Application of Text Analysis Technology in Aviation Safety Information Analysis

Junjie Liu, Huijuan Yan* and Yinlan Du
College of Flying Technology, Civil Aviation University of China, Tianjin, China

*Corresponding author email: 13920588569@163.com

Abstract. This research is aimed to obtain potential safety hazards quickly, accurately, and efficiently from a large amount of daily collected aviation safety information, and provide a clear improvement direction for safety risk control. By combing with the text analysis and machine learning, clustered a given type of aviation safety information according to its content is an important basis of mining information effectively. Taken the system failure/jam/fault events collected by Chinese civil aviation in 2017 as a sample. By means of text pre-processing, feature extraction by using logarithmic IF-IDF and k-means method under the environment of python3.6, an automatic clustering model of the sample information is established, and then the visualized results are output based on Multi Dimensional Scale (MDS). The analysis results show that text clustering and visualization can quickly and automatically file information, identify the similarity among sample information, easily lock key information, and provide targeted measures for the next step of risk management and control.

Keywords: Aviation safety information; Text analysis; Text clustering; Cluster visualization.

1. Introduction

In recent years, with the development of civil aviation industry in China, there are more and more channels to obtain aviation safety information, which has led to an increasing number of aviation safety information. Aviation safety information is very professional, so in-depth analysis requires strong industry knowledge. However, the lack of analysis methods and the shortage of personnel keep the analysis of aviation safety information at the stage of statistical analysis. Therefore, it is urgent to develop reliable and professional analysis tools.

Faced with a large amount of aviation safety information, some methods [1-3] can start from a specific event or event type, then analyse the cause of the event in detail, which is not able to obtain effective analysis results in a limited time. Therefore, domestic and foreign scholars have begun to use computers as tools for human thinking, mainly clustering algorithms, association rules and classification algorithms are used to automatically analyse aviation safety information. For example, the references [4-5] are based on the idea of clustering, by using neural network algorithms and latent semantic indexing methods to achieve the effective classification of airport noise annoyance model fuzzy sets and aviation report; Some scholars[6] used support vector machine algorithm to cluster human factors involved in aviation safety reports, and had achieved positive results; Ludovic Tanguy, Nikola Tulecki[7] applied Natural Language Processing (NLP) tools to aviation safety report management, they used supervised machine learning technology to find the cause of a single event and identify reports similar to the specified content. As a result, they have realized data / text mining and active analysis of interactive information; SD Robinson[8] used the Latent Semantic Analysis (LSA) process to identify correlations between corpora and proposed a computational method that allows the combination of macro data views and micro data views to successfully visualize the factors between...
corpora relationship; Rosa Maria Amalio Valdes, V. Fernando Gomez Comendador[9] used Bayesian machine learning methods and hierarchical structures to derive statistical estimates and prediction models with different complexity and goals to identify aviation safety risks in advance; Liu[10] classified aviation safety reports based on convolutional neural network algorithm to predict aviation safety risks.

The above analysis found that although the data mining technology [11] based on the traditional event analysis method had preliminary applications in aviation safety information analysis, most of them were based on association rules to find potential information relationships. There were few researches on using text clustering and visualization techniques to quickly extract key content of information. Therefore, based on the existing classification of aviation safety information, this paper used python as a tool, completed information clustering by using K-means and realized visualization based on MDS and word cloud diagram to discover hidden safety risks in aviation safety information. This way provided data-based improvement directions for risk management.

2. Text Analysis

Text analysis involves text classification, text clustering, and cluster visualization. Among them, cluster analysis mainly summarizes the relationships between data objects according to given requirements and rules from the data set [12], and completes the classification process. This paper is used key techniques of text analysis by python to automatically analyse information, including text normalization processing, cluster analysis, and visualization processes.

2.1. Text Normalization Processing Technology

2.1.1. Chinese word segmentation. In order to realize the feature input in numerical format, the text must be cleaned and standardized preprocessing. The Chinese text segmentation [13] is the first step in the normalization of text, which breaks down text data into smaller and meaningful components. This paper used rule-based jieba exact mode for word segmentation. Although jieba has the ability to recognize new words, a custom dictionary can prevent words such as "low oil level warning" from being forcibly separated to ensure segmentation accuracy.

2.1.2. Moved stop words list. Text data is unstructured and has various noises, such as prepositions, punctuation, adverbs, etc. If noise is not removed, it may affect the results of subsequent cluster analysis. Therefore, after the word segmentation, load and traverse the stop words list, clean the text words, and add or delete words in the stop words list as required to minimize text noise.

