Modeling Analysis of Network Spatial Sensitive Information Detection Driven by Big Data

Ruijuan Liu[1(✉)], Bin Yang[1], and Shuai Liu[2]

[1] College of Arts and Sciences, Yunnan Normal University, Kunming 650000, China
liuruijuan552@163.com
[2] College of Computer Science, Inner Mongolia University, Hohhot 010012, China

Abstract. The dissemination of sensitive information has become a serious social content. In order to effectively improve the detection accuracy of sensitive information in cyberspace, a sensitive information detection model in cyberspace is established under the drive of big data. By using word segmentation and feature clustering, the text features and image features of current spatial data information are extracted, the dimension of the data is reduced, the document classifier is built, and the obtained feature documents are input into the classifier. Using the open source database of support vector machine (SVM) and LIBSVM, the probability ratio of current information belongs to two categories is judged, and the probability ratio of classification is obtained to realize information detection. The experimental data show that, after the detection model is applied, the accuracy of the text-sensitive information detection in the network space is improved by 35%, the accuracy of the image information detection is improved by 29%, and the detection model has the advantages of obvious advantages and strong feasibility.

Keywords: Big data · Sensitive information · Spatial data · Information detection

1 Introduction

The rapid development of Internet has brought people unprecedented convenience of information. It is more convenient for people to get all kinds of information from Internet. However, the network also has a serious negative impact, in the good buckwheat coexistence of the network information is full of a large number of sensitive information. The harm of sensitive information, especially to teenagers, is enormous. At present, Chinese netizens under 24 years old account for about half of the total number of netizens, and the majority of them are college students and primary and secondary school students. Young people's self-control is not strong, and they are easily induced by bad network information in cyberspace. They are addicted to the network not to make progress, to throw away their time, and even to cause a lot of social problems. The network harm has caused the national relevant department to attach great importance to, and has adopted a series of measures. On March 26, 2012,

G. Gui and L. Yun (Eds.): ADHIP 2019, LNICST 301, pp. 1–11, 2019.
https://doi.org/10.1007/978-3-030-36402-1_1

the China Internet Association issued the <China Internet Industry Self-Discipline Convention>, which prohibits the dissemination of sensitive information in cyberspace. On May 10, 2013, the Ministry of Culture issued the <Interim Provisions on the Administration of Internet Culture>, which states that Internet operators providing sensitive information and other illegal information products should be strictly dealt with [1, 2].

The 21st century is a highly information-based era, with the rapid progress of science and technology, especially with the continuous acceleration of the construction of big data's network. The increasing bandwidth and transmission rate of the mobile operating network provide a large number of channels for the transmission of network data, and the content services and data services for intelligent terminals will also be continuously enriched. All these make the mobile network soon to follow newspapers and periodicals, broadcast, and so on. Television, like the Internet, has become an important carrier of cultural communication and information exchange. With the rapid development of communication technology and the remarkable improvement of network bandwidth technology, with the help of the Internet and mobile network, information exchange and dissemination account for more and more proportion, people can not only through the traditional text information, and also through the picture, video and other multimedia information dissemination and communication. Under this background, the detection and filtering of sensitive information in cyberspace has become an important task in network construction. Therefore, based on the current big data domain driver, a network spatial sensitive information detection model is established to analyze and detect the spatial sensitive information and improve the network security [3, 4].

2 Design of Network Spatial Sensitive Information Detection Model

In order to construct a practical sensitive information detection model, the correlation principle of information must be considered. The information in the same domain usually has some common properties, and the information in different fields has its own characteristics. Under the influence of big data, this paper firstly analyzes the characteristics of sensitive information, including bad text and bad image, in the context of related applications, and then extracts representative independent feature vectors from these information. An information filtering classifier based on these feature vectors is designed, and a pattern recognition method is adopted to achieve the purpose of classification and detection of sensitive information according to the probability ratio [5, 6].

