performing flights in the framework of the Polish Air Force must be recorded in IS TURAWA. The flying personnel consists of:

- cabin personnel, i.e. pilots, navigators, flight engineers;
- on-board personnel, i.e. stewards, aerial gunners, loadmasters or paramedics.

All necessary data illustrate the whole course of the service, from the start until the removal from service. Additionally, the resources are supplemented with relevant data before the commencement of the service, for example: describing the course of study at the aviation school. The scope of the collected data is illustrated in Figure 4.

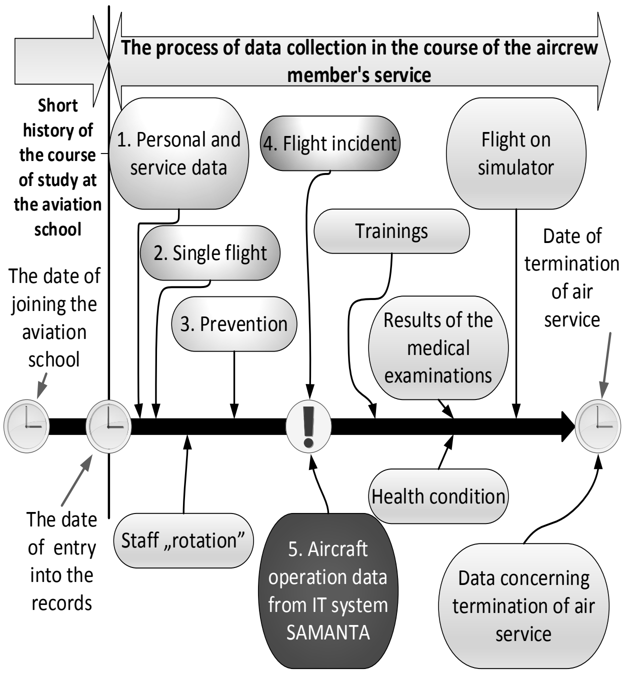

Fig. 4. Data collected during the service of the member of the flight crew

The most important data regarding every member of flying personnel include:

- **Basic data:**
  - personal data, i.e. name and surname, sex, date of entry into the records, date of promotion;
  - data regarding the course of service, i.e. the history of acquired specialities, history of occupied positions, rotation between units;
  - data on education and obtained certificates (i.e. from foreign languages).
- **Training:**
  - history of qualifications acquired, pilot licenses in a day and night, fulfilling the on-board position, etc.;
  - data of the inspections carried out in relation to the certain A/C type;
  - data on completed rescue and height training;
  - exam results from knowledge applied in relation to the certain A/C type.
- **Flying time:**
  - data on completed flights on A/C and simulators.
- **Health and leaves**:
  - history of medical examinations and medical boards, stays at hospitals and sanatoria;
  - data on completed training in condition centres (WOSZK), medical examinations in a low pressure test chamber;
  - data on the reasons and periods of incapacity for flights;
  - data on leaves.

Flights. Every flight performed within the Polish Air Force must be recorded in IS TURAWA. Due to storing the history regarding the course of service of the flying personnel it is possible to reconstruct the crew and the A/C status at the moment of performing the flight. The most important data characterizing the flight include:

- **Basic data:**
  - data regarding the unit, i.e. the place, identifier and name;
  - data concerning the chronology, i.e. when the flight took place, what was the flight duration, flight duration in a day and night according to VFR[2] or IFR[3];

[2] VFR – a flight performed according to regulations for flights with good visibility.

[3] IFR – a flight performed according to instrument flight rules.

  - category of flight, e.g.: training flight, on-demand flight, flight for the unit's individual needs;
  - data on the A/C and simulator, i.e. the number, type or sort of simulator.
- **Weather conditions:**
  - data on weather conditions, i.e. the visibility, special weather phenomena with division into particular flight stages.
- **Crew:**
  - data on crew, i.e. the list of persons taking part in the flight;
  - data on the functions performed by the crew and the distribution in the A/C.
- **Trainings and tests:**
  - data on completed trainings by particular crew members within the framework of training programme;
  - data on tests performed by particular crew members.

