WHICEB 2017 Proceedings

Wuhan International Conference on e-Business

Summer 5-26-2017

Risk Identification of Public Companies Based on Term Cohesion and Topic Visualization

Yiming Zhao
Center for Studies of Information Resources, Wuhan University, Wuhan, 430072, China, zym_0418@qq.com

Afeng Wang
School of Information Management, Wuhan University, Wuhan, 430072, China

Xue Xia
School of Information Management, Wuhan University, Wuhan, 430072, China

Xiangyun Si
Center for Studies of Information Resources, Wuhan University, Wuhan, 430072, China

Recommended Citation
Zhao, Yiming; Wang, Afeng; Xia, Xue; and Si, Xiangyun, "Risk Identification of Public Companies Based on Term Cohesion and Topic Visualization" (2017). *WHICEB 2017 Proceedings*. 18.
http://aisel.aisnet.org/whiceb2017/18

Risk Identification of Public Companies Based on Term Cohesion and Topic Visualization

Yiming Zhao[*1], *Afeng Wang*[2], *Xue Xia*[2], *Xiangyun Si*[1]
(1.Center for Studies of Information Resources, Wuhan University, Wuhan, 430072, China
2. School of Information Management, Wuhan University, Wuhan, 430072, China)

Abstract: The purpose of this article is to develop a procedure to identify risks in public companies based on cohesion relationships among terms and topic visualization. Prospectuses of public companies in the industry of computer implication services in China were collected and chapters of “risk factors” in those prospectuses were analyzed. Texts were split into 10 categories corresponding to different risks by coding subtitles of the texts and 10 sub text sets were formed. Ten categories of risk include market risk, operational risk, financial risk, products and technology risk, investment project risk, internal management risk, inter-control risk, human resources risk, industry risk, and political risk. Five major risks in the ten were visualized to identify topics. After the texts were cleaned and parsed, cohesion relationships among terms were expressed by proximity using cosine value. Relationships among each term and its related terms were characterized and grouped in visual spaces using multidimensional scaling (MDS). Topics were identified by clustering terms in a visual space while each topic corresponds to a specific sub-class of risk. A content analysis was employed to illustrate each topic in the visual space. The procedure to identify risks in public companies in our study enriches the analysis method system of public companies and provides supports for decision-making of the government decision-makers, enterprises’ managers and securities practitioners and the public investors.

Keywords: risk identification, public companies, listed companies, term cohesion, topic visualization, term clustering, text mining, knowledge discovery

1. INTRODUCTION

The public companies have the obligation to disclose information which is related to the production and operation, such as, prospectus, intention of offerings, annual report and announcement on major events and so on. These documents are important text materials that can be used to study the public companies. However, people usually feel confused when facing this huge amount of text information and it is more difficult for them to rapidly extract the desired information and knowledge that they are mostly concerned about.

Generally speaking, people read about 200 to 240 Chinese characters every minute while a prospectus usually has more than 200,000 Chinese characters[1]. This means that people will usually spend 667 minutes on completing the reading of a prospectus with 200,000 Chinese characters. Even if the users only need to get part of the information that they care about, it will also take huge amount of energy. For instance, if a user wants to get the information on public companies in computer application and service industry, he/she needs to read 496,343 Chinese characters of the prospectus and this will take him/her 2068 minutes (equivalent to 34.5 hours). In addition, only reading one time may be not necessary for getting the desired important information.

This paper aims to develop a procedure to analyze risks in public company using term cohesion and topic visualization method. Research questions are: 1) how to identify risks of public companies based on term cohesion in visual spaces; 2) what are the risk factors (topics) in Chinese public companies of computer application services?

This paper will contribute to: 1) enriching the method system of knowledge discovery on public companies, especially on the domain of risk discovery; 2) helping people acquire key information quickly from mass texts

[*] Corresponding author: Yiming Zhao, zym_0418@qq.com.

in an intuitive way; 3) effectively illustrating associative relationships among terms and topics in texts.

