Assessment Methods of Credit Risk of Agencies Based on Web Data

Xin-Yu Ma, Jian-Yun Shang,
Qian Zhu, Hua-Ping Zhang
Beijing Institute of Technology,
Beijing, China
E-mail: maxinyu2014@nlpir.org, sjy_w22@163.com
zhuqian2013@nlpir.org, Kevinzhang@bit.edu.cn

Yue-Ying He, Zhong-Hua Zhao
National Computer Network and Information Security
Management Center,
Beijing, China
E-mail: hyy@cert.org.cn, hyy@cert.org.cn

Abstract. Most of the existing studies in credit risk assessment of agencies are based on financial data, on the one hand which can't fully reflect the credit status of the agencies, and on the other hand the financial data of some agencies are difficult to obtain. In order to make up for the deficiency of the traditional assessment methods, this paper attempts to utilize Web credit data to assess the credit risk of agencies. Combined with text mining technology and sentiment analysis theories, a credit risk assessment of agencies model which is based on the comprehensive calculation results of sentence sentiment tendency value and evaluation words sentiment tendency value of agencies' Web data set is proposed in this study. To test the model, we grab four agencies' Web data and utilize the model to assess the credit risk of the four agencies. The experiment results are basically in accordance with the actual credit risk status of agencies, which indicates the feasibility of the model.

Keywords: Credit risk assessment, Web data, Sentiment analysis, Sentence tendency analysis.

I. INTRODUCTION

Credit risk, also called default risk, is the risk that the counterparty fails to fulfill the obligations under the contract and causes economic losses [1], and is one of the most important risks faced by China's credit institutions. With the development of China's economy, especially the rapid rise of the Internet finance, the proportion of credit transactions in a variety of commercial transactions is increasing and dishonesty phenomena is becoming more and more serious. In order to prevent credit risk, expand credit transactions, reduce transaction costs, solve the problems of information asymmetry in the transaction process [2], establishing a sound credit risk assessment system is very necessary and is also the request of social credit system construction.

Effective credit risk assessment mechanism helps not only to reduce the non-performing assets of credit agencies, but also to improve the agencies' management level of credit risk. Most of the existing studies in credit risk assessment of agencies are based on financial data [3-5], which can't do a multidimensional and comprehensive credit risk assessment of agencies. In addition, financial data is commercial privacy for most agencies, which is difficult to obtain for other agencies and individuals, resulting in information asymmetry and making the credit risk assessment model based on financial data is limited in assessing the credit risk of agencies. However, with the rapid development of the Internet, and particularly the implementation of "government information disclosure regulations", making it more easily to access agencies' credit data via the Internet, and increasing the dimension of credit data. The emerging of Internet credit data can be used as a supplement to the traditional credit data and make up for the deficiency of traditional credit risk assessment model. Utilizing data mining technology to construct the mechanism of credit risk assessment model based on Web data has become the mainstream trend of future credit assessment.

II. METHOD FOR GRABBING WEB DATA OF AGENCIES

Web credit data acquisition is the basis of credit risk assessment. Due to the dispersion and big volume of Web data, and is difficult to grab one by one. Therefore the study mainly takes the Baidu search results as the entrance to grab data related to agencies, such as news, Baidu know, Baidu Encyclopedia, post comments, etc. These data is relatively concentrated, with strong authenticity, and is worth to analyze.

In order to grab the WEB credit data of the agencies, a topical crawler based on Baidu search is designed in this study [6, 7], which is shown in Fig. 1. Get the URLs related to the topic through the Baidu search, and then use the Web link structure and content evaluation search strategy, utilizing the first in first out principle to determine the access sequence of links. Web search is an iterative process, firstly, crawler starts from a topic seed link and returns the relevant URL through the Baidu search engine, the HTTP protocol request and download the web and then pre-processor parses the web pages to extract URLs. And last parses the URLs sequentially according to the priority order returned through the search and then extracting the web page text [8, 9].

