A Clustering Analysis of News Text Based on Co-occurrence Matrix

Shan Liu*, Xinyi Fan, Jianping Chai
School of Information Engineering
Communication University of China
Beijing, China
e-mail: liushan@cuc.edu.cn

Abstract—**In this paper, we use the improved TF-IDF formula to calculate the weight of the feature word of News. The keyword extraction takes into account the factors such as the parts of speech of feature words and inverse document frequency (IDF). And the K-core theory is used to determine the range of keywords. This study analyzes the co-occurrence strength of news keywords in a certain period of time, and obtains the cluster analysis of the co-occurrence intensity distance of news keywords. And the clustering results are analyzed in the end. In this paper, the quantitative research methods commonly used in bibliometrics are used in the news field to analyze the news content.**

Keywords-component; Co-occurrence matrix; clustering analysis

I. Introduction

Co-word analysis is a kind of content analysis method. For the high-frequency keywords that represent the hotspots of the news, we can make a co-occurrence matrix of a high-frequency subject. The data in the matrix is the co-occurrence frequency corresponding to the word. After the appropriate calculation of these data into the correlation coefficient in the correlation matrix, we can show the degree of high-frequency keywords. Software such as SPSS and UCINET is generally used for clustering analysis of these keywords [1, 2].

In general, co-occurrence analysis of the keywords is to reveal the various relationships between the key words through the observation matrix of co-occurrence frequency between the keywords. But in this paper, the matrix is judged by the concept of nuclear, not only by human judgment any more so that the construction of co-occurrence matrix is more theoretical and scientific.

We'll get word vectors through automatic segmentation of news texts by the pre-processing and cleaning of the selected news texts where each word carries a part of the word tag, such as nouns, verbs, adjectives, words and other types. The contribution of different parts of speech to the expression of the subject is different, in which the most important part for the expression and recognition of the subject is verb and noun. In addition to the content of the news, we also add adjectives as an indicator. Therefore, we only consider these three parts of speech in the word frequency statistics, the other part of the word ignored. In the case of statistics, the messages are grouped into different unit panes according to the time of occurrence, taking several days as partition. Then, the frequency of the word in the same pane is counted to get a list of the total words in the time period. The list is sorted by word frequency. Words with highest frequency are retained for the detection of keywords, and the long tail part is eliminated to extract the characteristic words.

II. Selection of Feature Words

A. Calculation of the Weight of the Feature

In general, the method of calculating the weight of feature words is the TF-IDF formula as shown below in (1).

$$\mathrm{w}_{ik} = \frac{\log(tf_{ik}+1.0)*log(idf_k)}{\sqrt{\sum_{i=1}^{N}[\log(tf_{ik}+1.0)*log(idf_k)]^2}} \tag{1}$$

tf_{ik} refers to the number of occurrences of the feature words in the news texts;idf_k refers to the number of occurrences of $term_k$ in all documents; And N represents the total number of news reports.

Especially for some unique verbs, nouns, places, time, and characters as important elements of a news document, the title of the news document is more representative of the subject of the document. We increase the weight of these feature items in calculation based on the analysis. We use the improved TF-IDF (2) as shown below.

$$\mathrm{w}_{ik} = \frac{W*\log(tf_{ik}+1.0)*log(idf_k)}{\sqrt{\sum_{i=1}^{N}[\log(tf_{ik}+1.0)*log(idf_k)]^2}} \tag{2}$$

In the formula, W is the parameter used to control the size of w_{ik}, given a weight coefficient greater than 1(w>1) when calculating named entities and the news headers to increase the weight of the words in the corresponding document; N Represents the total number of news documents.

B. The Distinction between High Frequency Words and Low Frequency Words

In 1973, Donohue [3] studied the distinction of high frequency words and low frequency words in English articles on the basis of the expression of Burt's law. And the result is also applied to the Chinese article. The boundary formula of high frequency words and low frequency words is as follows:

$$\mathrm{n} = {\left(-1+\sqrt{1+8I_1}\right)}\Big/_{2} \tag{3}$$

978-1-5090-6352-9/17/$31.00 ©2017 IEEE

As can be seen from (3), the size of frequency n only depends on the value of I_1. (I_n describes the number of words with the same frequency). The calculation of the value of n of the content data to be analyzed can help to get the dividing line between high frequency words and low frequency words in each article.

C. Keyword Extraction

When an item occurs frequently over a certain period of time, and its frequency of occurrence is significantly increased over the previous time period, it means that it is associated with some of the newer news topics [4] to some extent.

$$\mathrm{G}_{ik} = \frac{f_{ik} - f_{i(k-1)}}{1 + f_{i(k-1)}} \quad (4)$$

In (4), G_{ik} denotes the rate of increasing rate of the word i in the k time window, and $f_{i(k-1)}$ is the frequency of the word i in the k-1 time window (the last time window).