2. LITERATURE REVIEW

Term clustering and visualization is one of the methods for knowledge discovery of public companies and the existing studies on knowledge discovery oriented to the public companies include:

2.1 Identification of financial fraud of public companies based on knowledge discovery

Identification of financial fraud of public companies based on knowledge discovery refers to collecting data from the balance sheet, cash flow statement and profit statement among the financial statement disclosed by the public companies to find out whether the public companies have fictitious income, recognize the income ahead of time, confuse the capital expenditure with revenue expenditure to adjust the cost, take advantage of accounting policies and accounting estimates to change or adjust profits and conceal the liabilities, use assets impairment, transactions of the affiliated parties and assets reorganization and other means to whitewash the accounting information[2]. The studies on this aspect mainly are: identification of financial fraud based on Logistic regression analysis[3], [4], identification of financial fraud based on neural network[5], [6], study on risk assessment of false financial report based on decision tree, support vector machine and fuzzy set[7] and so on.

2.2 Studies on identification of credit risk of public companies based on knowledge discovery

Identification of credit risk based on knowledge discovery can timely and accurately send alarms before the public companies have financial distress by building up financial distress prediction mode and continuously digging the valuable information of the public companies. Many scholars think that the essence of studies on enterprises' financial distress is the same as that of studies on risk identification and the only difference lies in the perspective of study. From the perspective of the bank, it is called as identification of credit risk; while from the perspective of the enterprise, it is called as financial distress. The studies on this aspect mainly are: early-warning of financial risks based on neural network[8], [9], analysis of financial early-warning based on decision tree[10], [11] and case-based analysis of financial early-warning[12] and so on.

2.3 Securities investment analysis based on knowledge discovery

Securities investment studies utilizing the financial data in regular reports of the public companies such as annual report and based on data-mining technology and knowledge discovery method mainly are: in the literature[13], stocks are grouped and high-quality stocks are selected on the basis of cluster analysis; in the literature[14], [15], cluster analysis of comprehensive profitability index of public companies on high-tech sector is conducted and thus blue chip and leading stock as well as the preferred investment objects among the high-tech sector are selected; in the literature[16], square sum of deviations is used to implement cluster analysis on fundamental aspects of the 31 randomly selected public companies; in the literature[17], cluster analysis method is used to analyze and classify the 40 public companies in petrochemical sector.

2.4 Existing methods for risk identification of companies

Risk identification is most important in risk management and the existing methods of enterprise's risk identification include risk list analysis method, risk source analysis method, standardized investigation method, financial statement analysis method, flow chart method, cause-effect diagram method, event tree analysis method and fault tree analysis method[18].

2.5 Summary

The existing studies on data-mining and knowledge discovery of public companies all use the structural data information and focus on the financial data mining while little attention to the highly non-structural text information is paid. In the real world, however, text is also an important information carrier. As a matter of fact, some studies show that 80% of the information is contained in the text documents. Therefore, more efforts should be made to have knowledge mining and discovery over text of regular reports of public companies.

3. METHODOLOGY

3.1 Data sources

Data of this paper comes from the text description of prospectus of all public companies in the computer application and service industry on “risk factors”. There were 97 public companies in the computer application and service industry up to September 2013[19], of which, 16 were listed companies at Shanghai Stock Exchange and 81 were listed companies at Shenzhen Stock Exchange. Raw data sets have 496, 343 terms.

3.2 Data cleansing and pre-processing

Prospectus is commercial text with complete format. It has clear chapter and section headings which indicate the theme and represent the core contents of the corresponding paragraphs. In this study, the section headings of all risk factors are extracted and coded to set up risk classification system and the texts under different headings are classified into the corresponding risk types before having visual analysis of the vocabularies in each type. Each type corresponds to a text set with 97 txt files and each txt file corresponds to a company. Text description of risks was contained in those txt files. This can effectively reduce the amount of vocabularies in each type and enhance the visualization effect.