After the above analysis, it is clear that after obtaining the information by big data driver, it is necessary to judge whether it contains text or image information first, and then pre-process the information. If the text content is included, the text is immediately segmented, and the word frequency is counted to extract the feature of the entry, and the text classifier is used to identify and judge the probability of the text belonging to the category. This probability is used as a parameter to describe the whole information. An upper and lower threshold is defined for the output of the text analysis module, which specifies that if the parameter is within a predefined range, it is not sufficient to

determine the category of information, and further image analysis is required. The image part is input into the image classification module for processing. If the parameter value is higher than the predefined upper limit, it indicates that the content has been judged to be sensitive information, and no longer needs to be processed by the image classification module, so it is directly masked. If the parameter value obtained by text analysis is lower than the predefined threshold, it indicates that the content can be judged as normal information and can be released directly without blocking. After the contents of the image to be detected are put into the image processing module, the extracted features and the parameters obtained from the text analysis are combined into a high-dimensional vector, and the classification results are obtained by the classifier decision-making. The flow diagram of the model is as shown in Fig. 1.

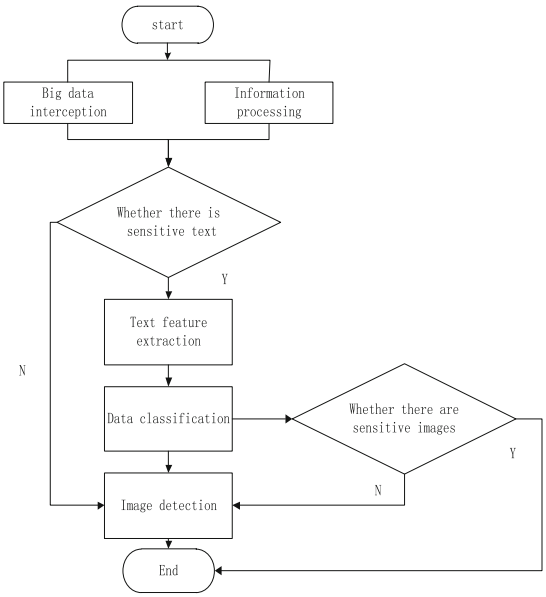

Fig. 1. Flow chart of sensitive information detection

2.1 Information Feature Extraction Driven by Big Data

Big data-driven text and images may extract a large number of different features, some of which are of great significance to information classification and sensitive detection, and some of them are not. If all the features are included in the sensitive feature vector, it not only increases the time-consuming of feature extraction and the computation of the classifier, but also may introduce noise to the classification. Therefore, first of all, we need to find an independent method for extracting and reducing the dimension of the sensitive information feature vector, that is, feature clustering algorithm [7, 8].

The clustering algorithm allocates every row of data in the current data set driven by big data to a group or a point in the hierarchical structure, and each data exactly

corresponds to a group, which represents the average level of the members in the group. Sensitive data feature extraction is to try to find new data rows from the data set and combine these newly found data rows to construct the data set. Unlike the original data set, each row of data in the new data set does not belong to a cluster, but is constructed by a combination of several features [9].

The essence of word segmentation technology in document feature extraction is to calculate the first few words with the highest TF x IDF value for all words appearing in the document as document features. According to the frequency of feature words appearing in two kinds of documents, the probability of each word belonging to two categories is calculated, and the calculated probability is stored in the database according to the fixed vocabulary order.

For a document to be classified, a vector representing the frequency of each feature word appears as the feature vector for the document, such as $(w_1.w_2.w_3.\ldots.w_n)$, where w_i represents the frequency at which the i-th feature word appears in the document. In the actual design, the N value is 1500, that is, from the current network data training sample, the total of 1500 words with the largest value in order of TF x IDF size are extracted from the current network data training sample as feature words. This includes 700 feature words from junk text and 800 from normal text. Then we do the TF x IDF calculation of the terms and the corresponding frequency in all kinds of documents, take the first 700 words and the first 800 words of the two kinds of calculation results, and generate two new dictionaries. The number of times a word appears in all articles and each article is recorded, and the data must be transformed into matrix form, in which each row represents one data item and each column represents an attribute of the data item [10, 11].

