Sensitive Data Discovery Technology Based on Artificial Intelligence

Jie SHI [a], ShuFang Cui *[b], Fei Chen [b], ChengTao Wang [b]

* Corresponding author: 228605126@qq.com

[a]China Todacco SiChuan Industrial LLC, SiChuan ChengDu, 610000

[b]WuHan Windoor Information Technology Co., LTD, HuBei WuHan, 430000

ABSTRACT: Enterprises will produce a large number of sensitive data with great value in production activities and management. If handled improperly, it will lead to data security incidents. Effective discovery and desensitization of these sensitive data is the premise of data security sharing. This paper refers to the relevant theoretical research and technical practice of sensitive data discovery at home and abroad, studies the discovery methods of sensitive data according to the application scenarios of manufacturing enterprises, and studies the sensitive data discovery method based on artificial intelligence for manufacturing enterprises which have frequently updated and unfixed sensitive data. Using the artificial intelligence technology to improve the sensitive data discovery, a sensitive data discovery method based on CRF-BiLSTM-CNN model is proposed. The experiment shows that it can provide a new solution for the current and future sensitive data discovery of manufacturing enterprises and provide support for data security governance and digital transformation of manufacturing enterprises.

Keywords: Data Sharing; Sensitive Data; Sensitive Data Discovery

1. INTRODUCTION

Companies produce large amounts of data during their production and operation, which contains a large amount of confidential and sensitive data, which is a digital asset and is of great value[1].It can be used for scenarios such as sharing exchange and data mining, and if improperly disposal will lead to data security events of sensitive data leakage[2]. Effective discovery and identification of these sensitive data is a precondition for achieving sensitive data desensitization, and it is a prerequisite for implementing sensitive data protection.

Enterprise data types can generally be divided into two categories, personal data and enterprise production business data.Take the personnel data as an example, including the name, gender, date of birth, ID card number, native place, nationality, education background, email address, mobile phone number, address and other fields.Business data mainly has five kinds, respectively for the core business data, general business process data, business process auxiliary data, industry sharing data and other reference data, the core business data contains marketing data, production data, quality data, financial data and other companies at the top of the core business data, the core business data contains a large number of sensitive data.

ICIDC 2023, June 02-04, Nanchang, People's Republic of China

DOI 10.4108/eai.2-6-2023.2334615

Manufacturing enterprises have a number of business departments and subordinate factories, according to the business needs of the many different business systems, distributed in different regions and sub belong to different departments.Some of these business systems have interconnection scenarios, and data flows, exchanged and shared between different business systems.At different stages of business system development, testing, and operations, different roles are exposed to different types of business data.Without the sensitive data identification of these data, the sensitive data cannot be further processed, and there is a risk of sensitive information leakage.With the large number and huge number of business data types, how to quickly and accurately identify sensitive data is a huge challenge.If the artificial way for sensitive data identification, cannot realize the new data continuous automatic identification, and the problem is low in identification efficiency, identification errors and omissions, need to study sensitive data automatic discovery method, establish a new identification mechanism, make the data in the process of the whole life cycle get good security control[3].

Data interaction between the application systems of manufacturing enterprises is increasingly frequent. In order to meet the needs of data sharing in the development, testing, production, application and other scenarios, it is necessary to automatically discover and identify the sensitive data and perform automatic desensitization, so as to maximize the utilization of data value and reduce the risk of sensitive data leakage.The traditional sensitive data automation discovery method is mainly based on rules[4-6]. This method is difficult to achieve and efficient, but it is only applicable to data with specific rules, and is not applicable to sensitive data that is frequently updated and the data format is not fixed.

2. RULE-BASED, TRADITIONALLY SENSITIVE DATA DISCOVERY

Sensitive data discovery is both manual and automatic.Manual identification methods mainly rely on the identification person's empirical conduct and pre-defined sensitive data dictionaries.The identification personnel can judge which attributes are sensitive data according to the predefined data model, and then use the form of sampling to find and identify the sensitive data.The speed of manual identification is much slower than that of automatic identification, and it is easy to appear omissions and errors, and the results of identification will be different due to the inconsistency of human subjective judgment criteria.The automatic mode can automatically identify the data in the database according to the specified some sensitive data samples or predefined sensitive data characteristics, improve the identification efficiency, reduce the error rate, and continuously discover new sensitive data[7-9].

