Data Mining in Promoting Aviation Safety Management

Olli Sjöblom

University of Turku, Turku School of Economics, Finland
oljusj@utu.fi

Abstract. Safety is a key strategic management concern for safety-critical industries and management needs new, more efficient tools and methods for more effective management routines. Effective methods are needed to identify and manage risks in both aviation and other safety-critical industries in order to improve safety. Analysing safety related records and learning from "touch and go" situations is one possible way of preventing hazardous conditions from occurring. The eventuality of an incident or an accident may markedly be reduced if the risks connected to it are efficiently diagnosed. With the aid of this outlook, flight safety has witnessed decades of successful improvement. This paper introduces aviation safety data analysis as an important application area for data mining. In this research text mining was utilised to study 1,240 flight safety reports testing three different systems, applying clustering to find similarities between reports, perhaps containing the indications of a lethal trend, without any presumption of their existence. All the different systems produced coherent results, proving that mining could extract information from unstructured data, which might not be possible with conventional methods.

Keywords: Management, Flight Safety, Data Mining, Text Mining, Analysis Method.

1 Introduction

Organisational decision making, especially in safety-critical systems, such as nuclear power and air traffic, is a complicated task. The top priority for the airline industry has always been the improvement of air safety [1]. Worldwide aviation is growing rapidly. Air traffic has generally been forecasted to grow 5 – 6% annually over the next two decades [2], or even over the next 10 – 15 years, the global air travel will probably double [3]. Consequently, the number of accidents will respectively increase if nothing were done to improve it, which development would, clearly, be unacceptable. This has been foreseen already at the shift of the millennium by the European Commission [4] that expressed the need to explore new and efficient ways in order to improve air safety.

The conventional safety tools and methods based on data collection have reached their peak performance because of their inability to create new knowledge. For further improvements new methods and tools, like data mining, are urgently needed [5]. The need for automated means to process data is increasing rapidly, because the amount of generated and stored unstructured data is increasing rapidly [6]. Usually, data accumulates faster than it can be processed [7]. To extract knowledge from the vast

K. Saranto et al. (Eds.): WIS 2014, CCIS 450, pp. 186–193, 2014.

amount of information and data that is now available, organisations search for the methods to making smarter decisions in order to achieve better results, are increasingly utilising the data assets as well as their advances in computational power with software combined with specialised analysts [8].

2 Management in Safety-Related Context

Kettunen et al. [9] emphasise the managerial challenges in the safety-critical industries, which are typically related to finding a balance between diverging demands and expectations, like economy- and safety-related objects without forgetting the priorities-setting and maintaining focus on these components. The key action is a continuous balancing between taking risks and allocating resources for risk management.

The significant development in database and software technologies, i.e. the warehousing of transaction data, has enabled the organisations to build a foundation for knowledge discovery in databases [10]. The unknown lethal factors brought into daylight could be eliminated; at least a significant part of them and a sufficient safety level could be reached with reduced investment allocation. For air traffic, there is theoretically no upper limit to allocate resources to safety in different forms, such as investing into hi-tech equipment and control systems on a redundant scale, investing in personnel training and developing directives and action rules to add safeguards. The relation between safety and cost efficiency could be illustrated explicitly comparing the costs between comprehensive maintenance programs and maintenance-induced accidents, the benefits that outweigh the accident costs [11].The process for allocating extra resources to special projects might become even more troublesome in case there are interdependencies among the projects [12]. Managing risk and safety has been problematic in air transport: very high levels of safety are too costly – high levels of risk are unacceptable. Therefore, safety reports have been collected through decades to investigate and assess risks and to define risk standards, which are consistent with the value systems of the society [13, 14].

3 Flight Safety

Air traffic is full of incidents and deviations that do not contain any hazard as such, but need to be investigated to find out potential lethal trends. These undesirable, but very minor events are valuable investigation subjects for risk and safety specialists to build an understanding about their causes and to detect unsafe trends. Investigation also reveals whether countermeasures are warranted and how to reduce or eliminate potential accidents [15].The appearance of similar recurring cases (a cluster) may indicate a hazardous trend that should be analysed very carefully to find out whether a real danger exists or not. The possibly existing lethal trends are trying to penetrate through the layers of defences, barriers and safeguards (Figure 1) that, fortunately, usually stop them from proceeding. Because serious incidents and even accidents do happen, it can be presumed that after a certain amount of time they pass all the layers but the last one; then they will pass the last layer as well, which leads to accidents. The latest studies on aviation suggest that text mining can be utilised to detect these trends [16], i.e. the chains of events that lead to accidents if intervention does not occur [17].

