Emotional Tendency Dictionary Construction for College Teaching Evaluation

Weiwen Shen(✉), Shuping Zhang
Wuhan Technical College of Communications, Wuhan, Hubei, China
`weiwenshen1001@21cn.com`

Abstract—To deal with a large amount of emotional information, the text was analyzed by constructing a sentiment dictionary. The mining system was used in the teaching evaluation of colleges and universities. Taking the high-quality emotional dictionary construction algorithm as the research object, a general emotion dictionary construction method based on function optimization was proposed. This method transformed the universal emotion dictionary construction problem into a function optimization problem and used a simulated annealing algorithm to solve it. A universal sentiment dictionary was constructed using the Modularity optimization method in the discovery of complex web communities. In addition, the traditional Modularity method was improved. The results showed that the improved method only compared the Modularity values for all bipartite cases and optimized them. This not only made Modularity method applicable to this problem, but also greatly reduced the amount of computation. In summary, the traditional information clustering method is improved. By making full use of the relationship between the emotional words and documents in the source domain and the target domain, the domain emotional dictionary in the target domain is established.

Keywords—sentiment analysis, emotional dictionary, function optimization, iterative enhancement

1 Introduction

Internet technology is one of the greatest innovations in human communication. In 1969, Professor Leonard Kleinrock of the University of California, USA, connected several computers to switches and routers. Over the past 40 years, the Internet has grown from the original small internal intranet (ARPANet) to today's household Internet. Thus, Internet leads people into a new era of information. With the introduction and application of many new concepts such as e-government, e-commerce, home appliance informatization, the network constantly changes the way people live, work and learn. It has become an inseparable part of the development of various industries and people's lives.

With the rapid development of the Internet and the advent of the web2.0 era, the network has gradually changed from a simple static information carrier to a platform

for people to express their opinions and display their individuality. More and more people are willing to record their feelings (blogs) on the Internet, express their views on policies (news reviews), post their own evaluations of products (product reviews), and so on. Compared with traditional structured data, its representations are mostly in the form of unstructured or semi-structured comment texts. Faced with such a large amount of emotional information-rich texts, a quick way to analyze attitudes on large-scale texts was proposed. Text orientation analysis research is carried out in this context. In the business world, it provides valuable information for companies in market analysis, market research, and customer feedback. In the field of management, it can help leaders understand the feedback from the public on various policy measures. The government and the masses have a positive interaction.

Emotion is a broad concept. It involves people's views and evaluations, including the evaluation of human behavior relative to social standards, the evaluation of products relative to national and industry mandatory standards, and aesthetics. By mining and analyzing subjective information such as positions, opinions, opinions, emotions, likes and dislikes in the text, the emotional tendencies of the text are judged. The sentimental tendency of a text refers to the positive or negative tendency of the text and the intensity of its emotional tendency. From the current research, textual emotion categories are usually divided into two categories (positive, negative) or three (positive, negative, and neutral), and most of them consider two categories. Positive (supportive, healthy) attitudes and positions in the theme are defined as positive categories. Negative (opposing, unhealthy) attitudes and positions in the subject are defined as negative categories. Neutral attitudes and positions in the theme are defined as neutral categories.

Text mining is applied to the evaluation of college teachers' educational technology ability training. Multidisciplinary approaches such as text mining, information extraction, information retrieval, machine learning, natural language processing, statistics, corpus linguistics, ontology, cognitive science, and visualization techniques are applied. The teacher training model was innovated and the quality of training was improved. The teacher training in colleges and universities is truly implemented to meet the training needs of teachers.

2 State of the art

At present, text orientation analysis research has become a hot spot of interest for many researchers. In recent years, several international top conferences (AAAI, ACL, CIKM, COLING, SIGIR, WWW, and SIDKDD) in the fields of natural language processing, artificial intelligence, information retrieval, data mining, and web applications have included the related papers of text sentiment analysis. Text orientation analysis has a broad field, including: subjective classification, emotional polarity, semantic orientation, perspective mining, opinion extraction, sentiment analysis, and emotional abstraction.