Segmentation strategy of forward maximum matching is selected to segment the initial text set. In the segmentation process, the stop words are filtered out: on the one hand, the function words without independent meanings or including few semantic contents are screened out, such as, the pronoun, the article, the conjunction and the preposition; on the other hand, the commonly-used topic words such as “company (companies)”, “market” and “risk” are screened out.

3.3 Representation of term cohesion and visualization

Cohesion relationships among terms are recorded in term-document matrixes in which each element is the value of the term frequency in corresponding document. As in the equation (1), n represents amount of texts in C_q, k represents amount of word entries in C_q, a_{ij} is the matrix item and value of a_{ij} represents frequency of the word i in text j.

$$M_{Cq} = \begin{pmatrix} a_{11} & .. & .. & a_{1n} \\ a_{21} & .. & .. & a_{2n} \\ & \dots & a_{ij} & \\ a_{k1} & .. & .. & a_{kn} \end{pmatrix} \tag{1}$$

Use absolute frequency representation and use the frequency of the words in the text as the value of component in the text vector and then use the differentiation equation of high frequency words and low frequency words that put forward by Donohue according to Zipf's second law in 1973 to determine the threshold of screening out low frequency words. The equation[20] is shown as follows:

$$Threshold = \frac{-1+\sqrt{1+8I_1}}{2} \tag{2}$$

Of which, I_1 represents the amount of words that only show once in the text.

In the project process, S-Stress and RSQ are used to reflect the project quality and the degree of information loss. When the S-Stress is closer to 0 and the RSQ is closer to 1, it indicates that the project quality is higher and the degree of information loss is lower and further suggests that the clustering effect of topics in MDS space map is satisfactory.

4. RESULTS AND DISSCUSSION

4.1 Analysis of the text set

After data cleansing and pre-processed, the text set in our study were split into 10 categories corresponding to different risks by coding subtitles of the texts and 10 sub text sets were formed. Ten categories of risk including market risk, operational risk, financial risk, products and technology risk, investment project risk, internal management risk, inter-control risk, human resources risk, industry risk, and political risk. According to a former study on the same text set, the top three most important risks in public companies of the industry of computer implication services are: market risk, products and technology risk, and financial risk[21].

4.2 Visual analysis of risk factors in Chinese public companies of computer application services

4.2.1 Market risk

In the category of market risk, the cosine proximity measure was used to represent cohesion relationships among terms and the Minkowski distance measure was employed in the MDS analysis. The stress value was 0 and the corresponding RSQ was equal to 1. Four clusters emerged in the visual space (Figure 1) including 23, 33, 5, and 13 terms respectively.