2.1.3. Feature extraction. The meaning of feature extraction is to use text analysis technology to screen out the content that can best express text information. Vector Space Model [14] is a very useful concept and model for processing text data. Mathematically, it is assumed that there is a text D in the text vector space VS. Each text dimension and number of columns will be the total number of different terms or words in all text in the vector space, therefore, the vector space can be expressed as:

\[ VS = \{W_1, W_2, \ldots, W_n\} \]  (1)

n is the number of different words in the entire text. Therefore, we can represent the text D in vector space as:

\[ D = \{w_{D1}, w_{D2}, \ldots, w_{Dn}\} \]  (2)

\( w_{Dn} \) represents the weight of the nth word in text D. This weight is a quantity value, which can be the frequency of words in the text, the average frequency, or the TF-IDF weight.

TF-TDF [15] still has certain advantages in the weight calculation method, the mathematical definition of calculating tf-idf is as follows:

\[ tfidf = tf \times idf \]  (3)

tf represents term frequency, the word frequency of any text is how often the word appears in a
particular text; idf represents inverse document frequency, its mathematical expression is:

$$idf(t) = 1 + \log \frac{C}{1 + df(t)}$$  \hspace{1cm} (4)$$

$idf(t)$ represents the idf of the word $t$, $C$ represents the total amount of text in the database, and $df(t)$ represents the frequency of the number of texts containing the word $t$. It can be obtained from the value of $idf(t)$ that frequent occurrences of words in the database are not necessarily valuable. It may be useful information that there are few occurrences in the database and more occurrences in specific texts.

The final TF-IDF metric used in this paper is tf-idf normalization. We also need to divide the tf-idf matrix by the L2 norm of the matrix for normalization. The L2 norm is also known as the Euclidean norm, which is the square root of the sum of the squared weights of tf-idf for each word, then we got:

$$TF-IDF = \frac{tfidf}{\|tfidf\|}$$  \hspace{1cm} (5)$$

2.2. Formatting Author Names

2.2.1. Feature extraction. K-means clustering algorithm [16] has the advantages of simple and scalable large amount of data, and is therefore widely used, which is a centroid-based clustering model. The basic idea of the algorithm is: Suppose there is a data set $X$ with $N$ data points or samples, we separate $N$ data points into $K$ separate clusters $C_k$ which do not want to intersect. And each cluster can be represented by the average values of the sample clusters. These averages are the centroids $\mu_k$ of the clusters. They are not restricted by the actual data points of the $N$ samples of $X$. The algorithm selects these centroids and constructs a clustering model by inertia optimization, and the mathematical expression is:

$$\min \sum_{i=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$  \hspace{1cm} (6)$$

In clustering and centroid, the optimization steps are as follows:

Step 1: Select $K$ random samples in dataset $X$ and select $K$ initial centroids;

Step 2: Update the cluster by assigning each data point or sample to its nearest centroid. The mathematical expression can be:

$$C_k = \left\{ x_n : \|x_n - \mu_k\| \leq \forall l \|x_n - \mu_l\| \right\}$$  \hspace{1cm} (7)$$

Step 3: Obtain new cluster data points for each cluster through Step 2 and recalculate to obtain new clusters, that is:

$$\mu_k = \frac{1}{C_k} \sum_{x_n \in C_k} x_n$$  \hspace{1cm} (8)$$

The above steps are repeated iteratively until Step 2 and Step 3 do not change, and finally the clustering effect is better.

2.2.2. Cluster visualization. In order for the clustered information to be displayed intuitively and visually in low-dimensional coordinates, to grasp the distribution of similar events as a whole, and to realize the function of visually classifying the research objects, so the information needs to be visualized. Aviation safety information is unstructured, so it is best to use Multiple Dimensional Scaling (MDS) to reduce non-linear dimensions. The core idea of MDS is to use a distance matrix to obtain the distance between data points, such as cosine similarity [17], Euclidean distance [18]. MDS
attempts to use the high-dimensional features in the vector to convert to low-dimensional feature representations, and uses the matplotlib tool to draw graphics.

3. Sample Analysis
In this paper, the information of domestic unsafe incidents in 2017 is selected, because of its strong objectivity of mechanical failure, and the reported information content is more true and reliable. Therefore, taken the system failure/jam/fault event type information with the largest number of mechanical reasons in 2017 as the sample, then 931 pieces of sample information is sorted out and extracted text features. Finally, quickly realizes information clustering and visualization.