Aircraft incidents. The system has to record every aircraft incident[4], which occurred during performing the flight as well as the data from reports compiled after making an analysis by appropriately designated commissions. The scope of data of the delivered report depends on the significance of the impact of the event on the level of flight safety. The most important data related to aircraft incident are:

- **Basic data:**
  - data on the location and time of the occurrence;
  - class of aircraft event, which defines most of all the scope of effects;
  - data on the commission examining the aircraft event and the persons supervising its work from the Committee for Investigation of National Aviation Accidents (KBWL LP).
- **Circumstances, course and reasons of the incident:**
  - data regarding the circumstances and the course of the aircraft incident;
  - data providing reasons of a dangerous situation, that are the statistical codes that arise out of from 5M model of managing the risk of flight safety as well as the additional description. The direct and the main reason is recorded.
- **Flight information:**
  - data on performed flights by the A/C involved in the aircraft incident, i.e. the information resources described in the group Flights;
  - data on the A/C i.e. the information resources described in the group the Aircraft.

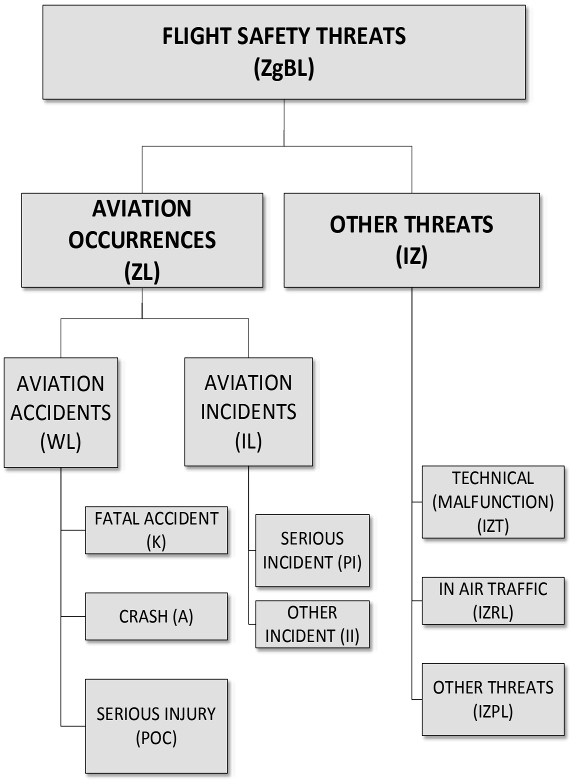

Fig. 5. The classification diagram of flight safety threats [4]

- **Bird strikes (if they were involved in the aircraft incident):**
  - data on birds, i.e. the size, species and number;
  - data on the flight phase, in which the aircraft incident occurred, i.e. the speed, altitude and type of the flight phase;

  data on the effects caused by bird strikes.
- **Prevention measures:**
  - Data on prevention measures implemented so as to minimise the risk of occurrence of the same incident.
- **Extended data (in the case of the extended report of the commission examining the aircraft incident):**
  - data regarding the description of A/C behaviour and the course of aircraft event;
  - data on weather conditions prevailing at the moment of aircraft incident;

[4] Situations exerting an influence on flight safety, e.g.: engine failure, unretained spatial separation, pilot lost consciousness.

- data on losses in the equipment and flying crew;
- data describing the conducted examinations;
- data regarding the reasons and flaws;
- conclusions of the commissions examining the aircraft incident.

Aircraft threats. Except for aircraft incidents, in IS TURAWA there are also recorded the observed aircraft risks, i.e. the situations or potential sources which may have a negative impact. The most important recorded data regarding aircraft hazards include:

- **Description and reason of the hazard:**
  - hazard description;
  - data on the reasons of hazard occurrence, i.e. the statistical code which arises out of the 5M model as well as the additional description.
- **Risk index:**
  - the history of changes in risk index, i.e. estimating the probability of occurrence and the scope of potential effects.

Aircraft. In the event of an air incident in IS TURAWA is recorded the current state of A/C involved in the situation that occurred. Information on technical parameters of the A/C are collected on the day of an accident in IS TURAWA. The most crucial data regarding the A/C contain:

- **Basic data:**
  - Registration data of the A/C and its subassemblies, that is its type, version, number, manufacturer and year of production.
- **Technical service life:**
  - Data on A/C technical service life and its subassemblies, i.e. what work was done from the beginning of the operation.
- **Overhauls and servicing:**
  Data on overhauls, servicing and bulletins.
- **Failures:**
  - Data on failures that occurred and its consequences.

Prevention. In IS TURAWA there are records of all preventive measures undertaken after aircraft incidents or in the case of detecting aircraft threats. The most important registration data in the framework of prevention measures included:

- **Prevention document:**
  - data on who and when managed preventive activity;
  - data on what was the reason for introducing prevention;
  - data on the relationship between the prevention document and aircraft incidents.
- **Preventive recommendations and their addressees:**
  - data on who, until when and for what preventive activities is responsible as well as about the entities implementing them;
  - data on the implementation of the preventive order, i.e. the date and potential remarks;
  - data on related documents issued for the purpose of the implementation of preventive order.