In recent years, text-oriented analysis techniques have gradually been applied to many fields. For example, the business intelligence system Pulse developed by Microsoft can use text clustering technology to extract the user's view of product details

from a large amount of comment text data. The product comment mining system Opinion Observer uses the rich customer comment resources on the network to analyze and process the subjective content of the comments. The individual characteristics of the product and the evaluation of the consumer are extracted and a visual result is given.

The purpose of the text subjective analysis study is to separate the objectivity document describing the fact from the subjective document expressing the opinion. The supervisor expression in a document is distinguished from the objective statement. Many studies have shown that text subjective classification has a very close relationship with document tone classification. Before classifying document mood, subjective classification technology can remove part of documents which are irrelevant to mood expression or difficult to classify classifier. Xie et al. [1] pointed out that the document size was greatly compressed by subjective classification techniques. However, the compressed document can obtain a tone classification result similar to the original text.

Chou et al. [2] proposed two lexical semantic orientation calculation methods based on semantic similarity and semantic correlation field. The tendency of the target vocabulary is obtained by calculating the similarity between the target vocabulary and the marked vocabulary in HowNet. Jung et al. [3] used HowNet as the benchmark. The semantic tendency of the target vocabulary is determined by calculating the degree of association between the target word and the reference word. Yuan et al. of the City University of Hong Kong studied the automatic acquisition of Chinese polar words based on the work of Turney. Zadeh et al. [4] proposed the polar coordinate method of word preference and used the equalized mutual information method to explore the self-propensity of words independent of context. Phu et al. [5] used the NEAR operator provided by the search engine AltaVista to get the number of pages containing two words. It is normalized as the co-occurrence rate of these two words, and then the word similarity is calculated. Product reviews are classified at the document level, i.e., "recommended" and "not recommended". However, the product has multiple attributes. Consumers are only interested in one or more of these attributes. Moreover, this situation is the most common, that is, the product has advantages in some respects and has disadvantages on the other. Therefore, Haselmayer and Jenny [6] believe that it is not enough to just classify the entire document. In many cases, even a single statement can express several opposite tone.

Song et al. [7] constructed an emotional dictionary by using semantic relations between words provided by semantic dictionaries such as Wordnet. The underlying assumption of this method is that two words with strong semantic relations have the same semantic tendency. In general, the construction of the semantic relationship of words is used. Then, the synonymous relationship in the semantic dictionary is used to add edges. Then, some methods based on graph theory can be used to calculate the "distance" between two words. Tsoulos et al. [8] used the word relationship provided by Wordnet to construct an undirected graph of words. Then, the word similarity is obtained by calculating the shortest path between the two words in the figure. However, this method only considers the synonymous relationship of words, but does not consider the antisense relationship and the hyponymy relations.

Product review mining systems also typically use opinion summarization techniques to summarize online product reviews by summarizing the polarity, extent, and related

events of the comments. With this technology, potential users can easily understand the current consumer evaluation of the product. Product manufacturers and distributors can also easily track consumer evaluations of products and the strengths and weaknesses of similar brands. The system uses three subsystems to process a series of online reviews about a product. The product attributes of the consumer's comments are identified. The tone of the comment for each attribute is judged to determine the praise and bad review. The above information is used to generate a summary of the comments.

The domain migration problem of classification is solved by combining the annotation text of the old domain with the unlabeled text of the new domain. The basic idea is as follows: First, the old domain classifier is used to classify n representative samples in the new field (n is the default parameter). These classified samples are then used to train the classifiers in the new field. Finally, the classifier is used to classify all test documents in the new field.

In summary, domain emotion dictionary construction is a relatively new research point in the field of tone analysis. The semi-supervised thinking in the field of machine learning is introduced into the emotional dictionary construction technology to try to solve the problem of domain emotional dictionary construction. The research of emotional dictionary construction technology is the basis of text orientation analysis research. It is of great practical significance to promote the development of text orientation analysis technology, to exert the potential of text orientation analysis and to promote its practicalization and commercialization.