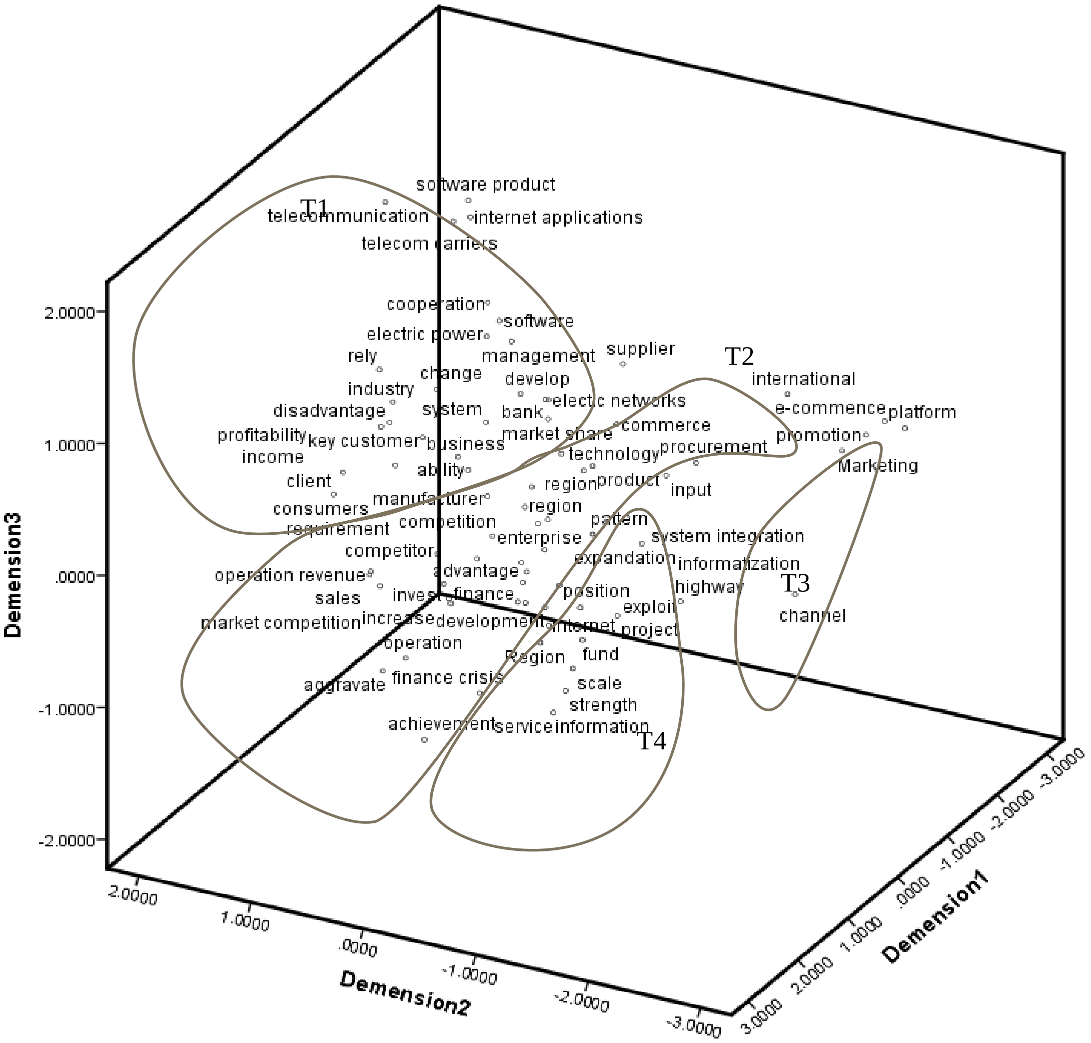

Figure 1. Visual clustering of market-risk-related terms

Topic 1 represents the risk of client-reliance. It results from reliance on the key customers, suppliers, the single market, electricity and telecommunications industry and other monopoly industries.

In terms of reliance on key customers, some listed companies rely heavily on key customers, as the top five customers accounted for most of the main business income, such as Shanghai Hyron Software Co., Ltd., Shenzhen Tianyuan Dic Information Technology Co., Ltd. etc. Some enterprises rely on a single market, such as the real risk of over-reliance on a single market of Shenzhen Das Intellitech Co.,Ltd. in southern China; The excessive reliance on the postal system exists in Hunan Copote Science Technology Co., Ltd.

In terms of reliance on specific industry, listed companies of Computer Application Services industry rely highly on the telecommunications industry, power industry, and banking sector. Business-intelligence Of Oriental Nations Corporation Ltd., for example, in the year 2007, 2008, 2009 and the first half of 2010, the proportion of revenues from the telecommunications industry were 95.55%, 92.55%, 97.31% and 96.43%. In the period of 2009-2011, the current operating income of Sinodata Co., Ltd., from the Agricultural Bank was 36.18%, 46.00% and 41.50%, showing that the company has a considerable degree of dependence on agriculture bank.

Topic 2 mainly reflects the risks of market fluctuation and market competence, including risks such as business cycle, prices, competition in the industry and the financial crisis.