Credit risk of agencies should be analyzed from different aspects, such as products, industry, human resources, financial status, customer sales, technology development, market reputation, etc. Therefore, after grabbing the data of agencies, relevant data of agencies under different topics should also be grabbed, so a "topic+ keywords" iterative grabbing mode is designed in this paper, namely when setting the crawler topics, by the way of "organization name + keywords" in order to comprehensively collecting the credit data of agencies.

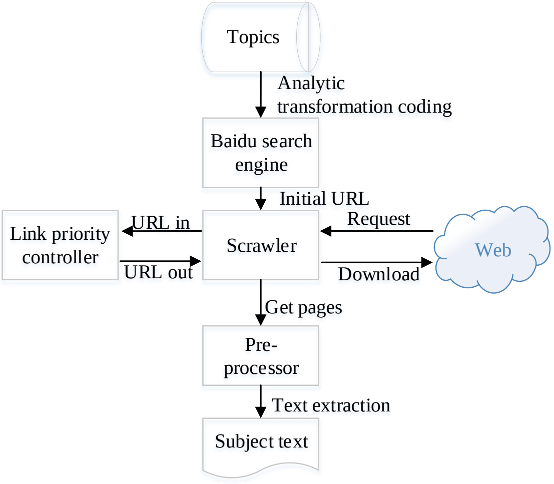

Figure.1. Principle of the topic crawler based on Baidu search

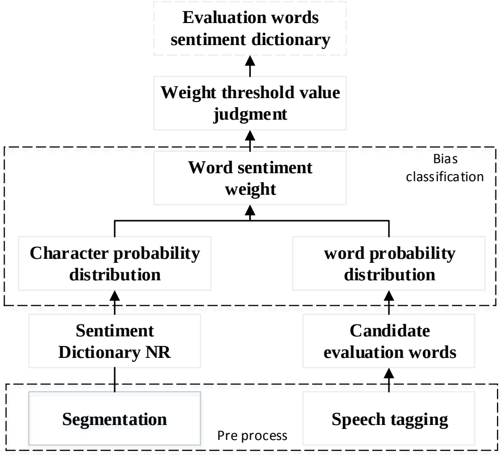
Evaluation words sentiment dictionary
Weight threshold value judgment
Word sentiment weight
Bias classification
Character probability distribution
word probability distribution
Sentiment Dictionary NR
Candidate evaluation words
Segmentation
Pre process
Speech tagging

Figure.2. Process of constructing evaluation words dictionary

III. CREDIT RISK ASSESSMENT ALGORITHM

Credit risk assessment of agencies in this study mainly achieved by computing the values of sentence sentiment tendency and evaluation words sentiment tendency in the Web data set of particular agency, namely if negative words or phrases that occur more frequently in a agency's data set, such as crisis, disputes, it indicates that there may be problems in the operation or management process of the agency and may be with the high possibility occurrence of credit risk.

A. *Construction of Evaluation Words Sentiment Dictionary*

Sentiment Analysis, also known as Sentiment tendency analysis or opinion mining, is a kind of method to get sentiment tendency or value attitude of authors from the text content [10]. It is the process of analyzing, processing, inducing and inference of the subjective text with sentiment tendency. The main work of sentiment analysis in this study is those extract evaluation words from the data set and then judge polarity and calculate the sentiment value of evaluation words. Based on a large number of previous studies, the extraction and discrimination of evaluation words is often an integrative work, which mainly includes corpus-based and dictionary- based methods. In this study, we choose the popular dictionary- based method [11, 12].

In existing studies, sentiment dictionaries are constructed through microblogging or specific corpus, is not suitable for our study, so in this paper an evaluation words sentiment dictionary of agencies is constructed based on existing NTUSD, CNKI.Net sentiment analysis dictionary and affective lexicon ontology. First, merge the three existing sentiment dictionaries, and perform word matching in existing Web data corpus, filtering out repetition, colloquial, low frequency, unsuited sentiment words in this study, and the rest of the sentiment words will to be used as evaluation words sentiment dictionary which called NR dictionary in this paper. Next, select the adjectives and verbs in the corpus as the candidate evaluation words to calculate its sentiment value, the value exceeds the threshold value will be joined into the evaluation dictionary, and constantly enrich the evaluation dictionary by iteration [13], as shown in Figure. 2.