In the years 2011–2017, as part of the operation of IS TURAWA, about half a million of flights were recorded, during which about 9000 events were recorded. The structure of the relational database of the system includes 261 tables and approximately 1,400 columns, which proves its complexity.

4. The impact of data structure on the problem of data mining

The proper formulating of the task of data mining shall include the following components:

1. Aim of the task – it should be determined what shall be achieved as a result of solving the problem with the use of data mining methods.
2. Model structure or pattern structure – a function form or a certain structure, which is searched for, shall be determined.
3. Objective function – the way of evaluating the quality of model adjustment shall be determined.
4. Optimization and searching methods – a method of searching for the optimum model parameters or its optimum structure.
5. Strategy of managing data – the way of organizing data during its searching and processing by the methods of data mining.

The approach, in which the method of data mining is defined on the basis of components, enables to build it with the use of interchangeable tools or algorithms provided from the field of artificial intelligence or statistics [12].

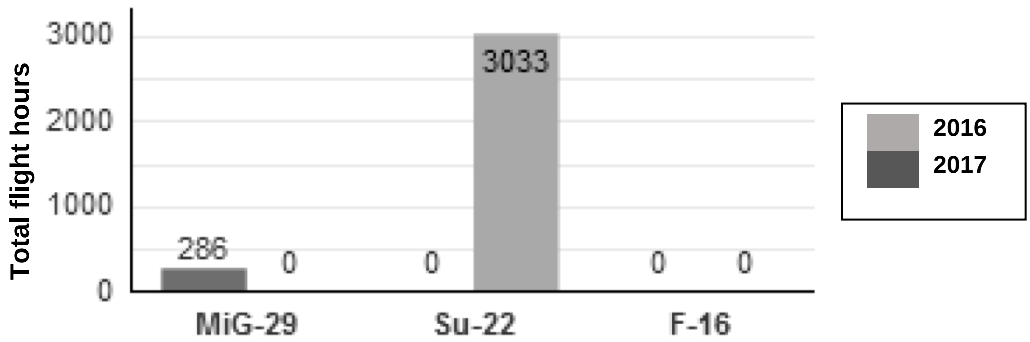

Fig. 6. The page fragment of IS TURAWA
that visualizes the average total flight hours
per number of aviation occurrences

In IS TURAWA there is few methods of data analysis that support depiction of the air safety level. For example, it is possible to show the average number of flight hours between aviation occurrences over a time period $p = \overline{1,P}$ for aircraft of a type $t = \overline{1,T}$ and an organization unit $o = \overline{1,O}$. This average can be expressed as:

$$A_{tpo} = \begin{cases} \frac{F_{tpo}}{E_{tpo}}, E_{tpo} > 0 \\ 0, E_{tpo} = 0 \end{cases} \quad (1)$$

Where $F_{tpo} \in \mathbb{R} \cup \{0\}$ is a total flight hours of aircraft of the type *t* over the time period *p* for the organization unit *o*, $E_{tpo} \in \mathbb{N}$ is a number of aviation occurrences of aircraft of the type *t* over the time period *p* for the organization unit *o*.

Another example are the Index of Other Incidents and the Average index of Other Incidents. Both indexes are used to illustrate the trend in the number of the Other Incidents depending on an intensity of flights. The Index of Other Incidents shows an average number of the Other Incidents per total flight hours in a given year $y = \overline{1,Y}$ for aircraft of types *t* and for the organizational unit *o*. It can be expressed as:

$$I_{yo} = \begin{cases} \frac{\sum_{t=1}^{T} c_t \cdot E_{tyo}}{\sum_{t=1}^{T} c_t \cdot F_{tyo}} \cdot S, \exists_{t,y,o} c_t = 1 \wedge F_{tyo} > 0 \\ 0, otherwise \end{cases} \quad (2)$$