3 Methodology

3.1 Word semantic propensity calculation based on function optimization

Two words with greater similarity are more likely to have the same semantic bias. In this way, the problem of semantic propensity calculation of words can be attributed to the division of undirected graphs, so that the sum of similarities of node subgraphs with the same symbol is the largest. At the same time, the sum of the similarities of the node subgraphs with different symbols is minimized. In this way, the semantic tendency of each word in the figure is determined. Based on this assumption, the problem is formalized as follows:

Definition: W is a set of words containing all semantic tendencies to be determined. The number of words included is N=|W|, and the connection weights of words i and j are:

$$W_{ij} = \begin{cases} 1 & \text{if there is an edge joining vertice } i, j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The emotional word classification problem is considered. Since the size of the subgraph is not known in advance, it cannot be assumed that the two subgraphs of the positive and negative tone are approximately equal in size. The graph is divided into "minimum segmentation", and the objective function must meet the following

conditions: First, reward the inner side of the subclass; second, punish the non-connected side of the subclass; third, punish the connected side between the subclasses; and fourth, reward the non-connected side between the subclasses. At the same time, the conditions that the objective function satisfies can be classified into two categories: The first condition and the second condition are used to increase the cohesion of the subclass. The third condition and the fourth condition are used to reduce the coupling between subclasses. In this way, a more extensible word semantic tendency calculation framework is obtained. The relationship between words is used to construct an undirected graph of words. Both dictionary-based and corpus-based methods are used. The problem of word semantic propensity calculation is transformed into graph partitioning problem and further transformed into function optimization problem. The "minimum segmentation" idea design is the objective function. The solution algorithm is constructed to solve the objective function. The simulated annealing algorithm is used to solve the problem.

Vocabulary similarity calculation is an important and basic work in the fields of natural language processing, information retrieval and information extraction. The goal is to measure the degree of similarity between words. Usually, the similarity value is defined as a real number between 0 and 1, and the larger the absolute value, the higher the similarity. At present, there are two main ideas. One is to use statistical methods to analyze the law of word distribution in large-scale corpus and obtain the similarity of words. The other is a dictionary-based approach, such as the English dictionary Word-Net and HowNet. The similarity calculation method based on corpus statistics and the word similarity calculation method provided by HowNet are used as the basis for constructing the undirected network graph of words.

As a huge corpus, the value of the Internet is recognized. The traditional method of calculating the similarity based on the word co-occurrence rate is appropriately changed so that it can be applied to the Internet corpus. In the following formulas, H(P) represents the number of returned pages obtained by entering the query P in the search engine. P∩Q represents a joint query of the word P and the word Q. Due to the noise in the network data, the two words co-occurring in some web pages may be accidental. To reduce this effect, the threshold is defined as c. If the number of pages H(P∩Q) returned by the joint query P∩Q is less than the threshold, the similarity between the words P and Q is set to zero. The formulas are defined as follows:

$$PCJaccard(P,Q) = \begin{bmatrix} 0, & if & H(P \cap Q) \leq c \\ \dfrac{H(P \cap Q)}{H(P) + H(Q) - H(P \cap Q)}, & & otherwise \end{bmatrix}$$

(2)

$$PCOverlap(P,Q) = \begin{bmatrix} 0, & if & H(P \cap Q) \leq c \\ \dfrac{2H(P \cap Q)}{\min(H(P), H(Q))}, & & otherwise \end{bmatrix}$$

(3)

$$PCDice(P,Q) = \begin{bmatrix} 0, & if & H(P \cap Q) \leq c \\ \dfrac{2H(P \cap Q)}{H(P) + H(Q)}, & & otherwise \end{bmatrix}$$

(4)

3.2 Markov chain description of simulated annealing algorithm

The idea of simulated annealing was introduced. The process of solving the problem is transformed into the process of searching for the optimal solution in the solution space of the objective function. Simulated annealing algorithm is a random optimization algorithm based on Monte Carlo iterative solution strategy. The starting point is based on the similarity between the annealing process of solid matter in physics and the general combinatorial optimization problem. The Markov chain is an important mathematical tool for analyzing simulated annealing algorithms. Its description is as follows:

Let $\Omega = \{s1, s2, \cdots\}$ is the solution space for all states. $X(k)$ is the value of the variable at time k. The random sequence is called the $\{X(k)\}$ Markov chain.

$$P\{X(n) = j | X(n-1) = i\}$$

(5)

For a one-step transition probability, the n-step transition probability is:

$$p_{i,j}^{(n)} = \Pr\{X(n) = j | X(0) = i\}$$

(6)

3.3 Experiment and analysis

The experiment used Chinese commentary data on three topics: emotional blog, movie review and laptop. All data was collected from relevant Chinese comment sites on the Internet. Comments on the same topic may appear on different review sites. To prevent duplicate samples from appearing in the data set, a specific collector is specified for a particular URL address. After the corpus is collected, it is extracted and converted into a unified text format, and the polarity is manually labeled. Finally, the experimental data set was obtained, as shown in Table 1:

Table 1. The composition of the experimental document test set

Domain	Total	Positive	Negative
Blog	1003	508	495
Movie Review	998	500	488
Notebook	901	450	451
Total	2892	1458	1434

The semantic tendency judgment of words has uncertainty and manifests in two aspects: First, some words have different semantic tendencies in different pragmatic environments. For example, "thinness" is a derogatory term; in the field of notebooks, it is a derogatory term. Secondly, for the same word, such as "tough", the judgment of different people is also different. To reduce the impact of the above factors, when generating a word test set from a document test set, a method of co-labeling by multiple people is adopted. When constructing a universal domain test set, words related to semantic propensity and pragmatics should be avoided as much as possible.

The relationship between the objective function and the accuracy is verified by experiments. In the iterative process of the simulated annealing algorithm, some of the solutions are intercepted. For ease of observation, these solutions are ordered by energy. A graph of the relationship between the objective function value and the accuracy of the semantic propensity is drawn. The results of this experiment are shown in Figure 1.

When some noise is excluded, it can be seen that the objective function is inversely related to the accuracy. As the energy value gradually decreases, the accuracy of the result is gradually decreasing. This illustrates the agglomeration of word networks constructed with word similarity. In addition, the rationality of the objective function is verified.

In the SOSA algorithm, the rate of decline of the initial temperature, and the number of inner loops of the program at each temperature value have a large influence on the convergence speed and the accuracy of the result. When using this algorithm, the parameters need to be adjusted to make a trade-off between the accuracy of the results and the convergence time. The effects of the number of iterations on the accuracy and convergence time of the results were tested separately, as shown in Figure 2 and Figure 3.
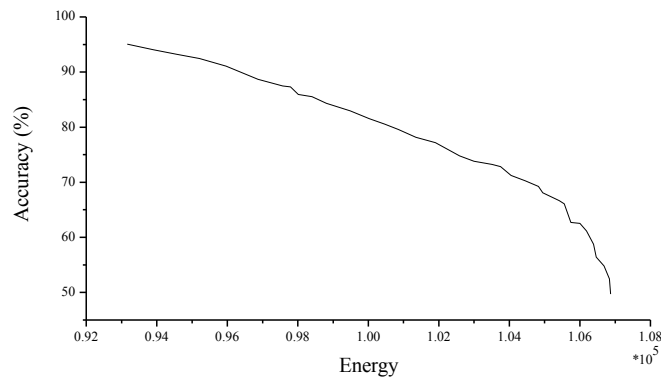

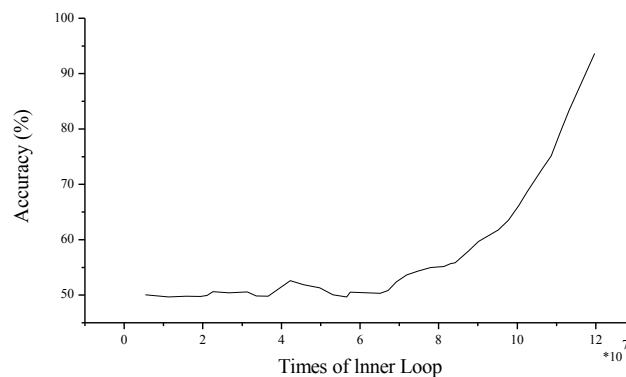**Fig. 1.** The relation curve of the energy value and the accuracy rate