Bias classification algorithm is used to calculate the sentiment value of evaluation words [14]. First calculate the frequency of candidate evaluation words w_i in the corpus, as shown in equation 1. And then calculate the probability distribution $P(z_i/c_i)$ of the character z_i in candidate evaluation words in the NR dictionary, as shown in equation 2. And finally calculation the sentiment weight of candidate evaluation words used the equation shown as 3 and 4. Because there are positive and negative sentiment words in the NR dictionary, the sentiment weight of evaluation words also contains positive and negative sentiment weight which is equal to the difference between the positive and negative sentiment weight. When the weight value is greater than 0, indicates a positive word, otherwise a negative word.

$$P(w_i) = \frac{C(w_i)}{\sum_{i=1}^{n} C(w_i)} \tag{1}$$

$$P(z_i/c_i) = \frac{P(z_i)}{P(c_i)}\,. \tag{2}$$

$$\begin{aligned} S &= \arg\max P(c/z_1..z_2...z_n) \\ &\approx \arg\max P(c, z_1..z_2...z_n) \\ &= \arg\max P(z_1..z_2...z_n/c)P(w) \\ &\approx \arg\max \prod P(z_i/c)P(w) \end{aligned} \tag{3}$$

$$S = S_p - S_n\,. \tag{4}$$

In order to assess the credit risk of agencies from different aspects, LDA topic model is used to do word clustering, and finally by artificial summary we classify the evaluation words of agencies into the aspects of security and fraud, the advantages, disadvantages, etc. Shown as in the TABLE 1.

TABLE I. EVALUATION WORDS SUMMARY IN DIFFERENT ASPECTS

Aspects	**Evaluation Words Example**
Security	safety, health, solve、perfect, experience, management, ability, comprehensive
Fraud	fraud, deception, fraud, deception, fraud, murder, gang, emergency
Advantages	best, advantage, win, energy saving, independent, strong, healthy, perfect
Disadvantages	disputes, competition, challenge, fall into, resignation, loss, unknown, problems
Innovation	innovation, autonomy, the latest, entrepreneurship, success, excellence, advanced
Illegal activities	bad, illegal, fraud, counterfeiting, pollution, recall, violation, error
Development	development, construction, growth, production, breakthrough, cooperation
Difficulties	problems, blows, difficulties, threats, crises, problems, resistance, even worse

B. Sentence sentiment tendency algorithm

The sentiment tendency of sentence, referred to ET in this paper, is decided by the number of evaluation words in the sentence. If the number of positive words is bigger than the number of negative words in the sentence, then the tendency of the sentence is positive, otherwise is negative. The calculation method is as equation 5, 1 represents the positive and -1 represents the negative. Due to the syntactic structure will affect tendency of sentences, in this study, we evaluate the positive and negative tendency of sentence by processing negative words, double negative words, collocation evaluation words, and interrogative words, turning structure in the sentence.

$$ET(sentence)=\begin{cases}1,\ \sum\limits_{\mathrm{i}=0}^{n}ET(word)>0\\ \text{-}1,\sum\limits_{\mathrm{i}=0}^{n}ET(word)<0\end{cases}. \qquad (5)$$

C. Processing of negative words

Generally, evaluation words itself reflect the tendency of the sentence, such as "happy" is a positive word, but there is often qualifiers ingredients in a sentence, such as "is not happy" or "is not very happy", the tendency of the sentence could be changed. In view of this situation, the study selects the negative words in CNKI.Net affective lexicon ontology to assist the judgment of sentence tendency. Combine these negative words and evaluation words, if there is a negative word in the front of evaluation word then change the polarity of the evaluation word, and as the method shown in equation 6. If there are two negative words in a sentence, then they are called double negative words. First determine the polarity of evaluation word in the sentence and then analyze the tendency of the sentence which contains double negative word. If there is two negative words, the sentence is considered to be in a double negative structure and it's an affirmative sentence and as the method shown in equation 7.

