A Detection Method of Online Public Opinion Based on Element Co-occurrence

Nanchang Cheng, Yu Zou(✉), Yonglin Teng, and Min Hou

National Broadcast Media Language Resources Monitoring and Research Center, Communication University of China, Beijing 100024, China
{chengnanchang, zouiy, tengyonglin, houmin}@cuc.edu.cn

Abstract. Discovering and identifying public opinion timely and efficiently from web text are of great significance. The present methods of public opinion supervision suffer from being rough and less targeted. To overcome these shortcomings, this paper provides a public opinion detection method of network public opinion based on element co-occurrence for specific domain. This method, considering the nature of public opinion, represents three main factors (subject, object and semantic orientation) that constitute public opinion by employing their feature words, which can be dynamically combined according to their syntagmatic and associative relations. Thus, this method can not only generate topics related to public opinion in specific fields, but also identify public opinion information of these fields efficiently. The method has found its practical usage in "Language Public Opinion Monitoring System" and "Higher Education Public Opinion Monitoring System" with accuracies 92% and 93% respectively.

Keywords: Element co-occurrence · Online public opinion
Syntagmatic relations · Associative relations

1 Introduction

At present, public opinion recognition and monitoring is a popular research field. What is the public opinion? Reference [1] regarded public opinion as "the sum of many emotions, wills, attitudes and opinions, in certain historical stage and social space, held by individuals and various social groups, to the various kinds of public affairs which are closely related to their own interests".

In short, public opinion detection is to check whether the content of the text connects with the public opinion. According to the definition of text classification in Ref. [2], public opinion detection is a branch of text classification, which means that whether the text contains public opinion information can determine it is public opinion or not. Public opinion detection is at the predecessor position of public monitoring. Only if the public opinion information is gathered in time, the further analysis of public opinion is available, which involves classification, hotspot identification, orientation analysis etc. Because public opinion is characterized by its abruptness, it is hard to predict what and where to occur. Thus, it is critical to detect and identify public opinion information in time. However, at present, there is a scarcity of the literature related to

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11256, pp. 322–334, 2018.
https://doi.org/10.1007/978-3-030-03398-9_28

the public opinion detection, and most publicized public opinion detection systems merely employ techniques such as text classification, information filtering and keyword retrieval [3]. To reduce redundancy, these systems graded the keywords. For example, we can input the word "housing removal" as the first-grade keyword, "Yun Nan Province" as the second-grade keyword (co-occurred word) and "Honghe Area" as the keyword to be excluded, which means that we want to find public opinions concerning the housing removal which happened in Yun Nan Province excluding "Hong He Area". The method of keyword grading and public opinion dictionary is of high speed when searching and identifying mass online information, and of high flexibility as well, for it permits the addition of batched keywords in a custom way according to user's needs. However, there are still two remaining problems: (I) the adding keywords must be the topics we have known, but yet the system lacks of ability in acquiring unknown public opinion information; (II) keywords only cover one point the text involves, which leads to the fact that it lacks enough tension to make sure that all the text extracted is concerned with public opinion information. These two drawbacks lead to high redundancy and high cost in processing texts.

Public opinion covers different fields in society, and in every field, public opinion shows unique characteristics. However, at present, public opinion detection method for a specific sub-field is still rare in literature, and the public opinion monitoring system mentioned above and other publicized systems are basically geared to all fields. Generally speaking, the more in detail the classification of sub-fields is, the deeper the research would go. Whole-field monitoring is one of the important reasons for the roughness in public opinion monitoring.

Therefore, in order to improve the roughness and the low specialization for sub-fields in the public opinion detection method, this paper, based on the nature of public opinion, and particularly imitating human's cognitive process of public opinion information, proposes the element co-occurrence method, a method that is specialized for online public opinion detection in subfields. This paper will take language public opinion detection as an example to illustrate this method and its detail implementation.

2 Relevant Studies

Studies related to public opinion detection mainly concentrates on the topic detection field. There used to be an international conference concerned with the evaluation of this field, whose name was Topic Detection and Tracing (TDT in short) [4]. In TDT, A topic refers to "a set of reports of a seed event or activity and its directly related events or activities" [5, 6]. Topic Detection (TD in short) task is to detect and organize topics that are unknown to the system [6]. Technically, statistical clustering algorithms are widely employed, such as K-Means [7], Centroid [8] and Hieratical Clustering [9], etc. Because of the mass calculation in clustering, when dealing with the massive online texts, it is rare to directly detect the public opinion related topics by clustering method.