The traditional sensitive data discovery method based on rules can be used for certificate number, bank card number, mobile phone number, mailbox, date and other types of data, and corresponding matching rules are required for each type of data[10]. The sensitive data is automatically identified according to the data features without depending on the definition of tables and fields.Compthe field data in the specified database table according on the data features of the rule configuration and calculate the matching degree.When the matching degree reaches the previously set threshold, the matching is deemed successful.The accuracy of sensitive rule-based data recognition mainly depends on the accuracy of matching rules, but

the accuracy and performance can not be achieved. The simpler the rule setting, the better the system performance, but the lower the recognition rate, and vice versa.

3. SENSITIVE DATA DISCOVERY BASED ON AI

In the actual application scenario of enterprises, the rule-based automatic discovery method can be adopted for the sensitive data with stock, irregular update and fixed rules in the data format; and for the sensitive data with incremental, frequent update and unfixed data format, the sensitive data can be automatically discovered based on artificial intelligence.Recognition methods based on artificial intelligence can find data without clear rules, such as addresses and proprietary names, and there is no need to make matching rules, but the accuracy and efficiency depend on the implementation of artificial intelligence recognition algorithm[11]。

Automatic entity recognition is a natural language processing technology that can identify human names, place names and other information.AI automatically named entity identification has statistical model-based methods and deep learning-based methods[12-13]. The main ones based on the statistical model are maximum entropy model, hidden Markov model and conditional random field model.Statistics-based methods to identify named entity recognition requires the contextual information of statistical words, such as word position, and collocation between words.As the application scope of the deep learning technology expands, the deep learning technology can be trained with word vectors from a large number of unannotated corpora, and then identified.Manufacturing enterprises can use artificial intelligence technology to automatically identify the sensitive data without fixed rules generated by business and production activities.

3.1 Conditional random field

Conditional random field (CRF) is a probabilistic structure model with the characteristics of discriminative model and undirected graph model. It can be used to divide and annotate sequence structure data, and is suitable for sequence analysis and annotation, such as named entity identification.

3.2 Convolutional neural network

Convolutional neural network (CNN) is a kind of feedforward neural network. It is a typical deep learning algorithm, with deep structure and contains convolutional computing.Its hidden layer is composed of pooling layer, convolutional layer and full connection layer.The convolution layer then extracts the input data with features, consisting of multiple convolutional kernels, and each element of the convolutional kernel corresponds to a quantity of deviation and a weight coefficient.After the feature extraction of the convolution layer, the feature map is passed as an output to the pooling layer, and used by the pooling layer for feature selection and information filtering.

Convolutional neural network can simplify the extraction of features, and can extract good features through model training.Convolutional neural network for text feature extraction process is shown in the following steps:

1）Convolutional layer: the raw data is filtered through the convolution check of the two convolution layers to extract the data features, reduce the number of parameters and reduce the training time through the local convolutional kernel and parameter sharing.

2）Pooling layer: After obtaining the text local structure features through the convolution layer, pooling through two pooling layers, merge the convolution layer features according to the semantic similarity to reduce the data dimension, improve the training speed and avoid overfitting.

3）Text feature extraction: the input of the model has a character vector with certain semantic features, and the text has n words. If the m-dimensional word vector is used to represent a single word, an n * m-dimensional character feature matrix should be constructed.

3.3 Bidirectional long-term and short-term memory network

Chain structure of long short-term memory network LSTM contains four interactive neural network layer, through gating state to control the transmission state, discard unimportant information and memory needs long time memory, suitable for predicting and processing events with long delay and long interval, suitable for the modeling of temporal data such as text data.

