Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 162 (2019) 102–110

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Information Technology and Quantitative Management (ITQM 2019)

# Applying Clustering and Co-occurrence Methods to Identifying Key Events and Their Relations in Chinese Stock Market

WU Jiachen[a], WANG Yaling[a], WANG Yue[a,*]

[a]School of Information, Central University of Finance and Economics, Beijing 100081, China

**Abstract**

Based on needs to extract effective information from massive financial news, this paper[†] uses text clustering to discover events and designs a news event correlation network model based on word vectors. We collect financial news of the whole year of 2017 and discover key news events automatically through text clustering. The extracted event keywords were modeled based on word co-occurrence analysis. The cosine distance of the word vector was used to measure the correlation between keywords, so that the model could contain more semantic information. Finally, an association network of financial news events was constructed to analyze the correlation between events. Preliminary experiments show that this method can extract news events objectively and measure the correlation between event pairs effectively.


Peer-review under responsibility of the scientific committee of the 7th International Conference on Information Technology and Quantitative Management (ITQM 2019)

*Keywords:* Event Detection; Word Embedding; Event correlation; Co-occurrence Analysis

## 1. Introduction

Today, with the rapid development of Internet and information retrieval technologies, as in many other fields, investors in the financial investment field have become accustomed to obtaining relevant market information by browsing reports on the Internet. The information contained in these news items influences investors' judgments on market conditions and decisions on investment. A great deal of research has shown that major news events have an immediate and measurable impact on the movement of related stock prices. [1]

Compared with traditional media, the speed and amount of online news has increased dramatically, making it

* Corresponding author. Tel.:+86-15010736698.
*E-mail address:* yuelwang@163.com.

[†]*This paper is supported by Central University of Finance and Economics 2019 First-class Discipline Construction Project Funding (Key Technologies and Application Research of Independent Controllable Block Chain), National Key Research Program of China (2016YFB1001101) and National Natural Science Foundation of China (61309030).*


10.1016/j.procs.2019.11.264

increasingly unrealistic for manual methods to collect, process and analyze. In order to summarize the useful information in financial news more easily and intuitively, it is necessary to find news events from the vast news texts. The clustering method can form multiple newsgroups in a large amount of news collected, and each news cluster represents one news event. We can judge the importance of the event based on the number of incident reports in the same length of time. At the same time, the occurrence of natural events is not isolated. Different news events often have interactions and may have a knock-on effect on stock market-.

It can be seen that the collection and mining of online financial news and the extraction of major news events are of great significance in identifying market sentiment, analyzing market conditions and studying market changes. Then, the correlation identification of the extracted news events has important guiding significance and auxiliary role in studying the influence of news on stock market fluctuations.

## 2. Related Work

Event discovery, that is, the identification of news key events is usually realized by text clustering, and can be divided into two categories: one is to find key events from a fixed-scale news set, including K-means clustering [2], hierarchical clustering, agglomerative clustering and many other methods; the other is to discover new hotspot events from the incremental news stream, which can be used for real-time topic recommendations on the Internet, such as online incremental clustering proposed by Allen et al [3]. On this basis, many improved models have been developed, for example, recalculating eigenvalue weights [4]; document representation based on word vectors [5]; and combining multiple clustering models [6]. Nowadays, more and more scholars are able to get rid of the dependence on artificially produced features by developing models based on multiple neural networks to achieve fully automatic detection of events [7].

There are many ways to identify various relationships between events. We can use elementary analysis, pattern matching and machine learning-based methods to identify temporal or causal relationships between events. However, there are few related studies to identify event correlation from the perspective of semantics. The co-word network model proposed by Jacobs regards words as the basic nodes in the word network, and describes the sentences, fragments, articles and themes through the combination of words [8]. Many scholars use this method to explore the relationship between topics [9]. Therefore, based on the vocabulary field of the event, the correlation between events can be mined. Shibata et al. used the co-occurrence method of predicate parameters to extract events [10]. Zhang Hui et al. proposed that event correlation can be expressed by the co-occurrence probability of some keywords [11], and the Chinese news event correlation model is realized. Lim et al [12] uses the chapter as the statistical window to judge the association by measuring the similarity of the word text vector.

