An Automated Recognition Model for Sensitive Information

Dongdan Guo, Hao Sun, Tao Zhu and Chang Cai

Travelsky Technology Limited, Beijing; Kunming Airlines Co., Ltd., Changshui International Airport, Kunming, Yunnan Province; civil aviation university of china, Tianjin
ddguo@travelsky.com; sunhao@travelsky.com; zhutao@shenzhenair.com;
18822129196@163.com;

Abstract. Civil aviation business system carries a lot of passenger information. How to strengthen the management and use of civil aviation passenger information has become an important issue for the civil aviation industry. For the sensitive passenger information with huge data type and quantity, it is extremely inefficient to comb and analyze it manually alone. In this paper, we establish an automatic sensitive information identification model to effectively identify the passenger sensitive information in structured and unstructured files, improve the identification efficiency of sensitive information in civil aviation business system and the ability of data security management.

1. Introduction
In order to build smart civil aviation, improve the operation efficiency of airlines and provide sincere service for passengers, civil aviation enterprises and institutions have vigorously promoted the construction of information technology in recent years. The increasing variety and quantity of information systems and the increasingly advanced technologies applied have also led to changes in the types of data stored. For example, the application of big data technology lead to the emergence of unstructured data types in previously structured data. Many of these business systems built by civil aviation enterprises and institutions carry passenger information. With the compliance requirements of laws and regulations related to network security and the development needs of civil aviation enterprises and institutions, how to strengthen the management and use of civil aviation passenger information has become an important issue facing the civil aviation industry. However, there are many systems in a civil aviation enterprise and institution, and the data types are not the same. Relying solely on manual sorting and analysis of data, the efficiency is quiet low, like this [1]. In this paper, an automatic sensitive information identification model is established to improve the identification efficiency of the sensitive information in civil aviation enterprises and institutions. At the same time, the adoption of automatic technical measures can effectively reduce the problem of information leakage caused by human participation and improve the data security management ability of civil aviation enterprises and institutions.

2. Grading and Classification of Civil Aviation Data
This paper mainly discussed the information of civil aviation passenger. Through analyzing the passenger information carried by the information system of civil aviation enterprises and institutions, the data fields involved in various information systems mainly include the following contents:

ID number, phone number, instant messaging account, email, home address, bank card number, name, flight number, e-ticket number, flight time, origin, destination, etc.

The sensitivity of information in different fields also varies. Referring to the Supreme People’s
Court, the Supreme People's Procuratorate on handle the illegal use of information network, help
information network criminal activities such as criminal cases to explain some issues of applicable law,
the GDPR, the personal information security information security technology standard (GB/T
35273-2017), such as regulatory standards, like this[2], and the passenger information in the black
chain illegal sales situation, level of civil aviation passenger information can be divided into the
following:
- Very sensitive information:
  ID number, phone number, name and e-ticket number.
  After getting the above information, the illegal elements not only can implement the air ticket fraud,
  but also may cause other aspects of the passenger trouble.
- More sensitive information:
  Home address, bank card number, instant messaging account, email.
  If the criminals use this information, they could carry out human search on the passengers.
- General information:
  Flight number, flight time, origin and destination.
  If criminals use this information, they can count the number of passengers and so on.
  For multiple records, different combinations of fields have different importance, like this[3]. For
  example, ID number + phone number + name + bank card number + e-ticket number + home address.
  These fields are the core content of the passenger information, and their combination would accurately
  locate the passenger, resulting in the serious disclosure of sensitive passenger information.
  Sensitive fields can be graded according to the sensitivity of the combined content of the field
  information. Sensitivity is also assessed based on the number of core fields.

Definition 1: let \( L = \alpha(l_1) + \beta(l_2) + \gamma(l_3) \) (1)

If \( L \) can represent the sensitivity of the combined information according to the weights of different
types of sensitive information, \( L \) is said to be the sensitivity of the combined information. Where,
\( l_1, l_2, l_3 \) respectively refer to the amount of very sensitive information, more sensitive information
and general information in a complete record.

3. Main characteristics of passenger information

Through analyzing the passenger’s information carried by the information system of civil aviation
enterprises and institutions, the fields information of civil aviation passenger information has the
following characteristics:
- Some field information meets strict standard requirements: such as length, character type,
  format, etc. These fields can be marked as the first category. Such as ID number, phone
  number, e-ticket number, flight number, bank card number, origin, destination and so on. In
  particular, id card, bank card number and other field information to meet their own
  verification requirements.
- Some of the field information contains instant messaging tool name, email, time, and other
  keywords; this type of field is marked as the second category. For example, the information
  before or after the field’s name contains keywords such as QQ, WeChat, Email, time, etc.
- Some of the field information appears as text, and this type of field is marked as a third type.
  Such as name, home address, etc.

4. Automatic Recognition Model of Sensitive Information

For structured data, the sensitivity level of the data can usually be determined by the field name, like
this [4][5]. However, different developers have different definitions of fields, and it is impossible to
judge the data content even by the header of the field. In the big data environment, information
systems of civil aviation enterprises and institutions are numerous, and the database tables involved
are more complex. If only through the manual data sensitive degree of recognition and judgment, on
the one hand, the work efficiency is low, on the other hand, it may cause sensitive information leakage
secondary leakage. It is more difficult to comb through unstructured data manually. In order to
improve the efficiency of data combing, reduce the risk of data leakage and improve the ability of data
management, this paper proposes an automatic identification model of sensitive information, with the specific contents as follows.

**Figure 1. Automatic identification model of sensitive information**

The sensitive information automatic identification model is suitable for structured data and unstructured data, and the specific processes are as follows:

1. **Input files**: including structured data and unstructured data.
2. **Feature recognition**: firstly, the contents of the file are classified and recognized according to features in this module.

