ATLANTIS PRESS
Advances in Intelligent Systems Research (AISR), volume 151
2018 International Conference on Computer Modeling, Simulation and Algorithm (CMSA 2018)

An Analysis Model of Network Illegal Fund-raising Based on Advertisement Content Mining

Jianying Xiong*
Security Management Department of JiangXi Police College, Jiangxi Police Institute, Nanchang 330013, China
*Corresponding author

***Abstract*—Network Illegal fund-raising is a social problem and is more subtle and destructive. Government regulators should explore more effective control methods for early discovery. At the preliminary stage of the illegal fund-raising, in order to promote the investment effect and producing the spread effect in the network, it need to do more propagandizing activity. Investment advertisement content with illegal fund-raising risk may some certain characteristics. Risk mining is carried out for the advertisement content in this study. The advertisement content risk is defined as three factors: risk loss, risk probability and risk diffusion speed. Through the establishment of an illegal fund-raising advertisement content anomaly model, the illegal fund-raising behavior can be detected by the method in the early stage.**

Keywords—network illegal fund-raising; abnormal monitoring; risk assessment; content mining

I. INTRODUCTION

The illegal fund-raising means those companies, enterprises, individuals or other organizations raising funds to the public through improper channels. The behavior is not been approved, disregard of laws and regulations, and is the essence of the crime. The illegal fund-raising of the network refers to the illegal fund-raising in the network space, which has the characteristics of low publicity cost, fast propagation and wide coverage [1]. In recent years, many blind investors has been deceived by Internet illegal fund-raising, and has also harmed the national normal financial order and social credit system by taking the investment and financing, emerging industries, green environmental protection and other guise [2]. The illegal fund-raising of the network often has a strong confusion, and it is not easy to distinguish in the early stage. They use "high return" and "high interest" to mislead a lot of investors. It is difficult for the ordinary people to identify the illegal fund-raising of the Internet, and the government regulatory authorities are facing some new problems. The network of illegal fund-raising is often done quickly in a very short period of time. The investors from all over the country are mostly through the network to participate in. It is difficult to effectively screen and deal with mass and low value information by manual or simple data classification system. Data resources are difficult to be exploited and utilized effectively, which greatly restricts the efficiency of handling cases [3]. In the case of illegal fund raising, suspects will leave a lot of data traces and clues on line and offline, including network information and all kinds of public information. The behavior of illegal fund-raising on the Internet is always accompanied by warning signs. Based on the preliminary investigation of network information, investigators can take some illegal publicity enterprises into the monitoring list, rank them according to the risk level, and timely monitor the dynamic changes of the online public opinion of the list enterprises, then do early detection and monitoring [4].

II. RELATED WORK

Under the traditional mode of investigation, the main clue of illegal fund-raising crime is found by police, and the reports of the investors or investors' relatives. Under the strict organizational structure of the illegal capital collection group, the public security organs are limited to obtain clues. The connection with risk is more common. In the field of fund-raising behavior research, the main research trend is the research on the characteristics and identification of the crime such as fund-raising fraud. Study on the behavior of fund-raising risk is mainly about illegal fund-raising behavior in the legal regulation as the core, and rarely rely on data mining and artificial intelligence method for the identification of illegal fund-raising behavior, and the risk assessment model based on its characteristics [5,6]. So there is no much relevant academic research on early warning mechanism.

In order to cope with economic crime investigation network financial pressure, some applications in early warning of illegal fund-raising, for example, ShuLian MingPin anti-fraud product line (BBD Anti-Fraud) is a project of monitoring high risk enterprise pornography holographic portrait to assist the Beijing Municipal Finance Bureau. BBD Anti-Fraud is also can provide the appropriate visualization methods to show related enterprises, and provide information to the police. The platform is more focused on the collection, classification and statistics of the information of the network collection, so as to facilitate the collection and inquiry of the evidence for criminal investigation. Beijing Jinxin Financial Information Services Limited has launched online banking risk early warning platform based on the characteristics of the smoking index. According to the previous case of illegal fund-raising platform, comprehensive legal, illegal fund-raising feature word, the high rate of return, negative feedback index, the spread of the five dimensions index. Smoking index is a comprehensive score for all indicators, but not all crime cases will have obvious problems on most indicators. In many cases, most camouflages are good, but some key indicators are not good enough to be smoked. The information sources are more diverse under Internet environment. The public security organs realize the data patrol control in cyberspace and find the crime of illegal fund-raising