$$ET(word)=-ET(word)\,. \qquad (6)$$

$$ET(word)=ET(word)\,. \qquad (7)$$

D. Processing of double negative interrogative words

The tendency of interrogative sentence, rhetorical sentence and rhetorical sentence is opposite against its literal tendency. For example: "Do you think he is a good man?" The real meaning is "I think he's a bad man." So before calculating the tendency of a sentence, judge if there is interrogative structure in the sentence, if it contains interrogative structure, the semantics is opposite. The method is shown in equation 8.

$$ET(word)=\begin{cases}ET(word), Iword\notin sentence\\ -ET(word), Iword\in sentence\end{cases} \qquad (8)$$

E. Processing of turning structure

In the turning structure sentence, usually the tendency is opposite between the former expression and the latter expression, while the latter is emphasized. For example: "Although the advertisement is very moving, but the product is very poor." In a sentence with a turning sentence, the evaluation word tendency of the latter sentence is double treated and the tendency of the former sentence is unchanged. The method is shown in equation 9.

$$ET(word)=\begin{cases}ET(word), Bword\in sentence\\ 2ET(word), Eword\in sentence\end{cases}. \qquad (9)$$

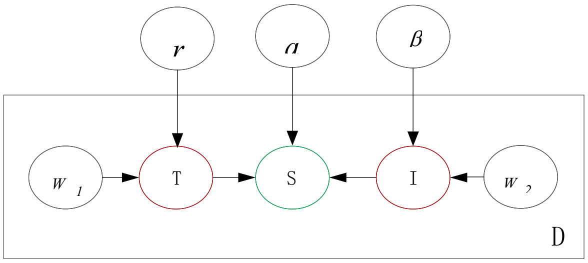

Figure.3. Credit risk assessment model of agencies

F. *Credit risk assessment model*

Credit risk assessment model of agencies is shown in figure 3. D represents for all the text data of an agency, S represents for the total score of D, that is, the credit risk value of an agency. T is the tendency value of all the sentences in D and I am tendency value of all the evaluation words in D, α, β and r are adjusting parameters. The calculations of T and I are respectively shown in equation 10 and 11.

$$T=\sum_{j=0}^{m}\sum_{i=0}^{n}\frac{\text{-}C(ns_i)}{C(s_i)} \qquad (10)$$

$C(ns_i)$ represents the number of negative sentences and $C(s_i)$ represents the total number of sentences in the data.

$$I=\sum_{j=0}^{m}\sum_{i=0}^{n}TF(w_{2i})\times V(w_{2i})=\sum_{j=0}^{m}\sum_{i=0}^{n}\frac{C(w_{2i})}{C(S_i)}\times V(w_{2i})\,. \qquad (11)$$

$TF(w_{2i})$ represents the frequency of evaluation words and $V(w_{2i})$ represents the sentiment value of evaluation words. $TF(w_{2i})$ is equal to the ratio of the frequency of evaluation words W2 to and the number of evaluation words in the text data. And the final credit risk value of agency can be described as equation 12.

$$S=T+\alpha I\,. \qquad (12)$$

In order to reflect the credit risk of agencies more directly, the study carries on qualitative analysis to the credit risk value; five levels are shown in table 2.

TABLE II. QUALITATIVE ANALYSIS STANDARD

Credit Risk Value	Credit Risk Level
>0	Very safety
-30 - 0	Safety
-65 ~ -30	Mild risk
-80 ~ -65	Moderate risk
<-80	Serious risk

IV. EXPERIMENT

In order to test the credit risk assessment method, in this paper, we choose four agencies for contrast experiment, they are China public opinion Strategy Research Institute (CPOSRI), United Nations Emergency Rescue Center (UNERC), Red Star Macalline (RSM) and Beijing Institute of Technology (BIT).

First, we use topic scrawler to grab Web data of four agencies from products, industry, human resources, financial status, customer sales, technology development, market reputation on the Internet, about 68.4MB. And separately calculate the sentence sentiment tendency and calculate T value according equation 10, and the risk value comparison chart as Figure. 4.