Fig. 2. The impact of the number of inner loop execution on accuracy
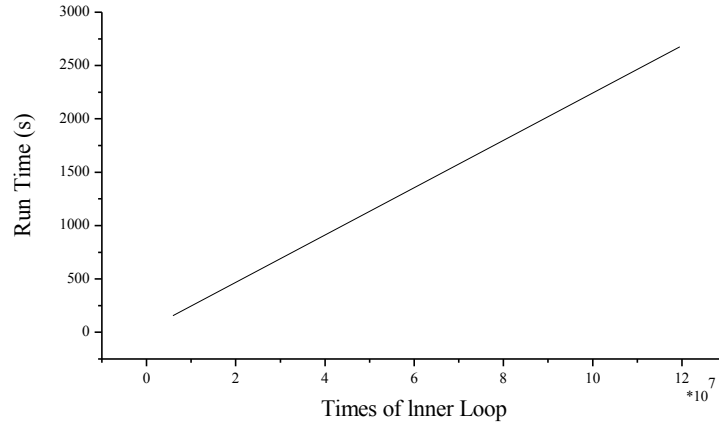
Fig. 3. The impact of the number of inner loop execution on running time

According to the above experiment, the initial temperature is set as the sum of the two similarities of all words in the word net. The temperature drop rate was set to 98%. At the same time, to ensure that the algorithm can try more possibilities at each temperature value, the number of inner loops is set to the number of words in the word network.

In Table 2, two methods based on HowNet similarity (HowNetPMI and HowNetSA) are compared. The proposed method can make full use of the global information between words. The simulated annealing algorithm can approximate the characteristics of the optimal solution with probability 1. The HowNetSA method has a larger performance improvement than the HowNetPMI method. By comparing the "SA" class algorithms (PCJaccardSA, PCOverlapSA, PCDiceSA, PCPMISA, HowNetSA), it can be found that PCDiceSA has an accuracy comparable to HowNetSA. When the number of corpora is sufficient, the co-occurrence-based method can achieve higher accuracy of word semantic propensity, which is consistent with the experimental conclusions in the existing research.

Table 2. Vocabulary experiment results

Approach	Test set	Accuracy	Approach	Test set	Accuracy
How Net PMI	Test set1	78.6%	PCDice SA	Test set1	92.6%
	Test set2	80.7%		Test set2	90.3%
	Test set3	81.2%		Test set3	90.1%
PCJaccard SA	Test set1	88.9%	PCPMISA	Test set1	89.7%
	Test set2	87.8%		Test set2	88.2%
	Test set3	88.2%		Test set3	87.7%
PCOverlap SA	Test set1	88.4%	How Net SA	Test set1	95.0%
	Test set2	86.9%		Test set2	90.3%
	Test set3	87.4%		Test set3	90.6%

To further verify the results of the semantic propensity calculations of the words, the words in the testset1 that have calculated the attributes are used to classify the documents. The document is considered together. The following method is used to calculate the accuracy of the document:

$$Accuracy = \frac{a + d}{a + b + c + d} \tag{7}$$

With the improvement of the accuracy of lexical semantic tendency calculation, the accuracy of document tendency classification is also improved, which proves the validity and practicability of the method.

A general framework for calculating the semantic tendency of words is proposed. First, the inter-word relationship is used to construct an undirected graph of words. The problem of word semantic propensity calculation is transformed into graph partition problem, and further transformed into function optimization problem and solved. According to this framework, a dictionary-based word similarity calculation method and a word undirected graph constructed based on the word co-occurrence rate similarity calculation method are tested. Then, the "minimum segmentation" is taken as the objective function, and the simulated annealing algorithm is used to solve the problem.