After the pre-processing as word segmentation, part-of-speech filtering and word frequency filtering on the news data, the meaningful verbs, nouns and adjectives are selectively remained. On this basis, the compound weights of the relative word frequency and word frequency increasing rate serves to evaluate the characteristic expression of a characteristic word W_{ik}:

$$W_{ik} = \alpha \ln T_{ik} + \beta \ln G_{ik}$$

(5)

The larger the value of W_{ik}, the greater the probability that the characteristic word is the keyword. In (5), parameters α and β are used to adjust the specific relationship between the relative frequency and word frequency increasing rate. The bigger β is, the more word frequency increasing rate plays the main role when α is fixed. In the opposite, the bigger α is, the more relative word frequency is considered when β is fixed.

According to the threshold, we select the feature words which have large weights to get a thesaurus after calculating the value of W_{ik} of the words in each time window. The thesaurus is characterized by its high occurrence frequency in the current time window and low occurrence frequency in the previous time window. After selecting the keywords, we can analysis the keywords in this way to construct the word co-occurrence graph. Through the division of the figure, we can get the news topic recognition.

D. K-core Matrix

The K-core concept comes from graph theory. The row and column in which the number of nonzero numbers in the co-occurrence matrix is less than k+1 are ignored. In the remaining matrix, because of the ignorance of some keywords, nonzero numbers on the row and column where the remaining part of the keywords are, is less than k+1. We recursively ignore the remaining row and column in which the number of nonzero numbers in the co-occurrence matrix is less than k+1 until all nonzero numbers on the row and column where the remaining part of the keywords are is less than k+1. Then the co-occurrence matrix between these retained keywords is a K-core matrix [5-10].

III. Research on Data of News Content

We take the news content from CCTV news as the dataset. The feature words are analyzed from the text of the News content. The news data is selected from September 1 to September 11, 2016. There are 6246 words and 60 news stories.

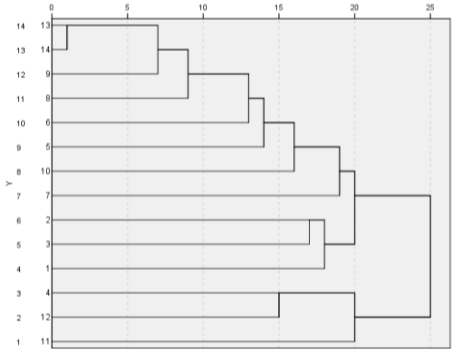

Figure 1. Co-clustering tree of the feature words.

(1. G20 summit 2.China 3.Hangzhou 4. Xi Jinping 5.world economy 6.increase 7.cooperation 8.inclusive 9.innovative 10. Developing country 11.business 12.The opening ceremony 13.interconnected 14.invigorated)

The number of 14 feature words are obtained in the processing of the co-occurrence matrix, grounds on the K-core theory of the social network. The co-occurrence data of these keywords are extracted to obtain the co-occurrence matrix of Table I below. We use Bibexcel software to get the co-occurrence of this table.

TABLE I. Co-occurrence Matrix of the Feature Words

	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***
1	200	24	27	19	20	17	8	13	14	8	5	9	10	9
2	24	116	22	12	17	17	8	12	12	6	2	5	9	7
3	27	22	92	15	19	16	9	11	11	8	4	7	9	8
4	19	12	15	66	13	10	6	9	7	6	4	9	8	6
5	20	17	19	13	65	16	8	12	12	8	3	6	10	9

	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***
6	17	17	16	10	16	55	8	13	12	7	2	3	10	8
7	8	8	9	6	8	8	32	7	6	5	2	2	6	5
8	13	12	11	9	12	13	7	31	9	6	2	3	10	8
9	14	12	11	7	12	12	6	9	27	6	2	1	9	8
10	8	6	8	6	8	7	5	6	6	20	1	2	5	5
11	5	2	4	4	3	2	2	2	2	1	18	4	2	1
12	9	5	7	9	6	3	2	3	1	2	4	17	2	1
13	10	9	9	8	10	10	6	10	9	5	2	2	13	8
14	9	7	8	6	9	8	5	8	8	5	1	1	8	10

1. G20 summit 2.China 3.Hangzhou 4. Xi Jinping 5.world economy 6.increase 7.cooperation 8.inclusive 9.innovative 10. Developing country 11.business 12.The opening ceremony 13.interconnected 14.invigorated