on the Internet through the data pre-warning platform and the network grasping technology [7,8]. For example, through the "data patrol control", we found that there is an illegal fund collecting feature sensitive words in a blog site, so that to find an early warning signs of the illegal fund-raising crime. Illegal fund-raising behavior always put lots of advertisement in network to promote people to invest. If you want to monitor illegal fund-raising in the first time, it can be carried out by mining the release of specific forums, investment website, the contents of the public number, the anomaly monitoring, in the early implementation of the illegal fund-raising behavior investigation[9,10]. In this paper, we analyze the investment advertisements and set up an abnormal model of illegal fund-raising propaganda, which can assist the regulatory authorities to quickly and effectively investigate illegal fund-raising behavior and deter crime.

III. CONTENT MINING MODEL

A. Model Description

There is no unified definition of risk for a long time, and people define the risk in the scope and field of their respective research. Taking the risk definition in the financial field as an example, the definition of risk is up to 14 kinds, including uncertainty, probability, loss, volatility and danger. For a crime with illegal fund-raising risk, there are certain characteristics of investment promotion content. Risk mining is carried out for content. The content risk is defined as three factors: risk loss, risk probability and risk diffusion speed. The risk value calculation of investment promotion content is (1). CR shows the value of risk, D means the deviation of earnings, P represents the similarity of risk characteristics, and V indicates the degree of diffusion.

$$CR=D*S*V \tag{1}$$

The definition and calculation methods for each variable of risk value are as follows:

(1) *Earnings deviation degree D*: Almost all financial frauds have a typical characteristic, that is, the higher interest rate. A higher interest rate always accompanied by the higher the risk. If set the profit threshold, the deviation is computed by the distance of actual commitment income and the threshold value. The profit deviation is positive to the distance.

$$D=(Pb-Tb)/Tb \tag{2}$$

(2) *Risk feature similarity S*: Preprocessing the content of the advertisement text, obtaining the keyword sequence of the text, comparing with the feature word library, and obtaining the feature similarity.

$$S=\frac{1}{n}\sum_{i=1}^{n} f_i * w_i \tag{3}$$

N is the amount of text feature words, f_i is the I-th feature word, and W_i is the weight of fi feature words in the feature word library. If the feature word does not exist in the lexicon, the weight is 0. The establishment of the feature word library and its weight is described in 3.2.

(3) *Diffusion degree V*: the degree of diffusion of advertisement content on the network has a positive impact on the risk. A greater of the diffusion degree will make the risk increasing faster. The factors of diffusion degree include the number of relevant content from network search engine and the ranking of similar searches. The diffusion degree can be classified by the fuzzy judgment of the related factors, and the classification rules are as follows:

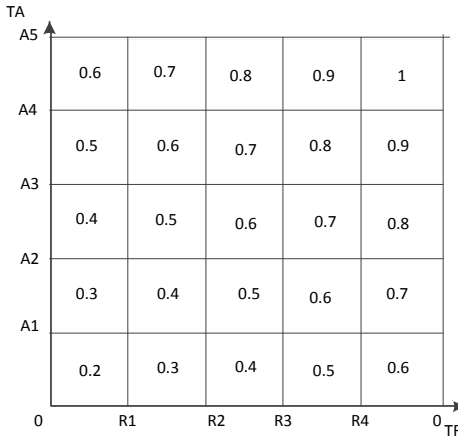

FIGURE I. CLASSIFICATION RULES OF DIFFUSION DEGREE.