For the current information text and image classification, the row of the matrix corresponds to all kinds of information, and the column corresponds to the word or pattern in the article, and each number in the matrix replaces a word in one. The number of times a given article appears. For example, get a matrix like Eq. 1.

$$W = \begin{vmatrix} 0 & sex & star & road \\ A & 3 & 1 & 0 \\ B & 0 & 2 & 3 \end{vmatrix} \tag{1}$$

This matrix represents the eigenvector sex three times in class A, and start appears twice in class B, and so on. There are two information to focus on from this matrix:

The first is allwords, which records the number of times a word is used in all articles, and it can be used to determine which words should be seen as part of a feature and the other is articlewords, which is the number of times a word appears in each article [7].

According to the definition of feature dimension information, the common words in all kinds of articles carry small amount of information, poor classification performance, and few words have little meaning to classification, so reduce the size of matrix. Words that appear in only a few documents should be removed, and words that appear in too many articles should be removed. Only words that meet the requirement of less than 60% of all articles that have appeared in more than three documents are considered here. After the above pretreatment, a document matrix with word counting information

can be obtained. The next step is to extract important features from the matrix to achieve the purpose of dimensionality reduction [12].

The non-negative matrix factorization method is used to factorize the document matrix and two smaller matrices are found so that the document matrix can be obtained by multiplying the two matrices. These two matrices are characteristic matrix and weight matrix respectively. In the feature matrix, each row corresponds to a document category and each column corresponds to a feature word. The numbers in the matrix represent the importance of a word to a document category. The function of the weight matrix is to map the document category to the document matrix, in which each row corresponds to a training sample, that is, a document, and each column corresponds to a document category. The numbers in the weight matrix represent the extent to which each document category is applied to each sample. This allows you to list the top 450 word features of the two document categories, which are the most important words in the document category, and you can select these 900 words as a text feature. That is to say, the text feature vector is reduced from 1500 to 900D, and the data document can be used as feature extraction document at this time.

2.2 Classifier Construction

In order to recognize sensitive information, the feature document extracted in the previous section should be regarded as a high-level semantic feature, and this mapping is accomplished by a system classifier [13].

Automatic classification is a new pattern classification and recognition technology introduced in statistical learning theory. It is defined as a set of training samples that have already been assigned class tags (these class tags need to be meaningful). To assign class labels to the new sample. In the network filtering technology, the classifier needs to accurately identify an infinite number of unknown samples, but in the learning stage of the classifier, it is impractical to collect a complete database of samples with rich representativeness and diversity. Therefore, classifiers need to start from a small number of learning samples, constantly learn and improve their classification performance. More and more learning methods are used in automatic classification, in which support vector machine (Support Vector Machines, SVM) method has advantages in solving problems such as small sample, nonlinear high-dimensional pattern recognition and so on, and has better generalization ability. Considering the influence of the small sample training set on the classifier, the computational complexity of the classifier and the processing time of the system, the SVM classification algorithm is introduced. The idea is to try to find a line as far away from all categories as possible, which is called the maximum interval hyperplane, as shown in Fig. 2 (For the sake of simplicity, only the example of linear separability is given. SVM can be extended to high dimensional space).

The basis for selecting this dividing line is: Two parallel lines passing through the corresponding coordinate points of each classification are searched, and the distance between them and the dividing line is made as far as possible. For new data points, the classification can be determined by observing which side of the boundary line it belongs to. It is to be noted that only the coordinate points located at the edge of the spacer are necessary to determine the position of the boundary, and if all the remaining

data is removed, the dividing line will still be in the same position. Coordinate points near this demarcation line are referred to as support vectors [14, 15]. Support vector machine (SVM) is an algorithm for finding support vectors and finding boundary lines by using support vectors.