Although TDT had stopped in 2004, related researches still go on. In recent years, Refs. [10, 11] proposed new event detection method based on topic classification and lemma re-evaluation respectively. But the TDT test corpus that they used have been carefully classified according to topics, however, in actual condition, online texts do

not have related information as classification and sub-topics to use. Reference [12] used keyword-based search method to detect the emergency events in Sina blog, and, by restricting time period and domain names to narrow down the search results and reduce the redundancy. This is similar with the keyword search method mentioned above. Reference [13] recognize sentences which contain critical information through hot words. Then apply clustering to all the sentences recognized to implement hot topic recognition. The probability that the hot topic belongs to public opinion is relatively high, which is related with this research. Though Ref. [13] reduced the computation amount from paragraph level to sentence level, hot word and sentence recognition still consume a lot.

To sum up, deficiencies of the current public opinion detection can be generalized into following 3 points: (I) Low specification. Most of the publicized systems are whole-fielded, which perform ineffectively in specific field. (II) Most of the publicized systems based on batched keywords or public opinion dictionaries, whose deficiencies have thoroughly revealed in Sect. 1. (III) Most of the statistic-based clustering method and other new methods are still at the theoretical level, are still rare in real public opinion detection.

3 Main Idea of Element Co-occurrence

According to the definition in Ref. [1], the public opinion is the sum of many emotions, wills, attitudes and opinions, in certain historical stage and social space, held by individuals and various social groups, to the various kinds of public affairs which are closely related to their own interests. It is obvious that the public opinion is composed of three basic elements: subject (people), object (various public affairs) and semantic orientation (the sum of emotions, wills, attitudes and opinions). "Element co-occurrence" starts from the essence of public opinion, representing each element by a feature word. Three kinds of feature words can be combined with each other dynamically to generate a topic that is related to public opinion in the certain field. For example, in language field, there are public opinion events as "traditional characters or simplified characters", "protect the dialect" and "letter words tumult". The relation of three kinds of elements represented by feature words can be shown as Fig. 1.

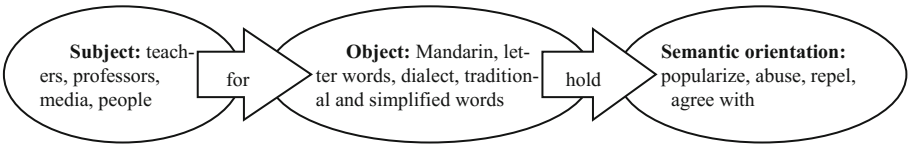

Fig. 1. Three kinds of feature words of public opinion on language and their relationship

The figure above expresses that: for language public opinion, subjects as professors and teachers hold opinions or attitudes as disagree, repel or agree for objects as letter words. In that, "for" and "hold" are the pre-set keyword in the pattern, and keywords in

three elements as "subject", "object" and "semantic orientation" are automatically extracted from the text or summarized according to the experience. Three types of feature words can combine dynamically with strong tension. Such combination can cover all public opinion that may appear in language public opinion field and can exclude most of the non-public-opinion information. The theoretical basis is Combinatorial Polymerization theory of Saussure.

Sweden linguistic Saussure pointed out that in language status, all of them are based on relationships [14]. The core of it is the sentence segment relationship and association relationship, which, in another word, combination relationship and aggregation relationship. Combination relationship refers to the horizontal relationships among language units that appear in language and based on linear basis; aggregation relationship refers to the vertical relationships among language units that may appear at the same position with same functions. According to this theory of Saussure, the dynamic combination of three kinds of characteristic keywords that mentioned above can generate different topics. For example, according to combination relationship, the system can generate topics as "teachers popularize Mandarin", "experts repel dialects", "media abuse letter words" and "people agree with simplified words" etc.; according to aggregation relationship, the system can generate topics as "teachers popularize Mandarin", "experts popularize Mandarin", "media popularize Mandarin" and "people popularize Mandarin" etc. As one can discover, element co-occurrence method is the simulation of the corpus of certain field's knowledge in human brain (combination relationship) together with the comprehension and expression generation of objects (aggregation relationship), which has strong topic generation ability. Moreover, so long as the topic is able to generate by this method, the effective identification of the topic is almost indeed. If, for a specific field, according to the characteristics of its public opinion, a corpus containing three kinds of feature words can be build, it will be possible to detect the public opinion in that field effectively. The generative ability of element co-occurrence is potential, when keywords appeared in a piece of text, this method can automatically ignore other words that are not related with the feature words and dynamically generate matching topics. For instance, after ignoring other words, for text piece "some post-90s students very like traditional characters", topic "students like traditional characters" can be detected.