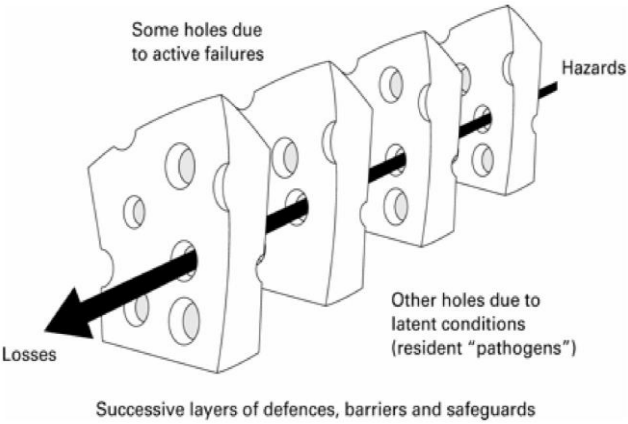

Fig. 1. The Swiss Cheese model [17, 18]

Searching for similar documents (clustering) is an essential mining function, able to reveal a recurring hazard that might lead to an accident. The clustering results have preliminarily proved its better performance compared with more traditional statistical methods [19]. These, often called "the nuggets of knowledge", are hidden in vast amounts of data and are practically undiscoverable with conventional techniques [20]. Using mining software, knowledge of data is combined by an analyst with advanced machine learning technologies to discover the relationships. Watson [20] also found that with conventional techniques it might take years to find meaningful relationships.

4 The Research Process

Finding trends from narrative data has required significant human involvement. Thus, the analysis process and its possible results rely on the skill, memory and experience of the safety officers [21]. Before data mining systems were developed, there were no tools for analysing textual data with computers. Data mining provides a worthy alternative in the selection of analysis methods in order to illustrate the safety indicators and to reveal undesired trends.

Clustering explores the data set and determines the structure of natural groupings without any preliminary assumptions. It is also directly applicable to Reason's Swiss Cheese model (see Figure 1). The literature in English gives several examples about using clustering in mining flight safety reports. The basic idea of cluster analysis is that all the texts within each cluster have a high similarity in content [22]. Using clustering as method, the main focus is the discovering and identifying of weak signals in the documentation.

The basis of all safety management is the systematic collection and analysis of operational data to identify and quantify potential risks [23]. The research data consisted of the narratives of 1,240 flight safety reports of deviation events from the years 1994-1996 in Finnish, provided by the Finnish Civil Aviation Authority (FCAA). The size of the narratives varied from a few words to a couple of sentences.

A three-year period containing more than 1,000 reports was considered creating a 'critical mass' for producing relevant and reliable mining results. As the material was more than 10 years old, it was guaranteed that the data was already statute-barred and there were no open cases [24].

Three different systems seemed to be appropriate for benchmarking. The author was aware of one prototype (GILTA), one commercial product (TEMIS) with a Finnish module prototype, and one commercial system (PolyVista) with encouraging results mining Spanish, which seemed worth testing in Finnish.

The structure of the research process is presented in Figure 2. The process contains several steps or phases that must be gone through to form knowledge from raw data. To be understandable the information must be presented with reports, graphs or in other suitable forms once found. Both information and knowledge can be refined from collected raw data using conventional tools, but in acquiring further value their limits will be reached. This is clearly seen in the goal of the whole process through which new knowledge can be synthesised from previously held knowledge.