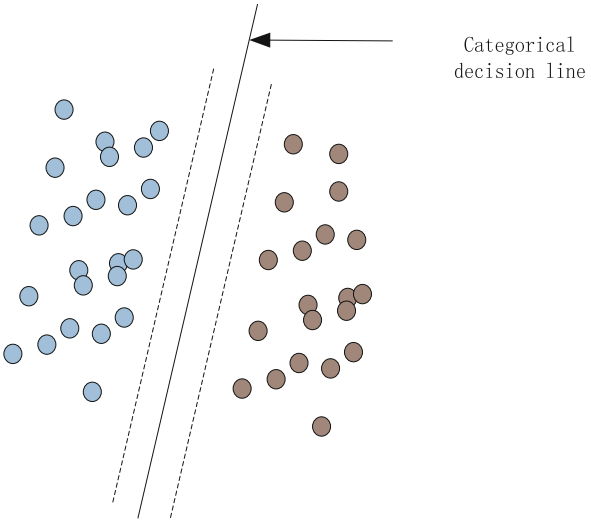

Fig. 2. Vector machine classification diagram

The solid lines in Fig. 2 show two possible decision planes, each of which correctly classifies the data into two categories. Similarly, for new data points, the classification of the new data points can be determined by observing which side of the boundary line it belongs to. The two dashed lines parallel to the solid line are the maximum offset positions of the decision face without causing misclassification. The distance between the two dashed lines is the classification interval of the decision face. The purpose of SVM is to find the decision surface with the maximum classification interval in all sample points. In fact, only the sample points at the edge of the spacer are necessary for classification, and in the case of linear separability, the function of the decision plane is $\bar{w} \cdot x - b = 0$. In which, x is any sample point to be detected, W and constant b are obtained by training the sample point. Let the input of SVM algorithm be a linear separable sample set $D = \{(y_i, x_i)\}$ be the classification of x. "+1" indicates that it is a positive example and "−1" is a counterexample.

The SVM problem is to find w and b that satisfy the conditions shown in Eq. 2, and the module of vector w is minimal (The classification interval is equal to 2 divided by the module of w).

$$\begin{cases} \bar{w} \cdot x - b > 1 & y_i = +1 \\ \bar{w} \cdot x - b < 1 & y_i = -1 \end{cases} \tag{2}$$

The classification function is required to classify all samples correctly, that is, satisfying Eq. 3.

$$y_i[(w \cdot x_i)] + b - 1 \geq 0 \tag{3}$$

Because the programming workload of the support vector machine algorithm is very large, an open source library called LIBSVM is introduced, which can train an SVM model, give the prediction, and test the prediction result with the data set. LIBSVM also provides support for radial basis functions and many other core methods, which is written in C++ and has a version of ava. We need to select the appropriate LIBSVM compiled version based on the platform you are using, and because the design model is developed in the windows environment, you need to include a DLL file named svmc.dll. The LIBSVM documentation details how to use classifier functions.

A minimum threshold is defined for each classification. For the data information to be assigned to a certain category, the probability must be greater than a pre-specified threshold compared to the probability for all other classifications. The default threshold is 3, which means that the probability of document classification for sensitive classes is at least three times higher than that for ordinary documents. This threshold can be adjusted according to the actual application of the individual.

2.3 Realization of Sensitive Detection

Because the design adopts a dictionary-based word segmentation method, firstly, a complete dictionary is needed, the documents in the training sample library are pre-processed and segmented according to the dictionary and the stop word list, two new dictionaries are generated after the word segmentation, the number of times the words appear in all articles and in each article are recorded, which are the preliminary text features. However, if the text is classified directly by these features, it will result in a high dimension of the feature vector and a poor classification effect. Therefore, according to the two dictionaries, each feature generated above is subjected to a dimension reduction process to obtain an independent feature, that is, the feature of the text classification.