Second, calculate the evaluation words sentiment tendency and calculate I value according to equation 11, and the risk value comparison chart as below Figure. 5.

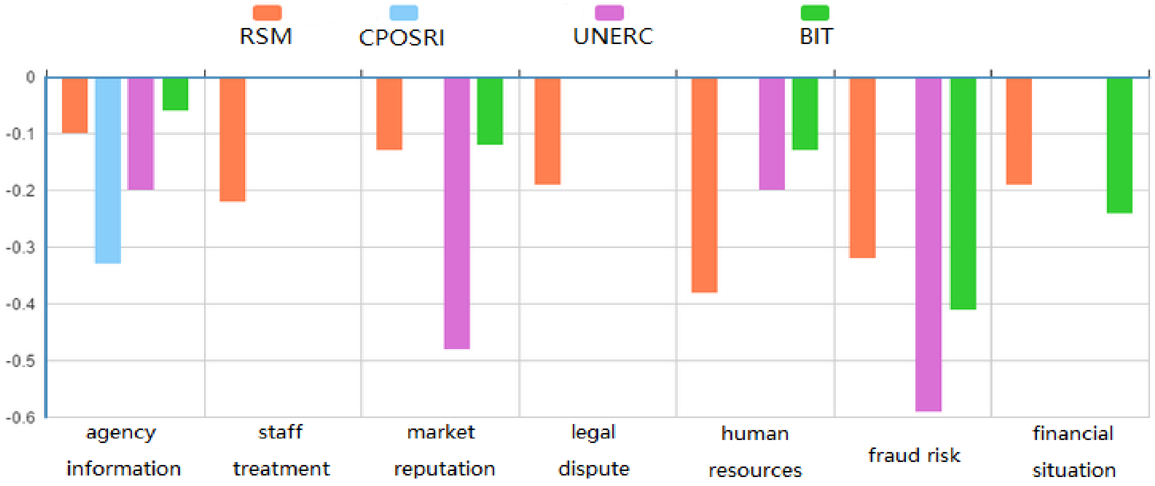

Figure.4. Sentence sentiment tendency value

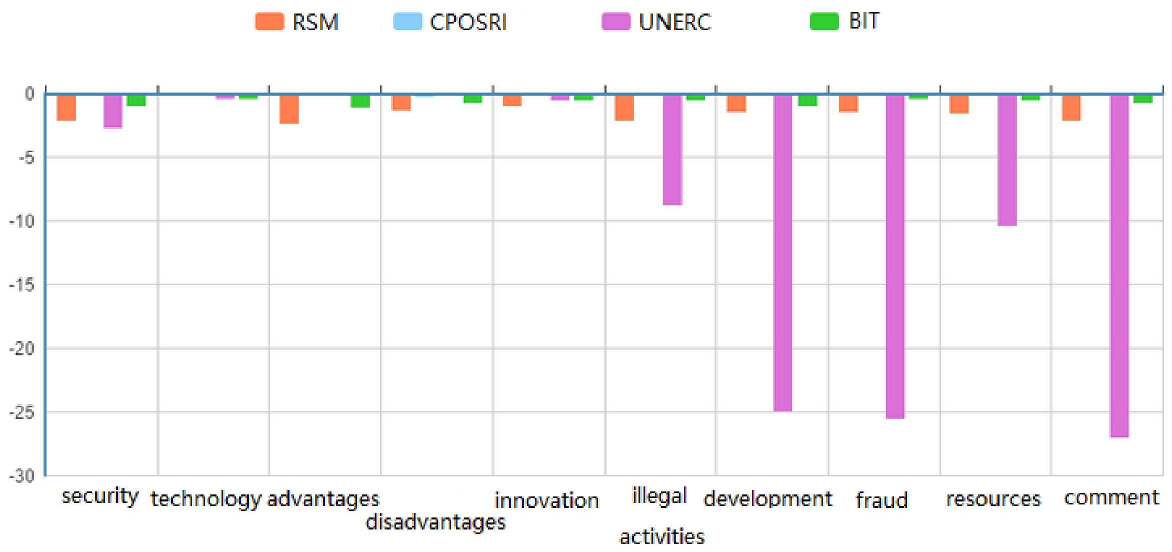

Figure.5. Evaluation word sentiment tendency value

Finally, calculate total credit risk value S, and the result is shown in Table 3.