TABLE II. DISTANCE MATRIX BETWEEN TWO WORDS

	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***
1	0.00	0.84	0.80	0.83	0.82	0.84	0.90	0.83	0.81	0.87	0.92	0.85	0.80	0.80
2	0.84	0.00	0.79	0.86	0.80	0.79	0.87	0.80	0.79	0.88	0.96	0.89	0.77	0.79
3	0.80	0.79	0.00	0.81	0.75	0.78	0.83	0.79	0.78	0.81	0.90	0.82	0.74	0.74
4	0.83	0.86	0.81	0.00	0.80	0.83	0.87	0.80	0.83	0.83	0.88	0.73	0.73	0.77
5	0.82	0.80	0.75	0.80	0.00	0.73	0.82	0.73	0.71	0.78	0.91	0.82	0.66	0.65
6	0.84	0.79	0.78	0.83	0.73	0.00	0.81	0.69	0.69	0.79	0.94	0.90	0.63	0.66
7	0.90	0.87	0.83	0.87	0.82	0.81	0.00	0.78	0.80	0.80	0.92	0.91	0.71	0.72
8	0.83	0.80	0.79	0.80	0.73	0.69	0.78	0.00	0.69	0.76	0.92	0.87	0.50	0.55
9	0.81	0.79	0.78	0.83	0.71	0.69	0.80	0.69	0.00	0.74	0.91	0.95	0.52	0.51
10	0.87	0.88	0.81	0.83	0.78	0.79	0.80	0.76	0.74	0.00	0.95	0.89	0.69	0.65
11	0.92	0.96	0.90	0.88	0.91	0.94	0.92	0.92	0.91	0.95	0.00	0.77	0.87	0.93
12	0.85	0.89	0.82	0.73	0.82	0.90	0.91	0.87	0.95	0.89	0.77	0.00	0.87	0.92
13	0.80	0.77	0.74	0.73	0.66	0.63	0.71	0.50	0.52	0.69	0.87	0.87	0.00	0.30
14	0.80	0.79	0.74	0.77	0.65	0.66	0.72	0.55	0.51	0.65	0.93	0.92	0.30	0.00

The distance matrix between the two words is obtained by the results that 1 minus numbers of the co-occurrence matrix, as shown in Table II.

The distance matrix of Table II is input into SPSS software to form co-clustering tree diagram using system clustering method.

Let’s take a look at Fig. 1. It can be seen that the 13th and 14th keywords are closer, that it, the co-occurrence strength of words “interconnected”, “invigorated” is relatively strong. And the 8th and 9th feature words are followed to them. We know the theme of the g20 summit is “Toward an Innovative, Invigorated, Interconnected and Inclusive World Economy”.

The distance between the fifth, the sixth and the above four words is close. To say, the co-occurrence strength of words “the world economy”, “increase”, “inclusive”, “innovative”, “interconnected”, “invigorated” is strong, which shows that the growth of the world economy is inseparable from inclusive, innovative, interconnected and invigorated economic pattern.

At the same time, the first one, second and third keywords are very near, indicating that Hangzhou is the venue for the g20 summit and China is also one of the participants. The distance between the 4th, 11th and 12th feature words is not very far, while the 11th word is slightly far from the 4th and 12th ones. The explanation might be that Xi Jinping attended the opening ceremonies of the summit of which part of them are the opening ceremonies of business.

Ucinet Software is used for analysis of the co-occurrence matrix of the feature words. Drawing with NetDraw software, the matrix is normalized into a 0-1 matrix, and different truncation values are tested for clustering. Based on the value of experience, it has better effect when the truncation value is 7 in the process of clustering. Figure 2 is seen as below.

The relevance between the word “business” and the theme is not strong so “business” is outside the clustering figure. The connection between words “the opening ceremony” “the developing country” and other words is few and scattered; “Hangzhou”, “G20 summit”, “Xi Jinping”, “world economy”, “China” connect more closely, in which there is a close connection between “innovative, “invigorated”, “interconnected” and “inclusive” as the theme of the G20 summit.