Based on the core algorithm proposed by Zhang Hui [8], this paper collects news from a financial website, and uses an improved K-means clustering algorithm (remove multiple outliers that are far away from each cluster at each iteration) to find key events and corresponding event news sets from the news report stream; adopts natural language processing and text mining method extraction for each news cluster. The event keywords, and the vocabulary field of the event is established. The cosine distance of two word vector is used to measure the correlation between them. The correlation between events is studied by constructing the vocabulary field network of event keywords. Finally, it is proved by preliminary experiments that this method can accurately discover the correlation between key events and events in the A-share market.

## 3. Research Method

### *3.1. Chinese Natural Language Processing*

We collect a large amount of financial news report texts on a certain fixed time period on the Internet, which

are unstructured. Therefore, the text information needs to be pre-processed. Specific steps are as follows:

1) *Information filtering.* Use a text processing program to remove unnecessary information such as header information and meaningless characters, preventing them from interfering with the main information. If the text downloaded by a website has uniform website symbols or advertisements, these also need to be removed.
2) *Chinese participle.* Word segmentation is the first step in Chinese text processing. To ensure the accuracy of financial terminology word segmentation, we have added external dictionaries in finance and related fields to make the system have good domain adaptability.
3) *Remove stop words and perform part-of-speech tagging.* Import common stop words dictionary to filter out non-representative words in news texts, perform part-of-speech tagging, and remove non-informative words such as adverbs, conjunctions, prepositions, and interjections.

### *3.2. Event Detection using VSM and Clustering*

This paper uses text clustering to discover events. We choose the vector space model (VSM) to represent the text object, and each news article is represented by the words in the text and the corresponding weight:

$$V(d) = \{w_1, w_2, \ldots, w_n\} \quad (1)$$

In each dimension of the vector in the formula, $n$ is the total number of characteristic words, and $w_i$ indicates the weight of the $i$-th feature word in the document. The calculation method for $w_i$ used in this paper is word frequency inverse document frequency (TF-IDF). The calculation formula is as follows:

$$w_{id} = tf_{id} * idf_{id} = tf_{id} * log(N/df_d) \quad (2)$$

$tf_{id}$ is the word frequency of $t_i$ in document $d$, $df_d$ is the document frequency of $t_i$.

The K-means algorithm is often used for text clustering, but it relies on the initial value of K and the choice of the initial cluster center, and the "noise" data can cause the result to be offset. Our clustering goal is to find the key news of financial news in a certain period of time. The outliers that deviate too far from the cluster center have no meaning in the result, and also cause the mean deviation of the clustering calculation. Therefore, we choose the improved K-means algorithm to remove multiple outliers that are far away from each cluster cluster at each iteration, which can reduce the sensitivity of the algorithm to the anomaly object and improve the iterative efficiency. The global contour coefficient method is used to determine the clustering effect to determine the K value.

After the clustering is completed, if we want to get $N$ key events in a certain month, we can filter the *TOP-N* clusters with larger data volume as the news set of key events from the results manually. The rest clusters are considered is a collection of other events. After that, the hotspot words in the event news set are extracted as key concepts according to the size of the TF-IDF value. Based on these key concepts, we can manually name the event.

### *3.3. Event Correlation Modeling Based on Word2Vec and Cooccurrence*

The concept of the vocabulary field comes from the semantic field which can show the association of words in the same system. The total vocabulary can divided into large vocabulary within the system according to the difference of semantics. On the basis of these large vocabulary fields, it is divided into several small vocabulary according to the small difference in meaning, until it is divided into single vocabulary.

In the study of this paper, each news event can be represented by a vocabulary field composed of hot words of the news event. The vocabulary fields of different events are connected to each other by public words, forming a larger network of vocabulary fields. Therefore, we can study the correlation between financial news events through co-word analysis. The co-word analysis method defines the co-occurrence of words as different keywords as the keywords of a news episode. The degree of co-occurrence between words is defined as co-

occurrence. Then, the vocabulary field network can be described as an undirected graph, the node is the event keyword, the edge represents the far and near relationship between the two words, and the weight on the side is represented by the word co-occurrence degree.