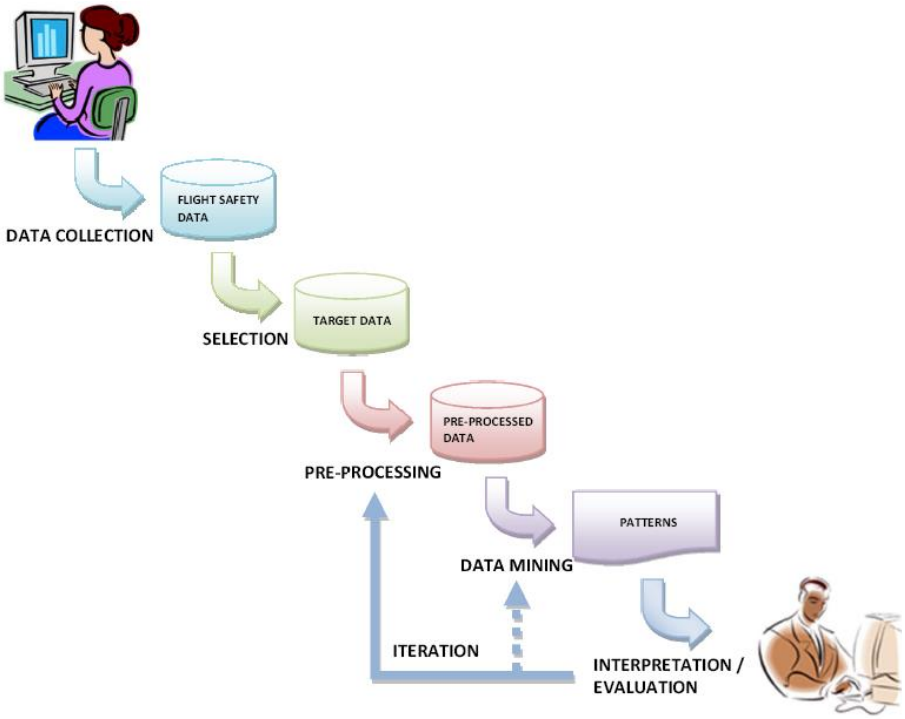

Fig. 2. The structure of the mining process

Mining is an iterative process although it makes no sense to increase the amount of rounds too much. The need for tuning, especially the definition of stop words and synonyms was discovered on the basis of the results of the first mining round. Some pure mistakes, like some common stop words and synonyms forgotten from the list,

were noticed. A more significant problem was the appearance of some frequently used "common" words (like 'plane' with its synonyms 'airplane' and 'aircraft') skewing the results. Their role in the data was carefully analysed [25], using a quantitative data analysis application called NVivo to get a deeper analysis.

In case there is a need to change the definitions of the data again, like in this study it appeared to be, the mining process proceeds from the interpretation/evaluation phase back to pre-processing as showed in Figure 2. As there was an obvious need to redefine the stop words and synonyms, the return was necessary. In the picture, a smaller 'loop' also exists, illustrated with a dash line. This path will be used in case there is no need to change the data itself, but for example when a big cluster ought to be re-clustered in order to receive smaller amounts of data to interpret and evaluate.

With structured data, the explanation of a case usually tells the truth to a certain extent, but completed with narrative data it can be close to 100%, at least theoretically. Mining combined with other methods will give significant contributions to the decision processes. Narrative text mining is demanding also because of the multiplicity of languages spoken in the world. Especially languages with small user groups, such as Finnish, have to wait for efficient tools being developed much longer than the major languages. The search technologies are challenged by inflected forms and compounds. In Finnish, for example, the words may have thousands of inflected forms and in addition to that, they can be parts of compounds in almost countless combinations [26]. For search technologies, English is an "easy" language.

The coherent clusters as the results of the second round were taken into more detailed inspection. The progress as the change of distribution can be recognised through the percentage of 'sense making'[1] clusters. Further, the average weight of the most important words of each cluster increased and the correspondent standard deviation diminished significantly. All these changes indicate the movement towards the aimed more homogenous clusters, thus more accurate information.

5 Results

Already the first round results looked promising. The smallest clusters began to produce some directly applicable information indicating that the sizes of the clusters play a significant role in the applicability of the results. This must, however, be scaled with the amount of production data. No preliminary definitions or limitations were made; the applied systems clustered the cases according to their basic determinations. Although the results of all the three systems were coherent, GILTA can be said to have produced the most accurate results clustering 1,240 flight safety reports into 100 clusters, their sizes varied between 58 and 1 report.

PolyVista processed the data determining the number of clusters first to be 6 and then raising it up to 20 in a second step, setting the score 100 for the most content describing word of the cluster and correspondent values to the others. When there were 20 clusters, the smallest of them contained 10 reports and the biggest 232, and in

[1] Clusters, from which information can be seen clearly as such.

that case, in eleven of them the scores of the three most important words were more than 50. In the last cluster containing 10 reports, the scores of the 10 most important words were 50 or more, which can be considered a good mining result.