   For the first type of field information: comb out the first type of field using regular expressions and length, character types, formats of matching data. The content sorted out is the type of field (such as: id card, phone number, electronic ticket number, flight number, bank card number, origin and destination number, etc.). In order to ensure the accuracy of data recognition, the field information such as id card and bank card number is verified.

   For the second type of field information: (1) use keyword matching to identify the header of structured data. Through the experiment, it is found that the recognition effect is poor due to the non-standard problem defined by the developer. (2) Keyword + pattern matching is used for recognition. Identify keywords in the file, such as QQ, WeChat, email, and other keywords, and then carry out pattern matching in the keyword attachment to identify whether there is sensitive information.

   For the third type of field information: this paper used the BiLSTM-CRF model for NLP recognition. BiLSTM-CRF model can identify Chinese people, place names, entity names and time. The recognition rate was 81%.

3. **Field recognition**: At this stage, annotate the sensitivity level of the fields’ type, which were identified in the previous stage (such as: ID card, phone number, instant messaging account number, email account number, home address, bank card number, name, flight number, e-ticket number, flight time, departure place, Destination, etc.). And annotate the number of each field.

   Through the above three steps, the recognition of sensitive fields in the file is realized.

5. **Identification of document sensitivity**

By automating the identification model of sensitive information, sensitive fields in files or databases can be quickly identified. At the same time, the sensitivity of the entire file or database also needs to be determined in order to determine the priority protected objects. This paper refers to the interpretation of the supreme people's court and the supreme people's procuratorate on several issues concerning the application of laws in criminal cases such as the handling of illegal use of information.
networks and the assistance of criminal activities of information networks. This paper refers to the data requirements of sensitive information in "Interpretation of the supreme people's court and the supreme people's procuratorate on several issues concerning the application of law in handling criminal cases such as illegal use of information networks and assisting criminal activities of information networks" to identify the sensitivity of files or databases:

1. Causing the disclosure of more than 500 pieces of track information, communication content, credit information and property information;
2. Cause the disclosure of accommodation information, communication records, health and physiological information, transaction information and other user information that may affect the safety of person and property more than 5,000 pieces;
3. Causing the disclosure of more than 50,000 pieces of user information other than those stipulated in the first and second paragraphs.

For structured data, count entries in civil aviation records after identifying sensitive fields. According to the calculation formula of sensitive data, the sensitivity degree of the whole database or file can be known.

\[ LR = LN = \{\alpha(l_1) + \beta(l_2) + \gamma(l_3)\}N \] (2)

Where, LR is denoted as the sensitivity of the file or database, and N is the number of entries recorded in the database.

For unstructured data, the number of entries in complete records cannot be known through simple statistics. This paper adopted the following method to determine.

Set \( R_1 = \{l_{11} \cdots l_{1i}\} \), \( R_2 \) represents the set of the number of very sensitive information fields, and \( l_{1i} \) represents the number of the i field of very sensitive information in unstructured data.

Set \( R_2 = \{l_{21} \cdots l_{2i}\} \), \( R_2 \) represents the set of the number of very sensitive information fields, and \( l_{2i} \) represents the number of the i field of very sensitive information in unstructured data.

Set \( R_3 = \{l_{31} \cdots l_{3i}\} \), \( R_3 \) represents the set of the number of very sensitive information fields, and \( l_{3i} \) represents the number of the i field of very sensitive information in unstructured data.

Then for the unstructured file sensitivity determination method is:

\[ LR = \alpha(\sum l_1) + \beta(\sum l_j) + \gamma(\sum l_k) \] (3)

Where, \( l_1, l_j, l_k \) respectively represent the number of different fields of each sensitivity level.

6. Conclusion
In this paper, an automatic recognition model of sensitive information was proposed and a sensitivity recognition method for structured and unstructured files was studied. The relevant work of this paper can help civil aviation enterprises and institutions, especially the information system is numerous, the amount of information is huge units. It would play a positive role in improving the efficiency of data sorting, reducing the risk of data leakage, and improving the ability of data management. It can also be used as an important method of data safety supervision of civil aviation. However, there were some problems in the sensitivity recognition of data files or databases in this paper. This paper simply carried out statistical analysis on a single field, without strictly considering the correlation between fields and data derivation. When processing unstructured processing files, complete records were not identified like them in structured data, and it was impossible to determine whether the fields were discrete or interrelated. Aiming at the above problems, this paper found through experiments and demonstration that in the case of massive data, the above problems had little impact on the recognition results. However, the file recognition effect for a small amount of data or records need to be further improved.

References
[1] Neerbek J., Assent I., Dolog P., Detecting Complex Sensitive Information via Phrase Structure in Recursive Neural Networks[C]// Pacific-Asia Conference on Knowledge Discovery & Data Mining. Springer, Cham, 2018.
[2] Luca Marelli, Giuseppe Testa. Scrutinizing the EU General Data Protection Regulation[J]. Science, 2018, 360(6388):496-498.

[3] Wang X, Wang D G, Ji-Yuan M U, et al. Research and Implementation of Data Masking System Based on Machine Learning[J]. Electric Power Information & Communication Technology, 2018.

[4] Wang R, Zhu Y, Chen T S, et al. Privacy-Preserving Algorithms for Multiple Sensitive Attributes Satisfying t-Closeness[J]. Journal of Computer Science & Technology, 2018, 33(6):1231-1242.

[5] Keane T. M, Naughton T. J. DSEARCH: sensitive database searching using distributed computing[J]. Bioinformatics, 2004(8):8.