The ordinate TA is denoted as the number of relevant content, the coordinates (A_1, A_2, A_3, A_4, A_5) is the five threshold value of Classification definition according to the experience. The abscissa *TR* is denoted as rank. The coordinates (R_1, R_2, R_3, R_4, R_5) is the five threshold value of ranking level according to the first 10 pages of search results. The threshold value is the quantity of content which is directly related to the content of the definition of value target. The value in the coordinate space in the graph is the corresponding value of the propagating diffusivity corresponding to the region in which it is located.

B. Feature Words Library

Although illegal fund-raising advertising has many forms, its essence includes investment profit information, and key words in its advertisement contents will directly contain some characteristic words or similar words with these key words. If mining the previous illegal fund-raising advertising cases, processing feature words, and expanding the synonyms based on the feature word list, we can build a complete and self-learning growth feature library. By comparing the target advertisement with the illegal fund-raising feature word library, the similarity degree is found to determine the possibility of the illegal fund-raising risk of the target advertisement. The process of building a feature word library is shown as shown in the diagram.

1) Collecting cases of illegal fund-raising advertising

As shown in the table I, we collect illegal fund raising cases as the data source of the characteristic words. The illegal fund raising advertisements use some normal financial marketing words such as understanding, lending and agency as fundraising pretexts to confuse users.

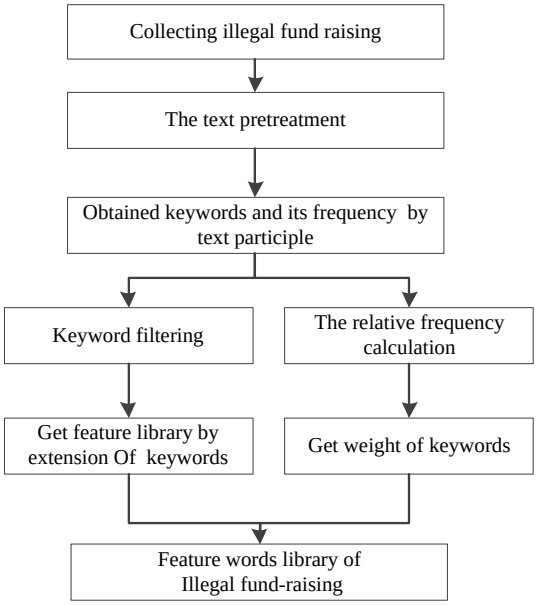

FIGURE II. THE PROCESS TO CONTRUST FEATURE WORDS LIBRARY.

TABLE I. CASE OF ADVERTISING TEXT

No.	Cases features words
1	"unsecured, unsecured credit", and promises low interest, the procedures for borrowing and lending are simple and quick
2	have many trust companies and Asset Management Co, have established long-term cooperation with many banks, hundreds of partners, released free loan information
3	"micro enterprise loan", "small enterprises loan", "city speed loan", "the same city personal consumer loan", "the same city personal credit loan"
4	"credit loans", "bank loans", "private loans", etc.
5	high return, investment 10%-25%, agent at all levels
6	Monthly interest rate of 2%~3%, signing the "membership agreement", "marketing agreement" and "chain marketing agreement"
7	Investment advisory business, financial consultation, loan consultation, customer service financing and financial activities advertising, paying interest or dividends or giving regular distribution of physical objects and other financing activities.
n	….

2) Text preprocessing and word segmentation

Because there are many HTML label, informal text content, hyperlinks, Latex symbols, brackets, the use of ## expression way of topic in the content from the data collected by crawler. We should delete them. Chinese text is separated by punctuation sentences, without any spaces between each word. For Chinese segmentation, the technology is relatively mature. For general text segmentation accuracy rate is very high, for the special field or special sentence, can use the way of adding thesaurus to find the new words.