The specific steps of the event correlation study based on co-word analysis are as follows:

*1) Extract event keywords.*

After text clustering, we have obtained a news episode containing a single event. The first n words with the highest frequency of the document are extracted from the keywords of each document in the corpus as the corpus keywords, which are used to represent the event keywords. According to the results of the extraction, it is possible to consider artificially filtering out common words in the financial and economic fields such as "financial" and "development", and it is also possible to extract the information-poor word dictionary and extract them before extraction.

*2) Establish a co-occurrence matrix.*

An $n \times n$ symmetric matrix is created for each event called a key weight matrix. The element $w(t_i,\ t_j)$ in $\boldsymbol{A}$ is the associated weight of the keywords $t_i$ and $t_j$.

There are many ways to calculate the co-occurrence weight, which is based on the co-occurrence word frequency, including the inclusion index, the similarity index, the Jaccard index, and the Salton index[13]. This paper uses the cosine distance of the word vector to measure the degree of relationship between words and words. First, We use all news texts to train the word vector of the word $t_i$, denoted as $\mathrm{v}(t_i)$. So that the co-occurrence degree contains more semantic information and the calculation is more convenient.

The formula for calculating the cosine similarity of word vectors is:

$$S_{ij} = \frac{v(t_i) \cdot v(t_j)}{|v(t_i)| \cdot |v(t_j)|} \tag{3}$$

$S_{ij}$ represents the cosine distance of the word vector, and $-1 \le S_{ij} \le 1$, $t_i$ represents a word vector, and $|\mathrm{v}(t_i)|$ represents a modulus of the word vector $\mathrm{v}(t_i)$. For the co-occurrence matrix of an event, the corresponding word vector is trained by the text of a single event set.

The associated weight $w(t_i,\ t_j)$ is the result of $S_{ij}$ normalization:

$$w(t_i,\ t_j) = \frac{S_{ij} - S_{min}}{S_{max} - S_{min}} \tag{4}$$

Among them, $S_{max}$ is the maximum value in the $S_{ij}$ obtained, and $S_{min}$ is the minimum value. Considering the time complexity of calculating the co-occurrence matrix, $n$ should not be too large.

*3) Event correlation modeling.*

When the same words are present in the vocabulary of two news events, it is considered that there is a correlation between the two events, and the correlation will increase as the co-occurrence words increase. The method of calculating the event relevance can be expressed as:

$$w(E_i, E_j) = \sum_{t_l \in T_l} w(t_l,\ E_i) \times w(t_l,\ E_j) = \sum_{t_l \in T_l} \left( \sum_{t_k \in E_i, l \neq k} w(t_l,\ t_k) \right) \times \left( \sum_{t_k \in E_j, l \neq k} w(t_l,\ t_k) \right) \tag{5}$$

$w(E_i, E_j)$ represents the correlation index between the event $E_i$ and the event $E_j$, and $T_l$ is a set of co-occurrence words between the event $E_i$ and the event $E_j$. $w(t_l\ , E_i)$ represents the associated weight of the associated word $t_l$ and the event $E_i$[11].

*4) Visualization of event-related networks.*

The event correlation network can be represented as the network diagram shown in Fig 1. The dotted ellipse of the outermost circle in the figure is the vocabulary field of an event. The vocabulary field overlap indicates two events, each node represents an event keyword, the solid node represents a word unique to an event, and the blank node represents a different event.