From the perspective of public opinion detection, the feature words of objects are most important. In the text, firstly, if only language-related words appeared, it is meaningful to discuss whether these belong to public opinion, and we can call them "topic words"; after that, the feature words with emotional inclination can be called "emotional words"; thirdly, the feature words associated with subjects, which are typically people as students, parents and teachers etc. Besides, the occurrence of public opinion requires certain time and space environment, and accordingly, their feature words are like "class, classroom, and school etc.", they also affect the public opinion detection, some even can replace the subjects, as "school popularizes Mandarin". In these condition, time and space feature words are similar to the feature words of subject, therefore, it is possible to combine these feature words into "people and environment" class, which, in short, "environment words". In three kinds of feature words, any of them alone cannot compose a public opinion topic directly, the co-

occurrence of two or more feature words is a necessity to compose a public opinion topic. Based on that, this method is called "element co-occurrence method."

Element co-occurrence method detects the public opinion towards constructing a discourse knowledge system related to public opinion in some fields. This method, instead of concerning a single point, concerns the combination of three basic elements that relate to the public opinion, which shows strong tension. Thus, this method has an essential difference with traditional detecting methods as keyword method or public opinion dictionary method. By batched keywords or public opinion dictionary, one can only search a point of public opinion, which is one-dimensional. For example, as "demolition incident", "Zhao Yuan murder" and "terror incident". Element co-occurrence is three-dimensional and is formed by the combination of three kinds of feature words to form different topics. Keyword grading method or public opinion dictionary method also concerned co-occurrence, but the co-occurrence in these methods is associated with some certain words. However, all elements in element co-occurrence method can combine with each other dynamically and have powerful topic generation ability. Taking advantage of this dynamic combination, element co-occurrence method endows the public opinion monitoring system public opinion alert function by discovering the unknown topic in real time.

4 Implementation of Element Co-occurrence

4.1 Extracting the Feature Words of the Three Types

The prerequisite of element co-occurrence is to establish three feature words sets. Feature words can be collected manually or be obtained by automatic searching method. This paper has 9436 texts (referred as X set) with 12.5 million words, among which 1836 texts are related to public opinion (referred as Y set) with 2.5 million words, and the rest 7600 texts (referred as Z set) are non-public opinion articles which are over 10 million words. Then the word segmentation system, CUCBst, is used to extract words and calculate word frequency. The words extracted from X, Y and Z are then graded according to the frequency: Grade 1 ($\geq$1000 times), Grade 2 (500–999 times), Grade 3 (100–499 times), Grade 4 (5–99 times), Grade 5 (1–4 times). To identify the feature words from tests related to public opinion on language issues, words extracted from Z set are compared with words from X set in their respective grades. Taking the word "language" as an example, its frequency in X set is 7161 times and thus is a Grade 1 word, but it only appears 62 times in Z set and is rated Grade 4. Without the process of comparing words' frequency according to their grade, it would be impossible to identify the feature words of public opinion on language issues.

The extracted words need to be further classified into topic words, emotion words and background words. An extracted word is identified as a topic word if it matches a term from Chines Term in Linguistic [15], and an emotion word is identified according to Emotion Term Dictionary [16]. Those that do not fall into these two categories are automatically classified as background words. Taking the Grade 1 words in X and Z sets as example, the extraction process of feature words is illustrated in Fig. 2.