Due to the Finnish module of TEMIS, no pre-processing was necessary. It created 26 clusters leaving out 8 reports as unclassified documents. The size of the clusters varied from 108 to 21 reports. The system was allowed by the operator to create sub clusters in case the size of the cluster exceeds 100 reports and therefore the first cluster was divided into two sub clusters with 58 and 50 documents. After the division, the biggest cluster included 78 reports.

In the target data, about 20 clusters could preliminarily be regarded as containing potentially lethal trends, e.g. a door opening during a flight. Others, like flying into Finnish airspace without air traffic control clearance and illegal smoking on board during the flight as well as gliding-related events can be mentioned. During the second data mining round, refining the definitions after the first one caused a fairly small but remarkable increase in the accuracy of the results. Narratives with a single word were excluded correctly from the clustering as an anomaly by the system. Despite their disparity, the contents of the clusters seemed to be very relevant and were used as material for a more accurate examination by human investigation to find out the existence of the potential hazard in similar recurring events. An additional detail is worth noting - all systems left out almost the same reports.

The testing process proved that data mining might be the only one for uncovering hidden information, supporting the premise that if lethal trends on the whole exist, it reveals important safety information from fast accumulating, vast amounts of data, not accessible with other methods, to be used as an essential factor for strategic safety management. Additionally, because it can find things or traits or tendencies we are not aware of, too, it is an essential tool for being used not only in strategic management but could also be used for allocating resources in safety management.

6 Discussion

Data mining does not give straight answers to the questions, but its role is purely a decision support system although that often provides, indispensable supplementary information for the decision making processes. Thus, the representation of the discovered patterns and the assessing of their value requires that they be consolidated with existing domain knowledge because their value or significance cannot be captured using mining tools. This is why the process requires human participants with vast experience of the subject.

To develop this study further, the other data fields left out in order to simplify the research process, in addition to the narratives only used in this study, might be taken into consideration in the data mining process – in order to gain more accurate results by increasing the coverage of the process.

Despite the fact that data mining has been available as an applicable method already since the 1960s, it is not as widely used as could be expected. Text mining was enabled later than the mining of numerical data being, however, on hand a couple of decades. There are, no doubt several explanations for this. One of the most

significant of them, concerning especially text mining, is that these tools are mostly language dependent, a fact which does not favour small language groups like Finnish that also happens to be a substantially complicated language from this point of view, notably requiring resources in order to develop functional tools. Another reason might be that the mining tools have not been as simple to use as the tools the authorities and other actors have been used to. In addition to this, successful data mining does not happen as facilely as the use of, for instance, Excel spreadsheet tools and functions but, as described more minutely before, requires a process including several steps or phases.

Although the accident ratio has diminished since the beginning of the 1960s steadily as well among general aviation as gliders and motor gliders, and there was even a period between 1996 and 2006 without lethal accidents among gliding and motor gliding, the situation is not satisfactory. The lethal accidents have returned to the aviation field having survived ten years totally without and the number of the accidents among ultralight aviation has increased remarkably. This for one's part is the reason why the Finnish Transport Minister ordered in April a wide mapping about the risks among leisure aviation to be made, to be completed at the end of September 2014. This unfavourable development evidently emphasises the significance of the need for sophisticated safety analysis methods and their development.

As mentioned, this study was made with no presumption of the existence of potential lethal trends. In case it is already known what will be looked for, business intelligence (BI) methods could be applicable. Although these systems allow the databases to be queried using numerous keywords to search for known cases of a certain type or their combinations, the results received from them are simpler to interpret compared with those of the mining tools. Additionally, business intelligence can also well be combined with data mining. In case through mining process something worth examining more minutely were found, this data could act as a query basis for BI tools that could pick up more accurate information on the type of cases found. When doing this way, thus applying BI tools in the second phase of the process, additional hazardous factors might be discovered in the data, guiding the safety specialist to those patterns that show where a potential accident could occur.