TABLE III. TOTAL CREDIT RISK VALUE

Agency	T	I	S	Qualitative analysis
RSM	-27.31	-1.53	-42.61	Mild risk
CPOSRI	-45.24	-3.33	-78.54	Moderate risk
UNERC	-75.6	-2.97	-105.3	Serious risk
BIT	-6.93	-1.96	-26.53	Safety

TIPS: α=10; β= -5; r= -0.5.

From the results we can see that the UNERC is with the highest credit risk, and it performed badly in market reputation and Fraud; RSM credit risk in human resources is high and BIT is with the lowest credit risk. By asking the advice of experts, the experiment result and the predicted situation is anastomose, which identifies the rationality of this assessment method.

V. CONCLUSION

In this paper, a credit risk assessment method based on Web data is proposed, by calculating the sentence sentiment tendency value and evaluation words sentiment tendency value in the data set, we propose a credit risk assessment model based on sentiment analysis. By experiment, we find the experiment results are in accordance with the actual situation which indicates the efficiency of the assessment model.

REFERENCES

[1]. Altman, E.I., "Predicting Financial Distress of Companies: Re-visiting the Z-Score and Zeta Models", Working Paper, NYU Salomon Center, June 1995.

[2]. Sha Lin, Jingsheng Lei, "An Empirical Study of DEA Model in listed small and medium enterprise credit risk", Research Management, pp. 158-164, Mar. 2010.

[3]. Xinhan Hu, Wuyi Ye, Boqi miu. "Variable selection in credit risk analysis model of listed company", Mathematical Statistics and Management, pp.1117-1124, Jun. 2012.

[4]. Xiaofan Zou, Jun Yu, Ying Qian, "Research on index system and evaluation method of enterprise credit assessment", Application of Statistics and Management, vol.24, pp. 37-44, Jan. 2005.

[5]. Xiaoli Zhang, Dawei Liu, "Corporate Credit Risk Analysis Method Based on Genetic Algorithms", Enterprise Economy, pp.77-80, Aug. 2012.

[6]. Chakrabarti S, Punera K, Subramanyam M, "Accelerated focused crawling through online relevance feedback", Proceedings of the 11th international conference on World Wide Web. ACM, 2002, pp. 148-159.

[7]. Cui W W, Liu S X, Tan L, et al. "TextFlow: towards better understanding of evolving topics in text", IEEE Transactions on Visualization and Computer Graphics, 2011.

[8]. Chakrabarti S, Van den Berg M, Dom B, "Focused crawling: a new approach to topic-specific Web resource discovery", Computer Networks, vol.31, pp. 1623-1640, Nov. 1999.

[9]. Huaping Zhang, Kai Gao, Heyan Huang, Yanping Zhao, "Big Data Searching and Mining", Science Press, Jan. 2014.

[10]. Yanyan Zhao, Bing Qin, Ting Liu, "Text sentiment analysis", Journal of Software, pp. 1834-1848, Aug. 2010.

[11]. Rosenberg E, Gleit A. Quantitative methods in credit management: a survey [J]. Operations research, vol.42, no.4, 42(4): 589-613, 1994.

[12]. Jiang Wu, Changjie Tang, Taiyong Li, Liang Cui. "Web financial text sentiment analysis based on semantic rules", Computer Application, pp.481-485+495, Feb. 2014.

[13]. Whitelaw C, Garg N, Argamon S, "Using appraisal groups for sentiment analysis", In: Fuhr N, ed. Proc. of the ACM SIGIR Conf. on

Information and Knowledge Management (CIKM). New York: ACM Press, 2005, pp. 625−631.

[14]. Choi Y, Cardie C, “Learning with compositional semantics as structural inference for subsentential sentiment analysis”, Lapata M, Ng HT, eds. Proc. of the EMNLP 2008. Morristown: ACL, 2008, pp. 793−801.