3) Feature words extraction

Remove the stop words such as "this", "and", and punctuation marks, that we don't want to introduce them to analyze text. The amount of commonly used words in Chinese stop word list is about 1208. Feature words extraction is the process of annotated words, including nouns, verbs, and adjectives and so on. The result of the Chinese word segmentation will include the result of this part of speech tagging, excluding the discontinuation words, pronouns, adjectives, etc., and keep the nouns and verbs. According to the synonym library, the feature word list can be expanded based on the key words.

4) *Computing weight value of feature words*

According to the number of cases appearing in the sample library, the weight of the feature words is calculated. If the words have more occurrences number, the weight value they will be set. The weight of the feature words is defined by frequency as computed as follow. Kfi is the number of the key words I, and N is the number of cases.

$$W_i = Kf_i / n \tag{4}$$

5) Setting up a feature word library

The feature words collection of illegal fund-raising is to transform the feature words of each case into the vector space model. Each feature word is a tuple composed of the word and its weight. The form is as follow, ($kw_1 \ldots kw_n$) is the list of the key words and their extended synonym combination, *W* is the weight of the keyword.

$$K = < (KW_1 \ldots Kw_n), W >$$

The expression of the feature word library is $KL = (K_1, K_2 \ldots .K_n)$

IV. EXPERIMENTAL ANALYSIS

We collect 100 advertisement cases of illegal fund-raising. 80 cases were selected as sample set, which was processed by text preprocessing and characterization, and then the illegal collection of feature words library was constructed. The other 20 are as the test cases. We also collect 20 normal network advertisement cases. So there are 40 network advertisement text as a test set. In the experiment, the precision rate P and recall rate R were used as experimental evaluation indicators, and the comprehensive evaluation index F value was selected as the harmonic mean of P and R to evaluate the accuracy of the evaluation analysis. The definition of these indicators is as follows: the correct number of illegal fund-raising advertisements is N_{pre}, which is judged by the system as illegal fund raising. The number of advertisements is N_{all}. The number of illegal fund-raising advertisements which is should be judged as illegal fund-raising is N_{ori}.

$$P = N_{pre} / N_{all} \tag{5}$$

$$R = N_{pre} / N_{ori} \tag{6}$$

$$F= 2*P*R/ (P+R) \quad (7)$$

1) Risk computation of illegal fund-raising

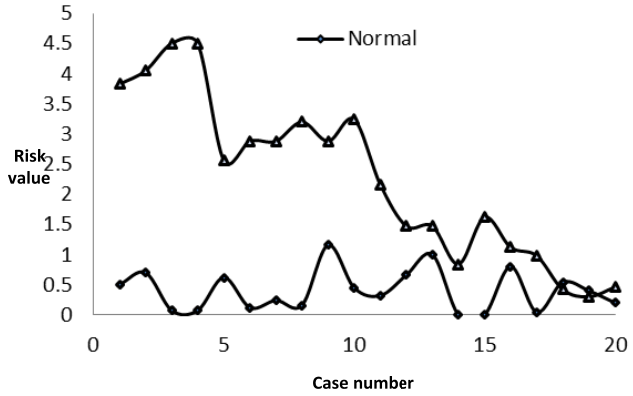

FIGURE III. RISK RESULT FOR EACH CASE.

Calculate the risk value of illegal normal sample through text mining on all of the advertisement content in test sample set, the results were shown as shown in the figure. There is a clearly distinction between most of the risk results of illegal fund-raising case and the normal case. But there is also some risk value is relatively close to the each other. It may be caused by the illegal fund-raising publicity is hidden, scarcely take propaganda, and avoiding the sensitive word or not directly in their advertisement. Some normal case is possible to take nonstandard advertisement in order to attract users.

2) Threshold effect on each indicator

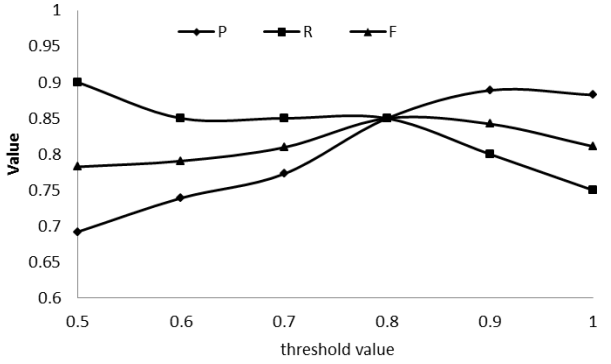

FIGURE IV. INDICATOR VALUE FOR EACH THRESHOLD.