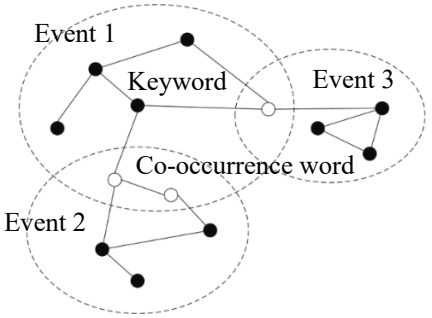


Fig. 1. Schematic diagram of event correlation network

## 4. Empirical Analysis

### *4.1. Experimental Data*

We uses the web crawler program to capture a large number of financial news highlights from the top financial network and the Eastern Fortune Network from January 1, 2017 to December 31, 2017, and to perform de-duplication processing to ensure a comprehensive news source. A total of 37,250 articles, with news release time and news headlines as document names, stored in months.

Due to the large number of events involved, the following experiment took November, which had a significant experimental effect, as an example. A total of 3,136 news texts were collected in that month as the experimental news set.

### *4.2. Event Discovery Experiment and Result Analysis*

This experiment used Python to add the word spliter of the financial dictionary to perform word segmentation, stopping the word and nonsense word, then calculating the TF-IDF value of the word, and using the Python third-party package gensim to build the Word2Vec model.

In the text clustering, set *K* as 5-15 for each experiment, and the contour coefficient results obtained are shown in Fig 2.

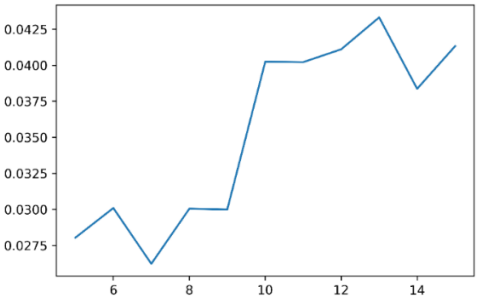


Fig. 2. Polyline graph of clustering contour coefficient under different k values

In the case of large K value, the contour coefficient can show the effect of clustering. Generally speaking, the larger the contour coefficient is, the better the result will be. As can be seen from the figure, when K> 9, the contour coefficient rises sharply; when K is equal to 13, the contour coefficient reaches its maximum value;

and when K increases to 14, the contour coefficient drops in a cliff-like way. Therefore, the optimal cluster number should be 13.

The more the number of event reports in a period of the same length proves that this event is more important. Therefore, five clusters with the highest content of documents are selected from the 13 clusters generated as the news set of key events in November, and they are ranked as cluster 11, cluster 10, cluster 4, cluster 2 and cluster 9 respectively according to the number of documents.

Next, we extract keywords from the news text set of key events according to TF-IDF value to name the event set, as shown in table 1. In the TOP-*N* keywords we extracted, there will be some words without the meaning of event direction, such as "investor", "capital", etc. The following table only lists the keywords that are useful for event naming in TOP-*N*. In order to verify the feasibility of this method to extract key events, we collected the events of November in the major annual events [14-16] summarized by the online media in 2017, and the results are also compared as shown in Table 1.

Table 1. Event set keyword recognition results

| Event No.( Cluster No.) | data size | Data rank | Keywords extracted | Events based on keywords | Events summarized by network media |
|---|---|---|---|---|---|
| 2 | 333 | 4 | Quotes, steel, liquor, chips | Steel plate, liquor segment, chip concept stocks performed well | Guizhou Moutai shares broke through 700 |
| 4 | 397 | 3 | Brokerage, asset management, research | New regulations for asset management | The new regulations for asset management began to solicit opinions |
| 9 | 297 | 5 | Half day, net, outflow, 100 million yuan | High market volatility | IPO application review "zero pass" for the first time |
| 10 | 484 | 2 | Zhao Wei, Longwei, Wanjia, Media | Zhao Wei and her husband were punished by the Securities and Futures Commission | Zhao Wei and her husband were punished by the CSRC |
| 11 | 966 | 1 | 360, shell, return, Jiangnan, Jiajie | 360 Backdoor Recombination Regression | 360 Backdoor Recombination Regression |

According to Table 1, it can be seen that the first four of the key events extracted by the experiment are generally consistent with the media summary, which proves that the experimental method is accurate and feasible.