Topic 3 reflects the risks of marketing and promotion. IT Product information service listed companies tend to be high-tech enterprises. High-tech products often face greater competitive pressures than the average products in the promotion of products. As the Chinese market is not mature now, there are some issues like regional division and the difficulties in expanding into new markets .If the company cannot adapt to the change of market competition in the distribution network, marketing strategies in market development, the company’s competitive advantage is likely to be weakened and will face the risk of market share declining, thus affecting the improvement of enterprise product sales.

Topic 4 mainly reflects the market development risks. If a company cannot correctly judge the situation of market and industry trends, and replicate the company’s existing mature business model successfully in other regional markets, it will have the risk of operation decline due to the market competition or falling behind competitors.

4.2.2 Products and technology risk

In the category of products and technology risk, the cosine proximity measure was used to represent cohesion relationships among terms and the Minkowski distance measure was employed in the MDS analysis. The stress value was 0 and the corresponding RSQ was equal to 1. Five topics, which include 11, 8, 29, 9 and 9 terms respectively, are found in the visual space of Figure 2.

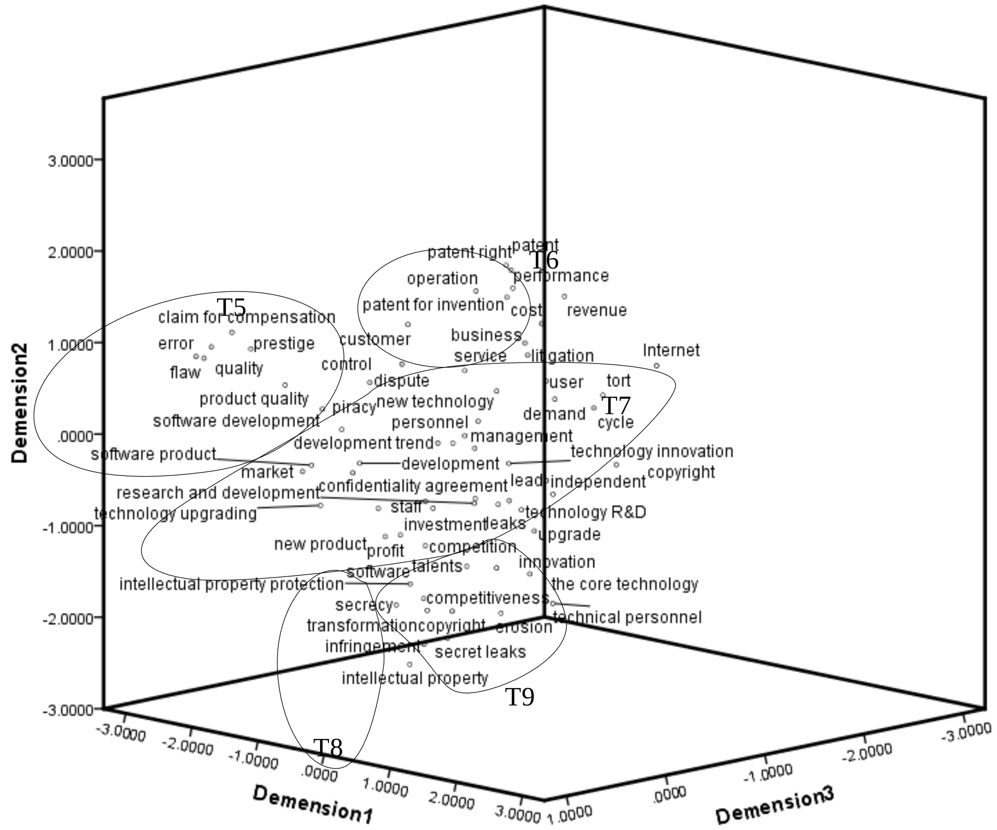

Figure 2. Visual clustering of products-and-technology-risk-related terms

Topic 5 primarily reflects the risk of product development quality. Since the software publisher and companies cannot completely eliminate software errors and defects due to the high complexity of the software, the disputes, claims or lawsuits arising due to quality problems will have a negative impact on the market reputation or market position of all listed companies.