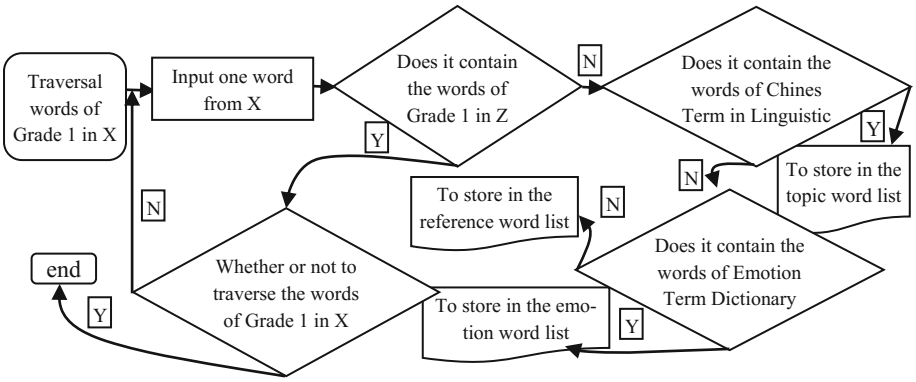

Fig. 2. Flow chart of three kinds of feature words extraction

The quality of feature set determines the accuracy and recall rate of public opinion detection. In order to guarantee the quality of feature word set, all of the feature words extracted automatically need to be manually confirmed.

4.2 Weighting Algorithm

Introduction to Weighted algorithm

The successful extraction of feature sets is the foundation of element co-occurrence method. Element co-occurrence is the main factor of public opinion judging, but it is not the only factor. In order to determine whether a given text is related to public opinion on language issues or not, the score of the text is affected by four factors: the normalized using rate of feature words, the co-occurrence of feature words, their location and the length of the text. When the score reaches a certain threshold, it can be determined that the text is related to public opinion on language issues.

Calculating the Weight of Feature Words

The importance of a word in a text set is usually indicated by its value of TF-IDF (term frequency-inverse document frequency). TF-IDF theory suggests that the importance of a word increases proportionally with the rising of its frequency in a certain text, but decreases with the increasing of the number of texts appearing in a corpus. That means the weight of special words appearing in only a few documents is higher than that of a word appearing in many documents [17]. However, the disadvantages of TF-IDF are obvious as it underestimates the importance of frequently occurring words in a certain domain. These words are usually highly representative and should be given a higher weight [18]. Therefore, this research chooses normalized using rate as an important quantification criterion. In fact, a text is more likely to be related to public opinion if the using rate of feature words in the text is high. For example, when feature words such as “Chinese” and “dialect” appear in the text, it is more likely to be related to the public opinion of language issues than a text with non-feature words such as “tone” and “syllable”.

Therefore, this paper employs the normalized using rate of feature words to determine its weight. The analysis of the feature words' using rate demonstrates that the using rates of the most frequently appearing feature words are generally ≥ 0.01, mid using rate is between 0.01 and 0.001, and the using rate of rarely used feature words is lower than 0.001. According to this finding, the weight of a feature word is defined into three grades, and each grade is given different points (3, 2 or 1). For example, the weight value of the word "language" is 3 points, and the weight value of "silk book" is 1 point. Table 1 shows ten typical features words from each of the three categories and their normalized using rates.

Table 1. Feature words and its normalized using rate.

No	subject	normalized using rate	sentiment	normalized using rate	person / background	normalized using rate
1	语言	0.330874223	问题	0.42520583	学生	0.318084654
2	汉语	0.204262811	规范	0.266159498	教学	0.107089852
3	汉字	0.177317285	重要	0.205760296	孩子	0.096642725
4	文字	0.108518912	保护	0.026511905	大学	0.095541486
5	语文	0.071111208	反对	0.017997309	小学	0.088119925
6	普通话	0.035162926	正确	0.013631831	学校	0.074040796
7	方言	0.03020151	严重	0.012919792	历史	0.072185948
8	中文	0.024844734	错误	0.012746114	考试	0.055044716
9	母语	0.011750716	缺乏	0.010717178	教师	0.047914166
10	繁体字	0.00595579	质疑	0.008350154	老师	0.045335821

The formula for calculating normalized using rates is as follows:

$$U_i = \frac{F_i \times D_i}{\sum_{j \in V} (F_j \times D_j)} \quad (1)$$

F denotes the frequency of the word and D denotes the distribution rate, and the denominator is the normalized term, V which denotes the set of all the homogenous survey objects (all word categories).

Calculate the Weights of Co-occurrence of Elements

Among the three kinds of feature words, topic words are the foundation: only when a topic word appears, an unknown segment of a text will be allowed to enter the next step of the analysis, otherwise, this segment will be abandoned directly. Thus the co-occurrence of the three types of feature words includes three cases: a. topic word + emotion word + background word; b. topic word + emotion word; c. topic word + background word. Case A, the co-occurrence of the three types of feature words, is most likely to be about public opinion. When there is only two types of

feature words appearing in a text, case b is more likely to be public opinion related than case c. Therefore, co-occurrence of feature words of different types is a very important weighting factor. The possibility of being related to public opinion is: a > b > c. Table 2 shows the co-occurrence of three types of feature words in clauses.