Among them, the first two key events are the same; in the third event, the news report focuses more on the research on the content and impact of the new regulations; the fourth cluster news collection contains three well-performing sectors, the keyword heat is roughly the same, but the daily market analysis can not be counted as an event. In a strict sense, the corresponding event should be the “Guizhou Moutai shares broke through 700” in the media summary. The keywords of the fifth cluster have no obvious event pointing, and most of them are the regular keywords that the investors pay attention to daily, and the event hypothesis that “IPO application review ‘zero pass’ for the first time” may be reflected in other clusters.

According to this method, a total of 45 events were identified throughout 2017. Generally speaking, in the

time span of this month, the extraction results of the two or three events before the key events are extracted by the test method of this article are more accurate, and then will be affected by the regular market analysis report. However, in terms of the selection method of key events, this experiment will be more objective than the media selection, reducing the impact of human factors.

### *4.3. Event Correlation Experiment and Result Analysis*

After extracting the TOP-N keywords of the event news set, we can establish a keyword co-occurrence matrix, calculate the weight of the associated event, and construct an event correlation network.

The first three key events that were better identified in November 2017 were selected. This experiment uses the TOP-15 keyword description for each event. In the following, the event 1 represents "360 Backdoor Recombination Regression", event 2 stands for "Zhao Wei couple is punished by the CSRC", and event 3 stands for "new asset management release". The cosine distance between the keyword word vectors is calculated as the co-occurrence degree and normalized to obtain a 15×15 co-occurrence matrix. The correlation between pairs of events is then calculated using equations (5). The results are shown in Table 2.

Table 2. Event correlation calculation results

| Event pair | Number of Co-occurrence words | Event correlation index $w(E_i, E_j)$ |
|---|---|---|
| Event 1 and Event 2 | 3 | 335.7009 |
| Event 1 and Event 3 | 5 | 332.1145 |
| Event 2 and Event 3 | 3 | 406.0615 |

It can be seen from the calculation results that event 2 and event 3 have the highest correlation and the most significant correlation. The other two groups have similar correlations.

Although event 1 and event 3 have the most number of co-occurrence pairs, the correlation between event 1 and event 3 is minimal. This is because in the calculation formulas (5) of the event correlation, the addition of different event key weights is an addition and the weight of the same keyword in different events is multiplied. It can be concluded that the correlation between a co-occurrence word and an event is more important than the number of co-occurrence words in the events.

In the event-to-event 2 and event 3, although there are not many common words, these co-occurrence pairs are highly correlated with the event, so this set of events is also considered to be highly correlated. The three co-occurrence words are "regulatory", "audit" and "shareholding", and event 2 and event 3 are "Zhao Wei and her husband were punished by the CSRC" and "new regulations for asset management", the correlation between them It is more common sense than the other two event pairs.

Finally, the event correlation network is visualized using Netdraw software, as shown in Fig.3.

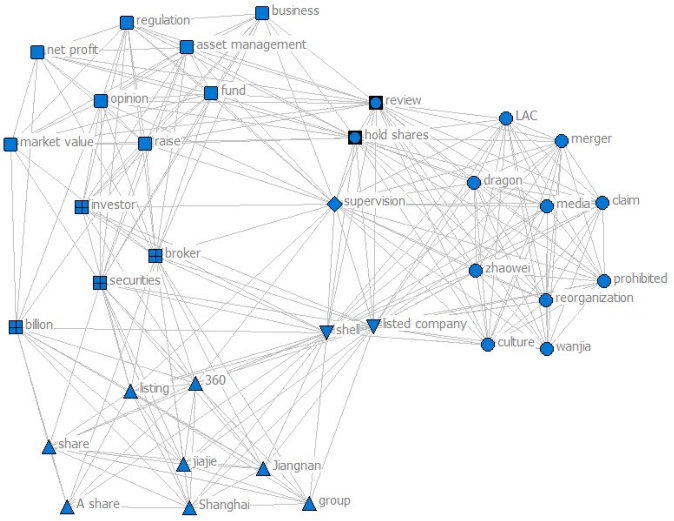


Fig. 3. Event vocabulary association network diagram (Upper triangles for event 1's keywords; circle for event 2's keywords; square for event 3's keywords. Lower triangles for event 1 and event 2's co-occurrence keywords; squares with crosses for event 1 and event 3's co-occurrence words; squares containing circles for event 2 and event 3's co-occurrence words; diamonds for event 1, event 2 and event 3's co-occurrence words.)