Independent features are used to treat the detected text for feature extraction. The extracted feature entries are combined into a high-dimensional vector to represent the text to be detected. Then the classifier is inputted into the classifier and the probability of the class of the information to be detected and the class of the document are obtained by decision-making. The probability ratio of documents belonging to two categories is judged, and if it is greater than 3, the information is directly judged to be sensitive; If it is less than 3, it continues to extract the image features, and combines the document probability and the image features into the feature vector of the whole web page information to the system classifier for further judgment to see whether it is sensitive information or not. The histogram feature extracted from each image is 192, the color aggregation vector is 48, the skin region feature is 4, the face feature is 2, plus the probability that the document obtained from the text recognition module belongs to

junk text and normal text, respectively. A total of 248 features. Combining these features into a high-dimensional vector, a feature vector $(t_1, t_2, \ldots, t_{248})$ can be used to describe the information to be detected. Then the high-dimensional vector is input to the system classifier for training or classification decision-making, so that the sensitive data can be monitored.

3 Experiment

The C# language is used to realize the sensitive information filtering system in Windows XP, and the classification performance of the system is tested with a large number of experimental data. The hardware environment of the development platform is CPU of Intel (R) Pentium (R) dual-core T2390d. 80 GHZ, 2.79 GHZ 1G memory, 256 m graphics card. Software configuration: windows XP professional version of the operating system, Microsoft Visual Studio C# integrated development environment.

The experimental data includes text and image, and the information filtering system is tested respectively. Among them, the text class sample uses the Chinese natural language open processing platform (http://www.nlp.org.cn), collected a total of 2270 texts, divided into three groups, each group of two categories, each class selected 450 features, the total selected feature vector dimension of 900. The images in the image library are mainly collected from the network, because compared to landscape images, character images are more difficult to classify because they contain human body and skin regions and sensitive images. So the two kinds of images are divided into two groups to train and test the system. Due to limited sources of information, the graphics library consists of only 570 images divided into five groups, the first of which consists of 60 sensitive images and 40 normal landscape images, the first for training and the other four for testing. After each set of images is tested, it is used to train the system to analyze the influence of the number of training samples on the classification results of the system, in which the linear SVM. is used for the SVM classifier. It is important to note that when training the image feature base of SVM classifier, because there is no input of the text classifier, the probability parameter of the document belonging to the classification of the 248 feature components is set to 0. 5.

3.1 Text Sensitivity Detection

The text sensitivity detection is carried out in the above experimental environment, and the contrast detection method is the traditional ULL detection model. Its monitoring effect is shown in the chart below. The results are shown in Fig. 3.

As can be seen from Fig. 3, the detection accuracy of traditional detection model is poor, while the detection accuracy of designed model is higher than that of traditional detection model.

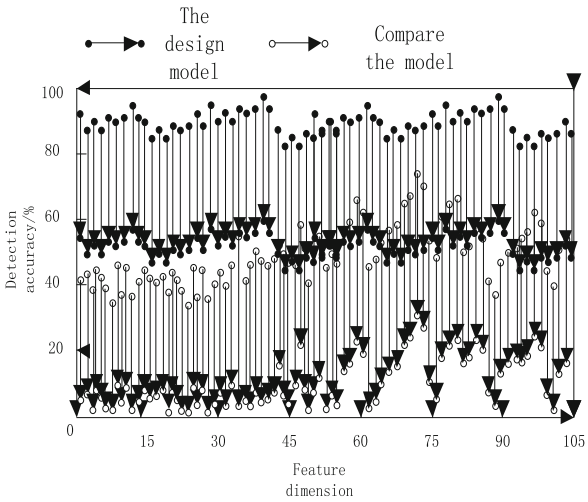

Fig. 3. Comparison of detection effects between the two models

3.2 Image Detection

According to the experimental method mentioned above, the network space image detection is carried out, and the detection results are as follows:

Table 1. Image detection results table

Data set	Design model	Traditional model
1	82	57
2	79	45
3	85	50
4	95	69
5	92	70
6	89	52
7	91	49
8	86	70
9	93	42

According to the Table 1 data, it can be seen that in all 10 sets of image data, the monitoring accuracy of the designed detection model is also higher than that of the traditional model. According to the comparison of the data, the accuracy rate increases by more than 29%.