LSTM can not code information from back to before, and in the actual text processing application, text has sequence feature correlation, the current moment words is not only affected by the past moment words, will be affected by the future moment words, in order to make the model contains the context information, using forward LSTM and LSTM back L S T M combined bidirectional long short-term memory model BiLSTM, training forward and backward sequence, respectively.Forward LSTM layer represents sequence sequence, backward LSTM layer represents reverse sequence, through the forward LSTM output to the

past information, through the backward LSTM output get future information, the forward and backward by combining BiLSTM layer output, BiLSTM structure can provide historical information and future information to the output layer at the same time, can solve long-term dependence, and both the context information, effectively processing sequence problem[14]。

3.4 The CRF-BiLSTM-CNN model

Sequence output can be obtained from the context feature information output by BiLSTM, but the sequence annotation problem has a strong dependence. In sensitive data identification, the tag sequence of each word has certain limitations. Model the dependence of the CRF model between entities after BiLSTM to deal with sequence annotation.Continuous words have spatial semantic information between them, and each word block in the sentence is interrelated and influential.In order to use this spatial semantic information, the convolutional neural network CNN is introduced, and the word vectors were trained with the CNN, with the spatial semantic features extracted, and then combined with the word vectors as the input to the BiLSTM.

The CRF-BiLSTM-CNN model is divided into word vector representation parts, convolution and pooling parts, and BiLSTM-CRF parts.

1）Word vector representation: the word vector representation part is below the model, and its main role is to convert the Chinese text into the form of the vector input into the deep neural network.The word vector in this paper mainly adopts the Word2vec word vector representation method.Word2vec word vectors have semantic information and have similar vector representations for semantically similar words.Word vector represents that each character of the training corpus is trained into a 100-dimensional vector using the Word2vec tool, yielding the input sequence, X is the K-dimensional vector.

$$X = \{x_1, x_2, \ldots, x_n\} \tag{1}$$

2）Convolution and pooling: The convolutional layer functions to extract the features of the matrix vectors.Features were extracted by using filters of different sizes.In this paper, the filter size is set to m, and then the vector features can be computed by the filter for the continuous m vectors in the input vector k.

$$k_i = \tanh(w_i + b) \qquad (i = 1,2,\ldots,m) \tag{2}$$

the w_i is input vector weight,$X = [x_i, x_{i+1}, \ldots, x_{i+m-1}]$ is a word vector with continuous m words,

, b is the bias matrix,by convolution Obtaining the features of each word in a sentence k_i.The sentence features is $K = [k_1, k_2, \ldots, k_{n+m-1}]$,useing window size l to Pooled sampling, geting a-dimensional feature vector Q.

$$Q = [k_1, k_2, \ldots, k_a] \tag{3}$$

and

$$a = (i + m - 1)/l \qquad (i = 1,2,\ldots,n) \tag{4}$$

3）BiLSTM-CRF: The forward LSTM unit uses the output of the decision function through the current word left information; the backward LSTM unit combines the output vector of the decision function through the current word right information and the forward output vector and inputs the conditional airport to obtain the final sequence.

4. VALIDATION

The verification environment is the open source deep learning framework TensorFlow1.2.0, the programming language is Python3.5, the CRF experimental tool uses the open source machine learning tool kit CRF + +, and the experimental data uses the People's Daily corpus. The experimental results are shown in the Table 1.

Table 1. Comparison of experimental results from different models

model	precision	recall
CRF	75.47	78.68
BILSTM-CRF	87.96	91.46
CRF -BILSTM-CNN	87.71	93.64

Where accuracy is defined as the proportion of the number of correctly identified sensitive data entities to the total number of identified sensitive data entities, and recall is defined as the proportion of the number of correctly identified sensitive data entities in the total number of sensitive data entities.

The experimental results show that the CRF-BILSTM combined model is much better than the CRF model. The BiLSTM can synthesize the relevant information of the context and significantly improve the performance of the model.The CRF-BILSTM-CNN introduces a convolutional neural network on the basis of the BILSTM-CRF model for feature extraction, which can better represent the sensitive data text and improve the model performance.Therefore, the CRF-BiLSTM-CNN can obtain a higher performance than the general deep learning models, which is suitable for the sensitive data discovery without fixed rules generated by the business operation and production activities.

5. CONCLUSION

With the integration of the two enterprises, the enterprise application systems have been widely used, effectively promoting the production and business activities, and these application systems produce a large number of sensitive data. How to automatically discover these sensitive data is the prerequisite for implementing data desensitization and data security sharing.This paper conducts research on sensitive data discovery methods based on artificial intelligence algorithm, which provides new ideas and solutions for sensitive data discovery methods of enterprises, and provides a new direction for sensitive data discovery of manufacturing enterprises.