Experiments show that the framework has good scalability and robustness. Scalability is reflected in the various methods of calculating word similarity and solving them using various heuristic algorithms. Since the simulated annealing algorithm can approximate the characteristics of the optimal solution with probability 1, the method has better robustness.

4 Result analysis and discussion

4.1 Word semantic propensity calculation based on Modularity optimization

Complex network community discovery research is an extension and deepening of the graph decomposition method. It is mainly for the case where the number and size of subgraphs are uncertain. At present, many methods have been proposed, such as edge density, median, and random walking. A representative approach is based on modularity optimization. The basic idea of Modularity is that a completely random network has no community structure. If a network has a good community structure, there is a division of the network, so that this division corresponds to a higher modularity value. By comparing the statistical differences between the real network and the random network, the method can find a more "natural" partition of the network, thereby avoiding the weakness of the traditional graph partitioning method.

Modularity-based optimization methods are widely used in community discovery. However, this method has not been used to calculate the semantic bias of words. This method can not only make full use of the global information between words, but also avoid the shortcomings of the figure division method which is easy to find ordinary solutions. This method is used to calculate the propensity of word semantics.

The following steps are taken for word semantic propensity calculations:

The construction of the word similarity matrix: two-word similarity calculation methods are used to construct the word similarity matrix. The first method is to use the similarity function provided by How Net. The second method is to use co-occurrence information of words in the corpus.

The calculation of the semantic tendency of words: based on the word similarity matrix, modularity is the objective function. It is divided into two disjoint subgraphs in such a way that the function value is extremely large. The modularity matrix is constructed by the word similarity adjacency matrix. A feature vector corresponding to the largest eigenvalue is found. Each element in the vector corresponds to each word of the semantic tendency to be calculated. These words are divided into two categories according to the positive and negative of the eigenvalues. For each type of word, the semantic tendency of the largest word in the class is manually determined, and the semantic tendency is used as the semantic tendency of the category. Words are continually exchanged between the two classes until the modularity value is stable. The semantic tendency of each word in the test set is returned.

There are two main ideas for the existing word similarity calculation methods. The first is to obtain the similarity of words by analyzing the distribution of words in the corpus. The second is a dictionary-based approach. The similarity calculation method based on corpus statistics and the word similarity calculation method provided by the semantic dictionary How Net are used to construct the word similarity matrix.

4.2 Experiment and analysis

To verify the rationality of the objective function, the modularity value of all solutions of the modularity optimization process and the corresponding semantic tendency calculation accuracy are recorded. The accuracy of the solution and the corresponding modularity value in the process of optimizing the termset3 are recorded in Figures 4 and 5.
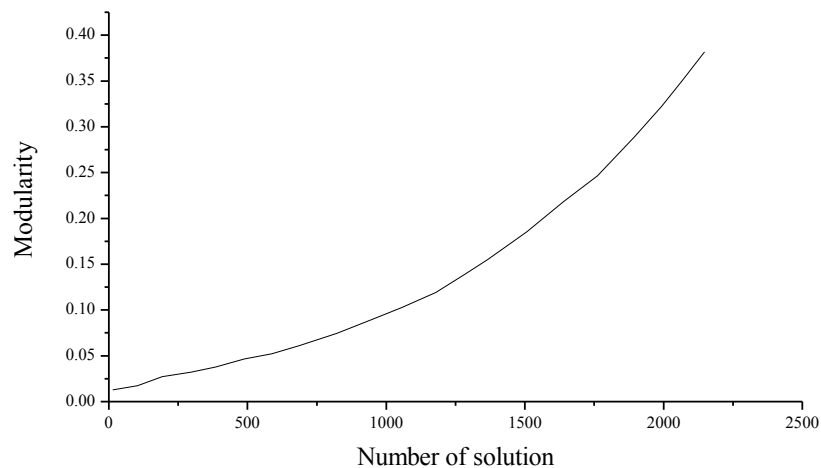