Topic 6 mainly reflects the risk of patent protection. Protection system for the current Chinese patent is not perfect, and the exercise of rights related to litigation and other remedies is costly, time-consuming. To some extent, the company's technology and patents is facing the risk of being violated. If the company's core technologies suffer a wide range of abuse, this will adversely affect the performance of listed companies in the industry.

Topic 7 mainly reflects the risk of technological advances and new products development. In new product development, if the enterprise cannot grasp the key point of products and market and adjust the direction of technology and products in time, or the new products cannot be quickly promoted, there will be a risk for technology and new product development.

Topic 8 mainly reflects the risk of infringement of intellectual property.

Topic 9 mainly reflects the risk of core technology leaking. For knowledge-and technology-intensive enterprises, technology development and business model innovation inevitably rely heavily on experts and core technologies. Therefore, the enterprises must rely on the core technology and technical staff that companies cultivate, introduce and accumulate, in order to maintain a competitive edge on the market. And the loss of core technical staff may lead to the loss or disclosure of core technology, which would affect the company's market competitiveness and technological innovation capability to some extent.

4.2.3 Financial risk

In the category of financial risk, the cosine proximity measure was used to represent cohesion relationships among terms and the Minkowski distance measure was employed in the MDS analysis. The stress value was 0 and the corresponding RSQ was equal to 1. Financial-risk-related terms assembled four topics which include 11, 21, 41 and 23 terms respectively (Figure 3).