Table 2. The co-occurrence condition of three feature words in clause.

No	Sentence	Subject	Sentiment	Person/background
1	The author did not respond to the question of why Lu Xun hated Chinese characters	Chinese character	Hate	Lu Xun
2	If it has a severe mistake in the usage of Chinese characters	Chinese character	Mistake	
3	About 1000 Chinese learners	Chinese		Learner

Table 2 shows that in example 1, the three types of feature words appear in one sentence, and thus can be determined as a public opinion related text; in example 2, with the co-occurrence of a topic word and an emotion word, it can be basically determined as a text about public opinion; and in example 3, only topic word and background word appear. This example might present some public opinion information of the international influence of Chinese or can be a part of the introduction of TCFL (Teaching Chinese as a Foreign Language) major of a school. Therefore, the sentence cannot be directly determined as containing public opinion information.

In most cases, the shorter the distance between different feature words is, the closer the syntactic and semantic relations of these words are, and thus the more likely they are public opinion related topics. In the examples above, the distances among feature words are short as they appear in clause. However, more than often, feature words are scattered over a sentence or even a passage. Therefore, to solve the problem of how to identify co-occurrence distance when feature words are scattered in different part of an article, this paper classifies the co-occurrence distance into four levels: article, paragraph, sentence and clause. Section 5 introduces the weighted algorithm used for the distance comparison in the four levels.

Apart from co-occurrence, the location of feature words in the text and the length of the text are also factors to be considered in the weighted algorithm. In terms of location, only the title and the text are considered in this paper. The weight of the feature words appearing in title is different from the words that appear in text. In the aspect of text length, since the score is higher when the text is simply longer than other texts, so it is necessary to constrain the factor. This paper uses the average length of texts of Y set to constrain it.

Additive Weighted Algorithm

Additive weighted algorithm needs to consider four factors: feature words weight, co-occurrence of three types of feature words, feature words position and text length. Algorithm needs to segment the text according to the co-occurrence distances among feature words. As stated above, this paper divides co-occurrence distances into four level: article level, paragraph level, sentence level and clause level. This section takes

sentence level co-occurrence distance as example, illustrates the process of algorithm. The segmentation of sentences employs "。？！" as boundary of a sentence. The Score of a sentence is shown in Formula (2).

$$\mathrm{Sen}_i = \sum_{a \in \mathrm{A}} (\mathrm{F}_a \times \mathrm{U}_a + \mathrm{P}_a) + \sum_{b \in \mathrm{B}} (\mathrm{F}_b \times \mathrm{U}_b + \mathrm{P}_b) + \sum_{c \in \mathrm{C}} (\mathrm{F}_c \times \mathrm{U}_c + \mathrm{P}_c) + \mathrm{G}_i \quad (2)$$

In the formula, Seni represents the score of a sentence, a, b and c represents three types of feature words respectively, F represents word frequency, U represents weight and P represents position score. The score of one exact feature word in a sentence which is included in feature word list equals to: word frequency (F) first multiplies weight (U), then the result of multiplication adds position score (P). Gi is the co-occurrence score of three types of feature words, co-occurrence of all three types is highest, then is subject + sentiment, the lowest is subject + background.

At last, the total score of a text is represented in Formula (3).

$$\mathrm{Text}_i = \sum_{k=1}^{n} (Sen_k) \times \frac{\mathrm{AL}}{\mathrm{L}_i} \quad (3)$$

Text_i represent the score of text i, AL represents the average length of all texts in set Y, and L_i represents the length of text i. the text score equals to all the sentence scores in the text, then multiplies the average length, at last divides the length of this text.

5 Experiment Result Analysis

5.1 Experiment Data

To test the performance of element co-occurrence method, 1200 texts whose length is around 1000–1500 words were picked. Among them, 160 texts were related to language public opinion.

5.2 Compute the Co-occurrence Distance and Threshold

At present, Precision (p), Recall (r) and F1-Measure (F1) are the factors to evaluate the effect of a classifier. Under different threshold, different precision and recall will get. To get the best recognition result, precision and recall under different threshold was computed. Also, F1-Measure was considered in computing the threshold.