4 Conclusion

The content-based sensitive information filtering is a hot spot research direction. It is actually a problem of information identification and information classification. It is an application of many-door disciplines such as text classification, image processing, computer vision, programming and pattern recognition. Human physiology and psychology are closely related, and have wide application prospect. The popularization of high-speed information high-speed highway has further promoted the development of this technology. At present, the sensitive information filtering technology has been paid attention at home and abroad, among which, the text classification technology has become mature, and the image classification technology is still in the research, is gradually refined and the mouth is perfect, and has produced many application and test examples. The sensitive information detection model proposed by the design can achieve the purpose of effective information detection by comprehensively using the relevant technology of the information classification.

Acknowledgment. The authors would like to thank State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System Director Fund (CEMEE2019K0104B).

References

1. Bi, J., Li, H., Xing, M.: Analysis on talent training mechanism and mode of cyberspace security under the background of big data. Mod. Vocat. Educ. **7**, 159 (2018)
2. Liu, S., Bai, W., Liu, G., et al.: Parallel fractal compression method for big video data. Complexity 2016976 (2018)
3. Wang, W.: Research on the current situation and countermeasures of cyberspace security under the background of big data. China Strateg. Emerg. Ind. **156**(24), 100+102 (2018)
4. Miao, L., Shuai, L., Weina, F., et al.: Distributional escape time algorithm based on generalized fractal sets in cloud environment. Chin. J. Electron. **24**(1), 124–127 (2015)
5. Tang, W., Wang, Y., Wang, J., et al.: Research on alienation control model of network public opinion information in the context of big data. China New Commun. **20**(10), 140 (2018)
6. Bing, J., Shuai, L., Yongjian, Y.: Fractal cross-layer service with integration and interaction in internet of things. Int. J. Distrib. Sensor Networks **10**(3), 760248 (2018)
7. Wu, J.: Research on college students' virtual cyberspace behavior management from the perspective of big data. Inf. Comput. (Theory Ed.) **7**, 223–224 (2018)
8. Lu, M., Liu, S., Sangaiah, A.K., et al.: Nucleosome positioning with fractal entropy increment of diversity in telemedicine. IEEE Access **6**, 33451–33459 (2018)
9. Xia, Y., Lan, Y., Zhao, Y.: Research on alienation control model of online public opinion information in the context of big data. Mod. Intell. **38**(2), 3–11 (2018)
10. Shu, X., Yao, D., Bertino, E.: Privacy-preserving detection of sensitive data exposure. IEEE Trans. Inf. Forensics Secur. **10**(5), 1092–1103 (2017)
11. Liu, S., Bai, W., Zeng, N., et al.: A fast fractal based compression for MRI images. IEEE Access **7**, 62412–62420 (2019)
12. Fan, P.: Analysis on network information security protection strategy in the era of big data. China New Commun. **20**(09), 134 (2018)

13. Zheng, X.: Promoting the construction of network trust system under the condition of big data intelligence. China Nat. People's Congr. **455**(11), 51 (2018)
14. Zhang, Y.: Development trend of information technology and network space security. Farmstaff **580**(08), 240 (2018)
15. Wei, W., Shuai, L., Wenjia, L., et al.: Fractal intelligent privacy protection in online social network using attribute-based encryption schemes. IEEE Trans. Comput. Soc. Syst. **5**(3), 736–747 (2018)