Where $c_t \in \{0,1\}$ equals 1 when user included the type *t* of an aircraft in the index calculation and 0 when not, $F_{tyo} \in \mathbb{R} \cup \{0\}$ is a total flight hours of aircraft of type *t* in the year *y* for the organization unit *o*, $E_{tyo} \in \mathbb{N}$ is a number of aviation occurrences of aircraft of types *t* in the year *y* for the organization unit *o* and *S* is a constant value that is used to scale the result. The Average index of Other Incidents shows an average number of the Other Incidents per total flight hours over a period of 5 years up to the selected year *y* for aircraft of selected types *t* and for the organizational unit *o*. It can be expressed as:

$$J_{yo} = \begin{cases} \frac{\sum_{i=[y\geq 5]\cdot(y-4)+[y<5]\cdot 1}^{y} \sum_{t=1}^{T} c_t \cdot E_{tio}}{\sum_{i=[y\geq 5]\cdot(y-4)+[y<5]\cdot 1}^{y} \sum_{t=1}^{T} c_t \cdot F_{tio}} \cdot S, \\ \exists_{t,i,o} c_t = 1 \wedge F_{tio} > 0 \\ 0, otherwise \end{cases} \quad (3)$$

Lack of sufficient tools depicting the level of flight safety leads to research for new solutions, which will be based on the collected data. Data mining methods seem to be the right choice.

Before selecting an appropriate method of data mining it should be examined which tables contain interesting data. Not all implemented system functions were deployed.

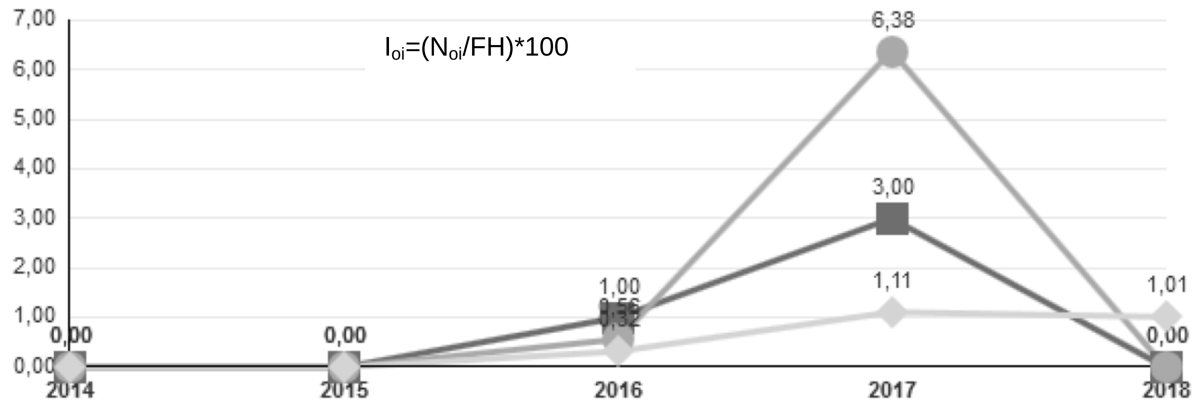

Fig. 7. The page fragment of IS TURAWA
that visualizes the number of Other Incidents,
the Index of Other Incidents and the Average index of Other Incidents

Due to this fact, the IS TURAWA database does not include all planned data. One should also remember about tables, that contains data resulting from operations on other data and closely related to them. These are the data that can be omitted. An example of such a situation is a sum of flight hours for a certain A/C types in a given month. Such data are used to determine and display the percentage of realization of the flight plan of the unit. There are also tables, which contain the dictionaries facilitating inserting data on the user forms. The example might be a screen of inserting a statistical code of the reason of the aircraft incident. Only the whole code gives some information. The dictionaries themselves are only used in the algorithm of determining a statistical code of the reason of the aircraft incident. It is also worth remembering about the constant development of the system, which generates changes in the structure of the database. An example here may be the causes of the aircraft incident. A new flight safety regulations forced changes in the dictionaries indicating the reasons of aircraft incident. However, to preserve the possibility to display aircraft incidents before the change, it was necessary to preserve old structures and dictionary values. This resulted in the creation of two different dictionaries defining the reason. It shall also be assumed that not all entered data are complete and correct.

The most common way of organizing data for algorithms of data mining is a matrix with *P* columns and *N* rows [12]. *N* rows means *N* objects recorded in the system, which are subject to analysis. *P* columns means *P* attributes describing the objects. Each value of the attribute represents a measure of a characteristic of the object that may be significant from the point of view of a person using data mining methods. There are however some problems. In case of having access to a relational database it is important to define appropriate attributes that can be extracted from database using SQL queries or more complex scripts. The number of chosen attributes might lead to the so-called „Curse of dimensionality", which might occur especially in the case of classification or regression task [12]. Another problem is the dependence between data that are hard to reproduce by the use of matrix and which might be necessary for the process of discovering knowledge. An example might be event sequences occurring in time, to which we can include:

- Aircraft incident;
- Flight;
- Inspection during a flight;
- Examination on medical board;
- Course of the conducted exercises in the process of training.