References

1. Liou, J.J.H., Yen, L., Tzeng, G.-H.: Building an effective safety management system for airlines. Journal of Air Transport Management 14(1), 20–26 (2008)
2. Netjasov, F., Janic, M.: A review of research on risk and safety modelling in civil aviation. Journal of Air Transport Management 14(4), 213–220 (2008)
3. Global Airline Industry Program. Analysis: The Airline Industry. Global Airline Industry Program [WWW-page] (2008), http://web.mit.edu/airlines/analysis/analysis_airline_industry.html [cited 2011 9.5.]
4. European Commission, Proposal for a Directive of the European Parliament and of the Council on occurrence reporting in civil aviation, Commission of the European Communities, Editor 2000, Brussels (2000)
5. Evans, B., Glendon, A.I., Creed, P.A.: Development and initial validation of an Aviation Safety Climate Scale. Journal of Safety Research 38(6), 675–682 (2007)

6. Delen, D., Crossland, M.D.: Seeding the survey and analysis of research literature with text mining. Expert Systems with Applications: An International Journal 34(3), 1707–1720 (2008)
7. Wang, X., Huang, S., Cao, L., Shi, D., Shu, P.: LSSVM with Fuzzy Pre-processing Model Based Aero Engine Data Mining Technology. In: Alhajj, R., Gao, H., Li, X., Li, J., Zaïane, O.R. (eds.) ADMA 2007. LNCS (LNAI), vol. 4632, pp. 100–109. Springer, Heidelberg (2007)
8. Cleary, D.: Predictive Analytics in the Public Sector: Using Data Mining to Assist Better Target Selection for Audit. Electronic Journal of e-Government 9(2) (2011)
9. Kettunen, J., Reiman, T., Wahlström, B.: Safety management challenges and tensions in the European nuclear power industry. Scandinavian Journal of Management 23(4), 424–444 (2007)
10. Blake, M.B., et al.: A Component-Based Data Management and Knowledge Discovery Framework for Aviation Studies. International Journal of Technology and Web Engineering 1(1) (2006)
11. Castro, R.: A Holistic Approach to Aviation Safety. In: Flight Safety Digest, pp. 1–12 (1988)
12. Kirkwood, C.W.: Strategic Decision Making. Wadsworth Publishing Company, Belmont (1997)
13. Janic, M.: An assessment of risk and safety in civil aviation. Journal of Air Transport Management 6(1), 43–50 (2000)
14. Sage, A.P., White, E.B.: Methodologies for Risk and Hazard Assessment: A Survey and Status Report. IEEE Transaction on Systems, Man, and Cybernetics SMC-10(8), 425–446 (1980)
15. Kirwan, B.: Incident reduction and risk migration. Safety Science 49(1), 11–20 (2011)
16. Sjöblom, O.: Data Mining in Aviation Safety. In: 5th International Workshop on Security. Information Processing Society of Japan, Kobe (2010)
17. Reason, J.T.: Human error: models and management. British Medical Journal 320(7237), 768–770 (2000)
18. Reason, J.T.: Managing the Risks of Organizational Accidents, 252p. Ashgate Publishing Limited, Aldershot (1997)
19. Saracoglu, R., Tütünkü, K., Allahverdi, N.: A new approach on search for similar documents with multiple categories using fuzzy clustering. Expert Systems with Applications: An International Journal 34(4), 2545–2554 (2008)
20. Watson, R.T.: Data Management: Databases and Organizations, 2nd edn. John Wiley & Sons (1999)
21. Nazeri, Z.: Application of Aviation Safety Data Mining Workbench at American Airlines. Proof-of-Concept Demonstration of Data and Text Mining, Center for Advanced Aviation Systems Development, MITRE Corporation Inc., McLean, Virginia, US (2003)
22. Rosell, M.: Text Clustering Exploration. Swedish Text Representation and Clustering Results Unraveled. School of Computer Science and Communication, p. 71. Kungliga Tekniska Högskolan, Stockholm (2009)
23. GAIN Working Group B, Role of Analytical Tools in Airline Flight Safety Management Systems, W.G.B.A.M.a. Tools, Global Aviation Information Network (2004)
24. Sjöblom, O., Suomi, R.: Data Mining in Aviation Safety Data Analysis. In: Rahman, H. (ed.) Social and Political Implications of Data Mining: Knowledge Management in E-Government, p. 349. Information Science Reference, Hershey (2009)
25. Lindén, K.: Word Sense Discovery and Disambiguation. In: General Linguistics, p. 191. University of Helsinki, Helsinki (2005)
26. Karlsson, F.: Yleinen kielitiede. Yliopistopaino, Helsinki (1994)