## 5. Conclusions

This paper aims to extract the effective event information in the massive financial news, use the text clustering method to extract news events, and extract financial news hot events more objectively, which helps investors to research market hotspot information and improve the efficiency of information collection. An event correlation model based on word co-occurrence and word vector is proposed. The semantic relationship between news events is mined according to event keywords, which provides a basis for future investors and regulators to predict the chain reaction of events. Experiments show that this method can extract news hotspot events effectively, and at the same time prove that news events happening in the same time period will produce a wide range of common discussions, making news events highly relevant on the semantic level.

## References

[1] TAKEDA F，YAMAZAKI H．Stock price reactions to public TV programs on listed Japanese companies［J］． Economics Bulletin，2006，13(7):1-7.

[2] Wang Qian, Wang Cheng, Feng Zhenyuan, et al. A review of K-means clustering algorithms. Electronic Design Engineering, 2012, 20(7): 21-24. (in Chinese)

[3] Gesang Duoji, Qiao Shaojie, Han Nan, et al. Single-Pass-based network public opinion hotspot discovery algorithm. Journal of University of Electronic Science and Technology of China, 2015, 44(4): 599-604. (in Chinese)

[4] Zhang Kuo, Li Zizi, Wu Gang, et al. A new event detection model based on lexical reevaluation. Journal of Software, 2012, 19(4): 817-828. (in Chinese)

[5] Zhang Bin, Hu Linmei, Hou Lei, et al. Chinese Event Discovery and Representation Based on Word Vector[J]. Pattern Recognition and Artificial Intelligence, 2018(3): 275-282. (in Chinese)

[6] Wan Xiaoxia, Zhao Jia. Research on Network News Hot Spots Discovery Based on Clustering[J]. Modern Computer Journal, 2015(9): 36-39. (in Chinese)

[7] Feng X , Qin B , Liu T . A language-independent neural network for event detection[J]. Science China(Information Sciences), 2018, v.61(09):81-92.

[8] Jacobs, Neil. Co-term network analysis as a means of describing the information landscapes of knowledge communities across sectors[J]. Journal of Documentation, 2002, 58(5):548-562.
[9] Tseng Y H , Ho Z P , Yang K S , et al. Mining term networks from text collections for crime investigation[J]. Expert Systems with Applications, 2012, 39(11):10082-10090.
[10] SHIBATA T, KUROHASHI S. Acquiring strongly-related events using predicate-argument co-occurring statistics and case frames [C]// IJCNLP 11: Proceedings of the 5th International Joint Conference on Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2011: 1028 – 1036.
[11] Zhang Hui, Li Guohui, Xu Xinwen, et al. News Event Correlation Modeling of Word Network[J]. Journal of National University of Defense Technology, 2014, 36(4): 169-176. (in Chinese)
[12] Park S C , Choi L C . An Automatic Semantic Term-Network[C]// International Conference on Computing. IEEE, 2008.
[13] Egghe L. New relations between similarity measures for vectors based on vector norm s[J ]. Journal of the American Society for Information Science and Technology, 2009, 60(2): 232 ～ 239.
[14] Yan Weimin. At the end of 2017, the six major events in China's stock market! [EB/OL].(2019-04-02)[2017-12-23]. http://www.sohu.com/a/213909036_651060
[15] Global Network. Top Ten Events in China's Securities Market in 2017 [EB/OL].(2019-04-02)[2018-01-02]. http://www.sohu.com/a/214142959_162522
[16] City Hunter. 2017 Top Ten Financial News of China A Shares [EB/OL].(2019-04-02)[2018-02-22]. http://www.zhicheng.com/n/20180211/203456.html