In the abovementioned case, both the order and the period between events might be important. Another problem are different values that might adopt certain data parameters. In the existing database are available the following types:

- numerical, e.g. the number of flight hours, visibility expressed in kilometers;
- time, e.g. the date of the event;
- category in the ordinal scale, e.g. the class of the incident or class of the pilot;
- category in the symbolic scale as a sex, pilot speciality, aircraft type;
- text, e.g. the description of the reason for the aircraft incident.

It is extremely important in the case of data mining tasks using the notion of distance or similarity between two data vectors. An example of such a task is a cluster analysis. By determining the distance it is indispensable to remember that certain parameters in data matrix might have different units, which might be solved by the standardization of values. Another problem seems to be the correlation between parameters that can be removed taking into account e.g. the Mahalanobis distance [12].

5. Conclusions

IS TURAWA is equipped with a wide range of data from the A/C operation process. The long-term use of the above system as a tool to maintain an acceptable level of flight safety allowed to collect enough data for the purposes of data mining algorithms. The above data may be used to formulate the task of data mining, the purpose of which will be the identification of the causes of the occurrence of aviation events or the evaluation of a level of the risk of identified threats. The analysis made it possible to discover database tables that do not have information relevant from the point of view of data mining.

6. References

[1] *Statistical Summary of Commercial Jet Airplane Accidents. Worldwide Operations 1959–2016*, Boeing, 2017.
[2] *Safety Report*, ICAO, 2015.
[3] Klich E., *Bezpieczeństwo lotów w transporcie lotniczym*, Wydawnictwo Naukowe Instytutu Technologii Eksploatacji, Radom 2010.
[4] *Instrukcja Bezpieczeństwa Lotów Lotnictwa Sił Zbrojnych RP*, Załącznik do decyzji Nr 67/MON MINISTRA OBRONY NARODOWEJ z dnia 9 marca 2015.
[5] Nanduri A., Sherry L., "Anomaly Detection in Aircraft Data using Recurrent Neural Networks (RNN)", in: *2016 Integrated Communications Navigation and Surveillance* (*ICNS*), Proceedings of Conference ICNS 2016, April 19–21, 2016, Herndon, Virginia USA, pp. 682–690, IEEE, 2016.
[6] Tanguy L., Tulechki N., Urieli A., Hermann E., Raynal C., "Natural language processing for aviation safety reports: From classification to interactive analysis", *Computers in Industry*, Vol. 78, 80–95 (2016).
[7] Sjöblom O., "Data Refining for Text Mining Process in Aviation Safety Data", in: *Software Services for e-Business and e-Society*, in series: IFIP AICT, 305, pp. 415–426, Springer, 2009.
[8] Milburn N.J., Dobbins L., Pounds J., Goldman S., *Mining for Information in Accident Data*, Final Report, Federal Aviation Administration Civil Aerospace Medical Institute Oklahoma City, 2006.
[9] Nazeri Z., *Application of Aviation Safety Data Mining Workbench at American Airlines. Proof-of-Concept Demonstration of Data and Text Mining*, MITRE, Virginia 2003.
[10] Demski T., *Czy leciał z nami pilot, czyli 'text mining' na przykładzie opisów wypadków lotniczych*, StatSoft Polska, 2006.
[11] Zieja M., "The analysis of computer systems dedicated for the assessment of flight safety", *Journal of KONBiN*, No. 4(7), 269–287 (2008).
[12] Hand D., Heikki M., Smyth P., *Eksploracja danych*, Wydawnictwa Naukowo--Techniczne, Warszawa 2005.

Analiza zasobów informacyjnych systemu informatycznego TURAWA w aspekcie metod eksploracji danych

B. SZAFRAŃSKI, M. ZIEJA, J. WÓJCIK, K. MURAWSKI

Artykuł poświęcony jest analizie danych pochodzących z systemu informatycznego TURAWA wspierającego zarządzanie bezpieczeństwem lotów w Siłach Powietrznych Rzeczypospolitej Polskiej. Siły Zbrojne dysponują systemem, który zbiera i przetwarza dane dotyczące całego personelu lotniczego, wszystkich wykonanych lotów oraz wszystkich statków powietrznych. Coraz większe możliwości w pozyskiwaniu danych oraz ciągły rozwój metod eksploracji pozwalają wydobyć nieznane dotąd informacje, które wraz z wnioskami otrzymanymi z analizy danych pozwolą w przyszłości na poprawę poziomu bezpieczeństwa lotów.

Słowa kluczowe: bezpieczeństwo lotów, eksploracja danych, TURAWA.