The risk threshold *THm* is set up to distinguish between two cases. When the threshold changes within a certain range, the corresponding *P*, *R*, and *F* index changes are shown in Figure 3. If change the threshold, the precision rate P has been increased, but the recall rate *R* has been declined. The threshold value changed from 0.5 to 0.8, *P* is more effective than *R*. a small threshold value always easy to obtain higher recall ratio. On the contrary, the threshold value changed from 0.8 to 0.1, *R* has advantage than *P*. When the threshold reaches 0.8, harmonic index reached the optimum. *P* and *R* reached 85%. It is a relatively good performance under the condition. Only a small amount of data analysis is done in the experiment, the data may be one-sided, but it can also prove the effectiveness of the aspect to a certain extent. In the early stage of Internet illegal fund-raising, the monitoring method is put forward, and it also helps to suppress the crime.

V. CONCLUSION

The Internet is a double-edged sword, which provides great convenience to our lives and is also a tool for illegal crimes to seek personal gain. Illegal fund-raising is a social problem. At the era of network, the concept of "everything can be quantified", which indicated that crime can be monitored and quantified in each stage. It makes early identification and management of risks of illegal fund-raising become possible. For the advertisement of illegal fund-raising behavior in the early stage, the content mining can assist the regulatory authorities to quickly and effectively investigate the the relevant departments and personnel. It is useful to crack down on illegal fund-raising in the early time.

ACKNOWLEDGMENT

This research was financially supported by Open Fund Project (NO.JXJZXTCX-018), in Collaborative Innovation Center for Economics Crime Investigation and Prevention Technology, Jiangxi Province.

REFERENCES

[1] Meng M., & Jiuhong Y. (2016). Research on gray system and quantitative model about the fund-raising ability to china private equity fund. *Journal of Shanghai Jiaotong University, 21*(3), 365-369.

[2] Bao, Y. (2016). Risk investigation and analysis of p2p network lending platform derived mode based on financial stability perspective. *West China Finance*.

[3] Meng, M., & Jiuhong, Y. U. (2016). Research on gray system and quantitative model about the fund-raising ability to china private equity fund. *Journal of Shanghai Jiaotong University*, 21(3), 365-369.

[4] Redmond, M. (2016). From “intrusive” and “excessive” to financially abusive? charitable and religious fund-raising amongst vulnerable older people. *Journal of Adult Protection*, 18(2), 86-95.

[5] Wadesango, N. (2014). Extent of teacher participation in school based fund raising activities. *Anthropologist, 17*(2), 319-325.

[6] Hosseinkhani, J., Chuprat, S., Taherdoost, H., & Moghaddam, A. S. (2013). Propose a framework for criminal mining by web structure and content mining. *Social Science Electronic Publishing, 1*, 1--13.

[7] Johnson, F., & Kumar Gupta, S. (2013). Web content mining techniques: a survey. *International Journal of Computer Applications, 47*(11), 44-50.

[8] Költringer, C., Dickinger, A., Martin, D., Rosenbaum, M., & Ham, S. (2015). Analyzing destination branding and image from online sources: a web content mining approach. *Journal of Business Research, 68*(9), 1836-1843.

[9] Arbelaitz, O., Gurrutxaga, I., Lojo, A., Muguerza, J., Pérez, J. M., & Perona, I. (2013). Web usage and content mining to extract knowledge for modelling the users of the bidasoa turismo website and to adapt it. *Expert Systems with Applications, 40*(18), 7478-7491.

[10] Strok, F. (2014). Modeling text similarity with parse thickets . *Procedia Computer Science, 31*, 1012-1021.