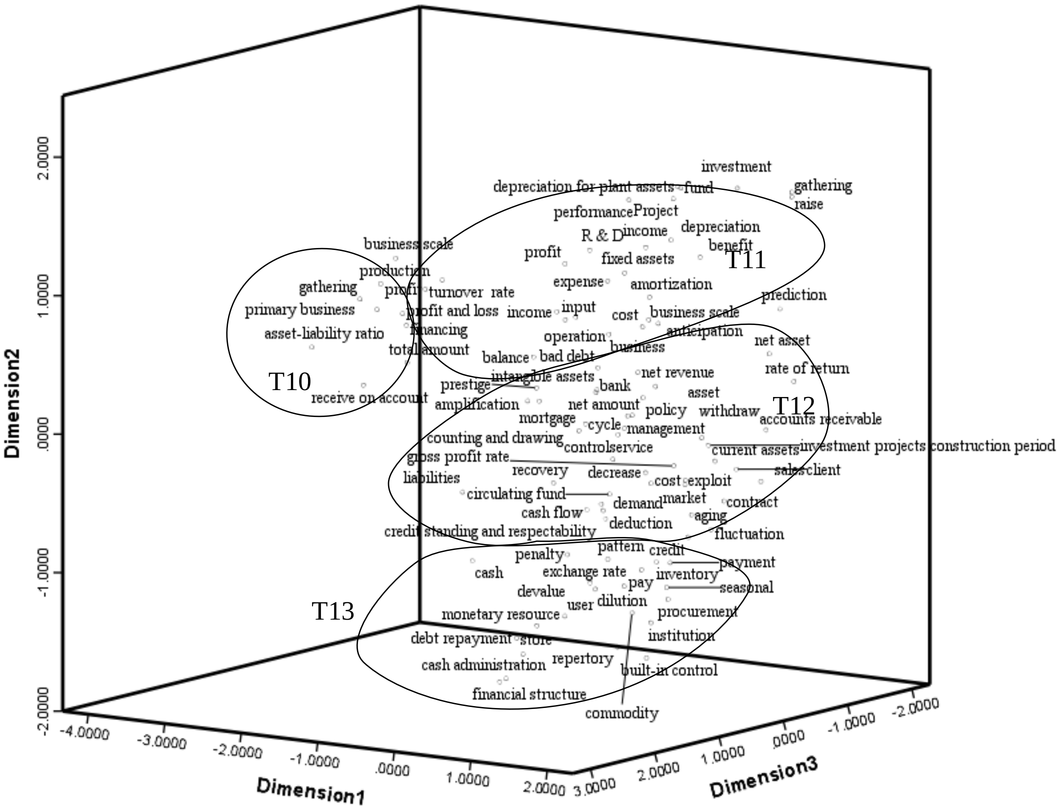

Figure 3. Visual clustering of financial-risk -related terms

Topic 10 mainly reflects the financial risks caused by the change of scale. After the company successfully raised funds through the stock market it will intensify efforts to implement new investment projects and accelerate the expansion of business scale. The debt scale will also expand, asset-liability ratio, asset-liability structure which is irrational and other issues will bring the risk of profit fluctuations in financial companies.

Topic 11 primarily reflects expense amortization and financing risks. In terms of investment amortization, companies will continue to increase investment in fixed assets at the beginning of the listing. In respect of financing, as the particularity of computer application services industry, the company’s assets consist of the smaller scale of fixed assets, the larger of current assets. With the expansion of industry competition and company size, once the company in the course of business is faced with a shortage of funds, it is difficult to obtain bank loans through fixed assets, so there are indirect financing risks to the company’s operation.

Topic 12 mainly reflects the risk of assets and liabilities. Potential topics of this sub category include receivables, ROE, and cash flow aspects.

Topic 13 mainly reflects the risk of internal financial control, financial risks such as revenue, inventory, expense internal control, inventory impairment, and cash management.

4.3 Implications

The procedure to identify risks in public companies in our study enriches the analysis method system of public companies and provides supports for decision-making of the government decision-makers, enterprises’ managers and securities practitioners and the public investors.

As to the government decision-making, the conclusions of this study offer basis for the government to formulate industry policies and polices to regulate the securities market.

As to the enterprise operation, enterprises’ managers have to investigate the industries in which they intend to invest before making decisions on which industry to enter, the enterprise merger and acquisition, risk investment, diversification and strategic cooperation. The traditional industry analysis methods include historical data analysis, survey research method, self-built database, industry comparison studies, industry dimension studies, industry grouping studies and industry valuation studies method. Knowledge discovery and visualization are improvement and supplementary of the industry analysis method system and offer a new perspective and tool to study information of the public companies. As to the managers of the enterprises to be listed, this study can offer enlightenment and suggestions for their preparation of getting listed in the stock market.

As to securities investment, the agency investors usually adopt “top-down” decision-making method for investment which mainly refers to screening out the stocks with performance growth faster than the average market growth by using the three-step analysis method, namely analyzing the macro-economic conditions first and then basic conditions of the industry and finally the basic conditions of the stocks. Such selected stocks usually have the potential of becoming the long-term best performers in the market. If the basic conditions of the industry are good, many business opportunities will also be brought to enterprises in the industry and carefully selecting the stocks with competitive strengths in the industry and holding such stocks for a long-term will be an inevitable choice of the agency investors. The ordinary public investors are lack of investment experiences and channels and methods to get the information on industry development. In this study, the information visualization is used to intuitively and vividly explain the industry risk, current development and development trend of the industry and demonstrate the core characteristics of the whole industry to the researches and publics who do not have profound understanding of the industry, thus providing a channel to rapidly know the industry development.

5. CONCLUSIONS

In this study, the proximity relationship of terms in transposed vector space is used to weigh the clustering characteristics of key terms and is projected into the visual three-dimensional space. The term sets classified by the tree-dimensional space are then used to express the topic. Multi-dimensional scaling analysis (MDS) is used to project the proximity relationship in the transposed vector space and calculate the position coordinates of terms in three-dimensional MDS space map. Under the premise of remaining the original topological structure as possible as practical, the terms and proximity relationship of terms, namely the clustering relationship, are expressed in the visual three-dimensional MDS space map.

One limitation of this paper is the bias of its data source. The data is collected only from public companies in one industry in Chinese. In future work, the author will investigate risks in different industry worldwide and improve the method and procedure proposed in this paper.