In order to evaluate the system, precision-recall curve and F1 curve were drew, as in Figs. 3 and 4. Generally, the better performance a system can achieve, the greater prominence of precision-recall curve should be. Figure 3 illustrates that when the co-occurrence distances of three types of feature words are at sentence level and paragraph level, performance of system is better than the condition when the co-occurrence distances are at clause level and article level. Moreover, in that, sentence level performs better than paragraph level, and clause level performs a little better than article level. Figure 4 illustrates that when threshold is at 90, F1-Measure gets the highest (0.94).

For the phenomenon stated above, this research suggests that: in normal understanding, the smaller co-occurrence distance of these three types of words, the tighter connection these words will be; however, it is usually hard for people to express the topic clear in a clause, three elements of a topic usually distributed at a larger range than a clause. Experiments showed sentence level is the best co-occurrence distance. According to the Experiment, this research set the co-occurrence distance of three types of feature words at sentence level with threshold at 90.

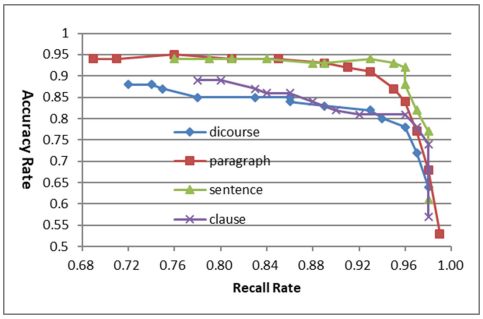

Fig. 3. Precision and recall curve of four lever

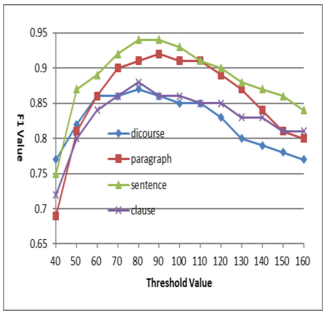

Fig. 4. Threshold and its corresponding F1 value of four level

5.3 Compute the Co-occurrence Distance and Threshold

The threshold was set to 90, and the result of the experiment is shown in Table 3.

Table 3. The experimental results when threshold is 90

Total texts	Identified texts	Relevant texts	Error texts	Precision	Recall	F1 value
161	165	153	12	0.93	0.95	0.94

Experiment Analysis

I. The language public opinion texts set adjusted from 160 texts to 161 texts. The reason is the detector found a not related text - China Daily: Beware of "Online Water Army" Kidnapping Online Public Opinion - from manually labeled as related texts; and found two language public opinion related texts - On the Revolution of "Country" and Nan Fangshuo: Chinese Shall Refind the Function of Talk - from manually labeled as not related texts. After careful analysis, the judgement of detector is right. Therefore, as one can see, the detector has a strong sense of objectivity.

II. The detector totally recognized 165 texts as relevance. In that, 153 are correctly judged, and 12 are misjudged. In the misjudged texts, 6 are education and culture category, 4 are music and dance category, and 2 are other categories. Through the analysis, the main reason of misjudgment is language public opinion also appears a lot in education and culture field, thus the feature word list of language public opinion

shares some words with education and culture field. How to detect and classify such texts of high similarity with language public opinion text is the top topic of further researches.

III. For texts like People are Stupid, But They Seem Great, which have a sense of ridicule, detector can correctly classify them. This proves that the detector has strong analysis and recognition ability.

5.4 Detector in Actual Use

We calculated the result of the detection system on randomly picked language public opinion in one-week size. The statistic reveals that the average precision of detector is around 92%. This system has been adopted by Department of Language and Information Management Afflicted to Ministry of Education and National Language Resources Monitoring and Research Center. Runtime of the system is more than 6 continuous years.

5.5 Element Co-occurrence Method in Tertiary Education Public Opinion

To valid the universality of element co-occurrence, the same effect was achieved by implementing this method at tertiary education online public opinion detection. As the result of the detection system on randomly picked tertiary education public opinion in one-week size shown that the average precision of detector in tertiary education public opinion detection reached 93%. This system has been adopted by National Tertiary Education Quality Monitoring and Evaluation Center which affiliated to the Ministry of Education Evaluation of Tertiary Education Research Center for Communication and Public Opinion Monitoring. Runtime of this system is more than 4 continuous years.