3.1. Information Preprocessing
Based on the characteristics of the research object, the dictionary of civil aviation terms is loaded, adopted jieba word segmentation in Python library, and the stop word list is loaded to delete invalid words. We ended up with 3757 words. Some terms are shown in table 1. Features represents feature terms, Index represents the position of feature terms, and table 1 also shows the type and character size of feature words. It is used tf-idf to calculate the weight values of each feature word, table 2 shows the weight values of "weather radar" in different information.

| Index | Type | Size | Features      |
|-------|------|------|---------------|
| 123   | str  | 1    | Right engine  |
| 281   | str  | 1    | Oxygen mask   |
| 306   | str  | 1    | Flare         |
| 564   | str  | 1    | Weather Radar |
| 637   | str  | 1    | pushback      |
| 893   | str  | 1    | Landing Gear  |
| ...   | ...  | ...  | ...           |

| Feature terms | Information title | Weights |
|---------------|-------------------|---------|
| 309           | 0.160             |
| 436           | 0.144             |
| 459           | 0.371             |
| 487           | 0.140             |
| 514           | 0.341             |
| Weather Radar |                   |
| 459           | 0.371             |
| 487           | 0.140             |
| 514           | 0.341             |
| ...           | ...               |

3.2. Cluster Analysis
The core idea of cluster analysis is to input the feature matrix and divide into groups based on the distance between the information. Import K-Means from the sklearn.cluster model in the Python library, set the parameter num_cluster to 13, that is, the clustering label is 13, and the clustering results after running are shown in Table3.
Table 3. Cluster results of system failure/jam/fault

| Clustering label | Key_feature                                      | Clustering number | Retrieved information               |
|------------------|--------------------------------------------------|-------------------|-------------------------------------|
| Cluster 0        | RVSM, Run, TCAS, Skin exit valve, Cockpit height | 117               | 3, 17, 32, 34, 44, 49…              |
|                  | Flaps, Position,                                 |                   |                                     |
| Cluster 1        | Approach, Trailing edge, Sensor                  | 86                | 7, 23, 25, 29, 79, 105…             |
| Cluster 2        | Rescue, Emergency, 09, Response, Level           | 29                | 12, 27, 39, 72, 89…                 |
| …                | …                                                | …                 | …                                   |

It can be known from Table 3 that each cluster label is assigned some information to quickly organize and archive. According to the key feature of each label to understand key content, and retrieve detailed information. For example, cluster1 contains 86 pieces of information, and the extracted detailed cluster analysis information is "flaps", "approach", "position", "backing edge", "sensor", etc. Refer to the key features in the cluster label, It can be timely and quickly grasped: In the system failure / jam / fault event type information, the aircraft is prone to cause asymmetrical flaps or jam events during the approach phase, and flap position sensing needs to be replaced after the flight. Therefore, according to the key features of the clustering label, the safety management or maintenance personnel should pay attention to check the flap position sensors, find out the method of eliminating safety risks in advance, and lay down preventive measures in time.

3.3. Cluster Visualization

Because aviation safety information describes events in the form of text, and the cosine distance is more likely to distinguish the similarity or difference between texts based on the content of text, the cosine distance is used as the distance matrix for cluster visualization. Visualization results based on MDS by using python are shown in Figure 1.

Figure 1. Clustering visualization of sample information

In the result shown in Figure 1, there are 13 cluster labels in total, and each cluster label has its own colour and symbol to distinguish the feature values in the legend box. Each cluster is described by its topic. The topic is used to define the cluster, in order to quickly retrieve each piece of information. For example, in the visualization results of clustering based on MDS, the information related to the "TCAS" fault is displayed as a red circular cluster label, and the information related to the "landing gear" fault is shown as a yellow pentagonal cluster label. Figure 1 reflects that the information with inherent relative relationship is linked together through visualization, and the similar features of the sample information are displayed in the form of graphics, which is more intuitive in effect and achieves the visual classification function of aviation safety information.

In order to fully understand the function of cluster visualization and its application in practice, detailed analysis will be performed on the information in a certain clustering label to clearly sort out the differences between the information. In the python environment, the cluster visualization diagram drawn using matplotlib can enlarge the target area, which makes clearly see the specific information contained in the clustering labels and other information close to it. For example, the clustering label
with the key feature of “TCAS” is enlarged, that is, the red circular area shown in the Figure 1 is enlarged, and the result is shown in Figure2.

Figure 2. Local visualization of clustering label related to "TCAS"

Figure 2 not only shows the frequency of failure of a specific system component of the aircraft, but also can mine the correlation between each other based on the distance between information. The same cluster label is mostly related to the same system fault, but the attributes of the same system fault are different. As the information shown in the Figure.2, the TCAS fault is mainly described, but it may be caused by problems such as the TCAS computer, antenna or jumping switch. Therefore, according to the coordinate position of the information in the low-dimensional space, visually discern the gaps between the safety information, retrieve and trace the more similar information, and comprehensively find the fault attributes of different system components involved in similar systems. The risk identification of multi-level faults in a certain system plays a supporting role.

Then using the micro word cloud online analysis tool to visualize the cluster label information, which forms a keyword cloud or keyword rendering, so as to filter more invalid words. Finally, the key theme of the information can be scanned at a glance during the information analysis processing. The main features of the information content in the red circular clustering label in Figure. 2 are shown above. The word cloud diagram is shown in Figure 3.