ACKNOWLEDGEMENT

This research was supported by the National Natural Science Foundation of China under Grant 71403190 and 71420107026. It is also supported by Research Fund for Academic Team of Young Scholars at Wuhan University (Big Data and Business Analytics, Whu2016013).

REFERENCES

[1] Buzan T. (2009). Speed Reading. Beijing: China Critic Press, 25.

[2] Zhou F. (2010). Review on the data mining models and methods to identify the fraudulent financial statements. Accounting and Finance, (3): 39-43.(in Chinese)

[3] Dechow P M, Hutton A P, Kim J H, Sloan R G. (2012). Detecting earnings management: a new approach. Journal of Accounting Research, 50(2): 275-334.

[4] Chen T. (2012). Analysis on accrual-based models in detecting earnings management. Lingnan Journal of Banking Finance and Economics, 2(1): 5.

[5] Chen H J, Huang S Y, Kuo C L. (2009). Using the artificial neural network to predict fraud litigation: Some empirical evidence from emerging markets. Expert Systems with Applications, 36(2): 1478-1484.

[6] Khan A U S, Akhtar N, Qureshi M N. (2014). Real-time credit-card fraud detection using artificial neural network tuned by simulated annealing algorithm. Proceedings of International Conference on Recent Trends in Information. Telecommunication and Computing, ITC. 113-21.

[7] Kirkos E, Spathis C, Manolopoulos Y. (2008). Support vector machines, decision trees and neural networks for auditor selection. Journal of Computational Methods in Sciences and Engineering, 8(3): 213-224.

[8] Lee S, Wu S C. (2013). A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis. Expert Systems with Applications, 40(8): 2941-2946.

[9] Jeong C, Min J H, Kim M S. (2012). A tuning method for the architecture of neural network models incorporating GAM and GA as applied to bankruptcy prediction. Expert Systems with Applications, 39(3): 3650-3658.

[10] Olson D L, Delen D, Meng Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. Decision Support Systems, 52(2): 464-473.

[11] Kwak W, Shi Y, Kou G. (2012). Bankruptcy prediction for Korean firms after the 1997 financial crisis: using a multiple criteria linear programming data mining approach. Review of Quantitative Finance and Accounting, 38(4): 441-453.

[12] Liu B, Sheng Z. (2003). Case-based reasoning for cheat risks diagnosis. Control and Decision, 18(4): 494-496.(in Chinese)

[13] Da Costa Jr N, Cunha J, Da Silva S. (2005). Stock selection based on cluster analysis. Economics Bulletin, 13(1): 1-9.

[14] Nanda S R, Mahanty B, Tiwari M K. (2010). Clustering Indian stock market data for portfolio management. Expert

Systems with Applications, 37(12): 8793-8798.

[15] Tsai C F. (2014). Combining cluster analysis with classifier ensembles to predict financial distress. Information Fusion, 16(1): 46-58.

[16] Li M, Li H. (2006). Application of cluster analysis in security investment fundamental analysis. Journal of Liaoning Normal University, (2): 145-146.(in Chinese)

[17] Li Q. (2005). Financial performance evaluation and cluster analysis of listed companies. Journal of Industrial Technological Economics, 24(8):146-148.(in Chinese)

[18] Xu J. (2016). Risk Management. 4th ed. Shanghai: Shanghai University of Finance and Economics Press.(in Chinese)

[19] Sina Finance. (2012). http://vip.stock.finance.sina.com.cn/mkt/#hangye_ZG87，2016-09-30.

[20] Donohue J C. (1973). Understanding Scientific Literature: A Bibliographic Approach. Cambridge: MIT Press.

[21] Zhao Y, Cheng B, Wang X. (2014). Knowledge Discovery in Texts Oriented to Specific Domain: Risk Classification Schema and Relationship identification of Risks in Public Companies. Journal of Intelligence. 3, 165-170.(in Chinese)