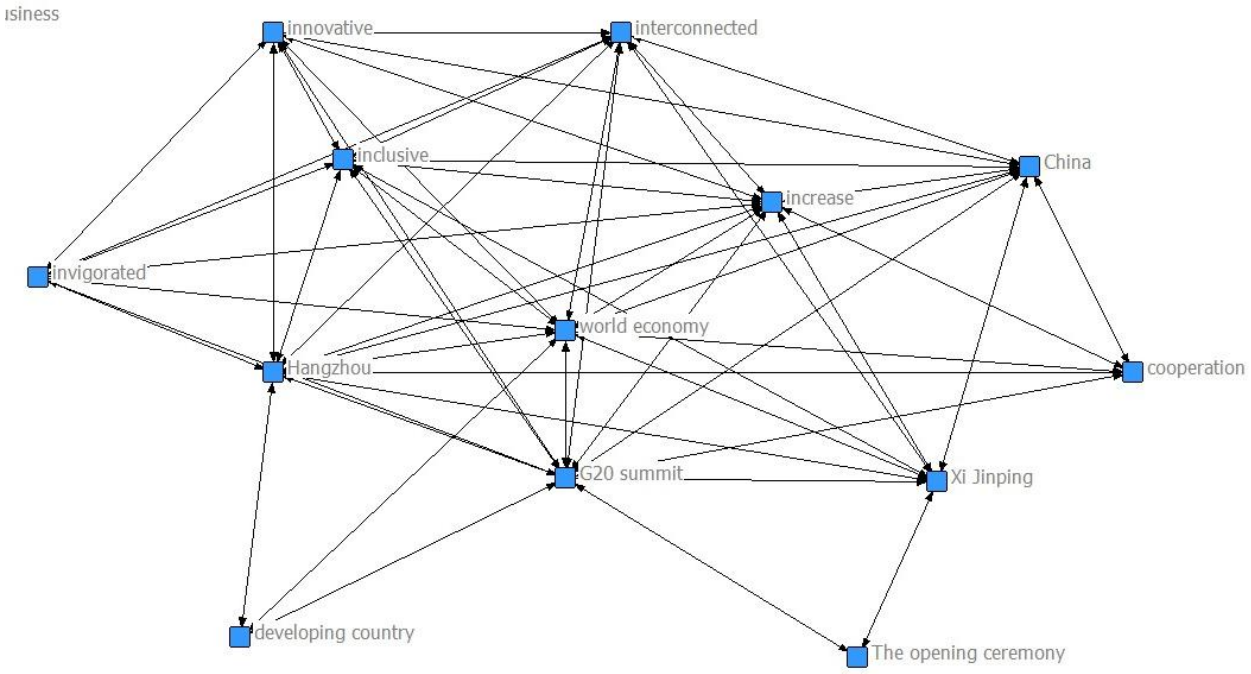

Figure 2. Clustering figure of feature words.

IV. COUCLUSIONS

In this paper, we consider the concept of time window in line with the characteristics of news events. The keyword extraction takes into account the factors such as the parts of speech of feature words and inverse document frequency (IDF). We choose nouns, verbs and adjectives as the feature words of the news and use K-core theory to determine the scope of the keywords. Besides, the improved TF-IDF formula is used to determine the weight of keywords. Data sources of the news content are from CCTV news for a period of time in 2016. We take selected feature words of high frequency of the news content which satisfy the K-core theory as the object of analysis. This study analyzes the co-occurrence strength of news keywords and obtains the cluster analysis of the co-occurrence intensity distance of news keywords. And the clustering results are analyzed in the end. We provide reference for keyword clustering of news text.

REFERENCE

[1] W. Zhao, and X. Hou, “News topic recognition of Chinese microblog based on word co-occurrence graph,” CAAI Transactions on Intelligent Systems, pp. 1-6, May 2012.

[2] C. Chou, S. Hsieh, C. Qiu, "Hybrid genetic algorithm and fuzzy clustering for bankruptcy prediction," Applied Soft Computing, vol.56, pp.298-316, 2017.

[3] E. Lord, M. Willems, F.-J. Lapointe, and V. Makarenkov, "Using the stability of objects to determine the number of clusters in datasets," Information Sciences, vol.393, pp.29-46, 2017.

[4] Zhao Wenqing, Hou Xiaoke, “News topic recognition of Chinese microblog based on word co-occurrence graph,” CAAI Transactions on Intelligent Systems, 2012, 7 (5): 1-6.

[5] Chih-Hsun Chou, Su-Chen Hsieh, Chui-Jie Qiu, "Hybrid genetic algorithm and fuzzy clustering for bankruptcy prediction," Applied Soft Computing, vol.56, pp.298-316, 2017.

[6] Z. Xu, Z. Xu, D. Li, and S. Li, “Topic detection and tracking for internet news,” Intelligent Computer and Application, pp. 59-65 Jan. 2011.

[7] Zhou Aimin, “The Cluster Analysis of Co-occurrence Strength in the Field of Knowledge Management in 2006,” Modern Information, 2008, 28 (5): 30- 33.

[8] Xu Zhikai, Xu Zhiming, Li Dong, Li Sheng, “Topic Detection and Tracking for Internet News,” Intelligent Computer and Application, 2011, 01 (3): 59- 65.

[9] Booth A D., “A Law of Occurrences for Words of Low Frequency,” Information and Control, 1967, 10 (4): 386-393.

[10] Bing Liu, Chunru Wan, and L.P. Wang, “An efficient semi-unsupervised gene selection method via spectral biclustering,” IEEE Trans. Nano-Bioscience, vol.5, no.2, pp.110-114, June, 2006.