Figure 3. Word cloud of clustering label related to “TCAS”

It can be seen from Figure3 that words such as “TCAS”, “RVSM”, and “cruise” are relatively large and prominently located, which indicates that the words in the information set are associated with many cases and are of great importance. The visualization method can grasp the key theme of the cluster label quickly and easily, and it is easy to help the analyst or manager to lock the key items and improve the work efficiency of the information analysis.

There are some words related to the stage of flight and region, such as "taxi", "cruise", "Chongqing", "Shenzhen" and other words, which shows that TCAS failures may be caused by a combination of multiple factors, the description of the system fault information is not specific enough, and the in-depth analysis of the system failure / fault is limited, and further detailed event report information needs to be obtained.
4. Conclusion
1) In the python environment, starting from the content of the information, the text analysis technology is used to process the sample information to realize the automatic identification and clustering of aviation safety information in a timely and fast manner, so as to reduce the time for manual review, sorting, and archiving of information.
2) Clustering and visualization of aviation safety information based on MDS, can grasp the distribution feature terms of information from the overall visual, to realize the function of visual classification of information, and retrieve and trace fault information similar to the same system or the same component faults, then find the multiple failure with the same system to master possible factors for a certain system failure, which offers a tool of the systematic hierarchical management of aviation safety risk.
3) Word cloud diagrams is used to visualize the information content in specific cluster label, it is easy to lock key information items, and discover multiple factors that may be related to system fault. Word cloud visualization can provide reference for the next step in developing targeted measures for risk management, so as to achieve the goal of early warning of security risks.

References
[1] Han Bing. Analysis of Human Factors in ATC Communication based on Shell model[J]. Online game world cloud education, 2013 (22): 24-24.
[2] Lve Chunyu, Fang Chunhua. Human Factor Analysis and Classification System (HFACS) and Accident Case Analysis [J] Journal of civil aviation flight college of China, 2009, 20(2):37-40.
[3] Liu Hanhui, Sun Ruishan, Zhang Xuishan. A Method of Basic Element Analysis of Aviation Accidents [J]. Journal of Civil Aviation University of China, 1997(3):1-9.
[4] Feng Xia, Zhang Congying, Lu Min. Learning Method of Airport Noise Annoyance Model Based on Fuzzy Neural Network[J]. Computer Engineering and Design, 2015(12):3396-3401.
[5] Feng Xia, Yan Guannan, Li Juanjuan. Research on Spectral Clustering Based on Latent Semantic Indexing[J]. Journal of Civil Aviation University of China, 2011, 29(3):47-51.
[6] Shi, Donghui, Guan, Jian, Zurada, Jozef, etc.. A Data-Mining Approach to Identification of Risk Factors in Safety Management Systems[J]. Journal of Management Information Systems, 2017, 34(4):1054-1081.
[7] Tanguy L, Tulechki N, Urieli A, et al. Natural language processing for aviation safety reports: From classification to interactive analysis[J]. Computers in Industry, 2016, 78:80-95.
[8] Robinson, S. D. Visual representation of safety narratives[J]. Safety Science, 2016, 88:123-128.
[9] Rosa Maria Arnaldo Valdes, V. Fernando Gomez Comendador, Luis Perez Sanz, Alvaro Rodriguez Sanz."Prediction of aircraft safety incidents using Bayesian inference and hierarchical structures." Safety Science 104(2018):216-230.
[10] Liu Mengna. Analysis of influencing factors and risk prediction of aviation safety report based on text mining [D]. Anhui University of Architecture, 2019
[11] Han Yan. The Method of Data Mining and Knowledge Discovery[J], 2007, 2(8):227-228.
[12] Wang Xiao. Research on uncertainty data clustering algorithm based on fuzzy sets [D]. Harbin university of science and technology.March.2018.
[13] Long Shuquan, Zhao Zhengwen, Tanghua. Overview on Chinese Segmentation Algorithm, 2009, 5(10):2605-2607.
[14] Lu Yuchang, Lu Mingyu, Li Fan, et al. Analysis and Construction of Word Weighting Function in VSM[J]. Journal of Computer Research and Development. 2002, 39(10):1205-1210.
[15] ZHANG W, YOSHIDA T, TANG X. A comparative study of TF*IDFF, LSI and multi-words for text classification[J]. Expert systems with applications, 2011, 38(3):2758-2765
[16] Xiong Kangping. Application of machine learning method in text clustering[J]. Electronic world, 2018, 556(22):43-45.
[17] Wu Yongliang, Zhao Shuliang, Li Changjing, et al. Text Classification Method Based on TF-IDF and Cosine Similarity [J]. Journal of Chinese Information Processing, 2017, 31(5):138-145.
[18] Using Eclidean distance in the comparative analysis of taxonomic abundance Joseph E.Beaver Journal of Archaeological Science: Reports 2019-Elsevierjournal 10.1016/j.jasrep.2019.03.022