5.6 Comparison of Other Similar Methods

Reference [19] proposed an improved single-pass text clustering algorithm called single-pass*. Their experimental results show that, compared to the single-pass algorithm, the improved algorithm achieved 86% average accuracy by the hot topic identification in Network. Furthermore, Ref. [20] used deep learning and OCC model to establish emotion rules to solve the problem of a lack of semantic understanding. Their work obtained 90.98% accuracy of emotion recognition in network public opinion. By comparing we found that the element co-occurrence method is significantly better than others.

6 Conclusion

This paper, based on the nature of public opinion, proposed an online public opinion detection method for specific field (element co-occurrence method), and gave the detailed implementation. Different with traditional methods, element co-occurrence method starts at people's recognition of public opinion. Through construct the language

knowledge system of a specific field, this method can not only generate specific field related public opinion topics, but also retrieve the related public opinion information of that field. Based on these, this system can effectively detect the public opinion information. Experiments show that, this method is able to implement in real use, and have relatively good universality.

Acknowledgement. This paper is supported by the National Language Commission (No. ZDI135-4), National Social Science Foundation of China (No. 16BXW023 and AFA170005).

References

1. Liu, Y.: Introduction to Network Public Opinion Research. Tianjin Renmin Press (2007)
2. Zong, C.Q.: Statistical Natural Language Processing. Tsinghua University Press, Beijing (2013)
3. Luo, W.H., Liu, Q., Cheng, X.Q.: Development and analysis of technology of topic detection and tracing. In: Proceedings of JSCL-2003, pp. 560–566. Tsinghua University Press, Beijing (2003)
4. Carbonell, J., Yang, Y.M., Lafferty, J., Brown, R.D., Pierce, T., Liu, X.: CMU report on TDT-2: segmentation, detection and tracking. In: Proceedings of the DARPA Broadcast News Workshop, pp. 117–120 (1999)
5. Hong, Y., Zhang, Y., Liu, T., Li, S.: Topic detection and tracking review. J. Chin. Inf. Process. **21**(6), 71–87 (2007)
6. Li, B.L., Yu, S.W.: Research on topic detection and tracking. Comput. Eng. Appl. **39**(17), 7–10 (2003)
7. Allan, J. (ed.): Topic Detection and Tracking: Event-Based Information Organization. Kluwer Academic Publishers, Norwell (2002)
8. Forgy, E.W.: Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics **21**(3), 768–769 (1965)
9. Allan, J.: Introduction to topic detection and tracking. In: Allan, J. (ed.) Topic Detection and Tracking. The Information Retrieval Series, vol. 12, pp. 1–16. Springer, Boston (2002). https://doi.org/10.1007/978-1-4615-0933-2_1
10. Hong, Y., Zhang, Y., Fan, J., Liu, T., Li, S.: New event detection based on division comparison of subtopic. Chin. J. Comput. **31**(4), 687–695 (2008)
11. Zhang, K., Li, J.Z., Wu, G., Wang, K.H.: A new event detection model based on term reweighting. J. Softw. **19**(4), 817–828 (2008)
12. Zhao, L., Yuan, R.X., Guan, X.H., Jia, Q.S.: Bursty propagation model for incidental events in blog networks. J. Softw. **05**, 1384–1392 (2009)
13. Chen, K.Y., Luesukprasert, L., Chou, S.C.T.: Hot topic extraction based on timeline analysis and multidimensional sentence modeling. IEEE Trans. Knowl. Data Eng. **19**(8), 1016–1025 (2007)
14. Switzerland, Saussure: General Linguistics. The Commercial Press, Beijing (1980)
15. Linguistic Terminology Committee: Linguistic Terms. The Commercial Press, Shanghai (2011)
16. Yang, J.: A research on basic methods and key techniques for monitoring public opinions on language. Ph.D. thesis, Communication University of China (2010)
17. Shi, C.Y., Xu, C.J., Yang, X.J.: Study of TFIDF algorithm. J. Comput. Appl. **29**(s1), 167–170 (2009)

18. Zhang, Y.F., Peng, S.M., Lü, J.: Improvement and application of TFIDF method based on text classification. Comput. Eng. **32**(19), 76–78 (2006)
19. Gesang, D.J., et al.: An internet public opinion hotspot detection algorithm based on single-pass. J. Univ. Electron. Sci. Technol. China **4**, 599–604 (2015)
20. Wu, P., Liu, H.W., Shen, S.: Sentiment analysis of network public opinion based on deep learning and OCC. J. China Soc. Sci. Tech. Inf. **36**(9), 972–980 (2017)