FUND PROJECT:

Data security sharing technology research project of cigarette manufacturing enterprises [(2020) 0068], Construction project of Network Security Management Center of Wuhan Ask Information Technology Co., LTD. [(2020) 0075]

Information System Life Cycle Management Project of China Todacco SiChuan Industrial LLC.[2020510000340518]

ABOUT THE AUTHOR:

Shi Jie (1982), Chengdu, Sichuan province, engineer,M.D., main research direction for network system, information security system and other infrastructure project technical plan and project implementation, infrastructure system operation maintenance and technical support, Chengdu jinjiang district, Sichuan province 2 paragraph 86, telephone 02867838126, email 47443634@qq.com;

Cui Shufang (1985), male, Hubei Xiantao, assistant engineer, B.D., the main research direction: network system, information security system process and other infrastructure project technical plan formulation and project implementation, infrastructure system operation and maintenance and technical support, 430040,13397149606, cuishufang@hjl.hbtobacco.cnt.email 228605126@qq.com

Chen Fei (1982), male, Wuhan, Hubei province, engineer, B.D., main research direction: Network security, IT Infrastructure and operation and maintenance management, Yellow Crane Building Science Park, 1355 Jinshan Avenue, Dongxihu District, Wuhan city 430040,13995584732, 47443634@qq.com;

Wang ChengTao(1971), male, Hubei province, engineer,M.D., main research direction: technical research and development of software design and development, system integration, information security and other aspects, Huanghe Tower Science Park, 1355 Jinshan Avenue, Dongxihu District, Wuhan City, 430040,18062510304, 47443634@qq.com;

REFERENCES

[1] Guoqiquan, zhanghaixia Key information infrastructure security protection technology system [j] Information network security, 2020, 20 (11): 1-9

[2] The National People's Congress Network security law of the people's Republic of China [eb/ol], 2016-11-7.

[3] Liu Jin Sensitive data recognition method based on data characteristics [j] Information and communication, 2016 (2): 240-241

[4] Zhanghainan, wudayong, Liu Yue Chinese Named Entity Recognition Based on deep neural network [j] Chinese Journal of information technology, 2017, 31 (4): 28-35

[5] Fengyanhong, Yu Hong, sun Geng Named entity recognition method based on blstm [j] Computer science, 2018, 45 (2): 261-268

[6] Geng Xin Construction of safety assurance system for industrial control system in tobacco industry [j] Tobacco science and technology, 2017, 50 (12): 99-105
[7] Li Lu, Tao Linghan Safety challenge of industrial control system [j] Confidential science and technology, 2016, 000 (012): 34-36
[8] Geng Xin, Geng, Xin, et al Construction of safety assurance system for industrial control system in tobacco industry [j] Tobacco technology, 2017
[9] Yangruichao, yuejianhui, hang Xiao Discussion on the safety protection of industrial control system under the environment of industrial Internet [j] Network security technology and application, 2020, 000 (005): 108-109
[10] Zhang Shu Industry 4.0 and intelligent manufacturing [j] Mechanical design and manufacturing engineering, 2014 (8): 1-5
[11] Chu Lu, Hong Liu, "Generative Adversarial Networks Based Industrial Protocol Construction in the Fog Computing", Local Computer Networks (LCN) 2018 IEEE 43rd Conference on, pp. 453-456, 2018
[12] Xiao Long, Fang Yong, Dai zongkun, Yang Wei, Cai Heng Risk analysis of information system based on fuzzy neural network [j] Computer application research, 2006, (5): 137-139
[13] Hu Yong, Qi Gang, Chen Lin, Yang Wei Information system risk quantitative evaluation index system [j] Journal of Sichuan University (NATURAL SCIENCE EDITION), 2006, 43 (5): 1048-1052
[14] Zhang Tao, Hu mingzeng, yunxiaochun, zhangyongzheng Research on computer network security analysis modeling [j] Journal of communications, 2005, 26 (12): 100-109