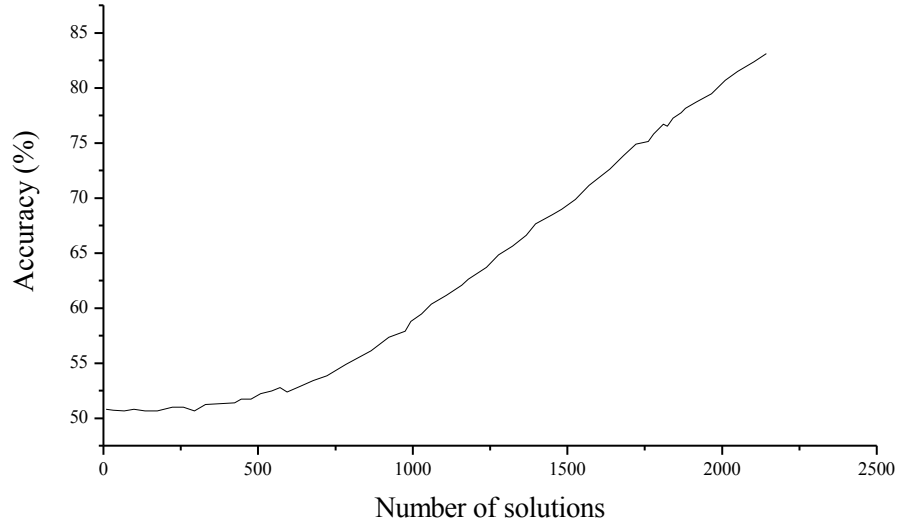**Fig. 4.** Q value change curve

Fig. 5. Changes in the accuracy of Q value

As can be seen from Figure 4 and Figure 5, as the Q value increases, the accuracy of the semantic tendency of the words increases, and the magnitude of the change is basically the same, thus verifying the rationality of the objective function. In addition, when the Q value is small (less than 0.05), the degree of coincidence of the Q value and the curve of the accuracy is lower than the case where the Q value is large. When the Q value is small, the community results are relatively insignificant and the nodes may be classified into any subclass. As the Q value increases, the community structure becomes more and more obvious, and nodes can more easily find suitable subclasses. In this experiment, the maximum value of Q was 0.38. In general, a Q value greater than 0.3 usually indicates that a better network partition was found.

Table 3. Experimental results on the co-occurrence test set

Approach	Test set	Accuracy	Average Accuracy	Approach	Test set	Accuracy	Average Accuracy
PMI	4	0.457	0.453	K–L	4	0.618	0.604
	5	0.456			5	0.598	
	6	0.446			6	0.596	
Proposed Approach	4	0.592	0.583				
	5	0.583					
	6	0.576					

The overall test is based on the accuracy of the modularity optimization method on the test set generated by HowNet and the co-occurrence test set. As a comparison, the PMI method and the K-L method are implemented. As can be seen from Table 3, on the three co-occurrence test sets, the K-L method and the proposed method are

significantly better than the PMI method using only the local information between the words and the reference words, since the global information between the words is fully used. Moreover, the accuracy of the proposed method is higher than that of the graph decomposition method (K-L method). In addition, the accuracy of the proposed method on the three test sets generated by different scale corpora is basically stable.

5   Conclusions

To reduce the dependence of the general sentiment dictionary construction algorithm on the benchmark words, a general emotion dictionary construction method based on function optimization is proposed from the perspective of graph division. This method transforms the general sentiment dictionary construction problem into a function optimization problem and solves it by using the simulated annealing algorithm. The results show that the method has higher accuracy and is relatively insensitive to the change in the number of reference words. To solve the problem that the graph partitioning method is easy to fall into the local extremum, the Modularity optimization-based method in complex network community discovery is used to construct the general sentiment dictionary. The traditional Modularity method is improved. The Modularity values for all binary cases are compared and optimized. This makes the Modularity method suitable for this problem, and greatly reduces the amount of computation. Experiments show that the method has higher accuracy and is relatively insensitive to the change in the number of reference words.

6   References

[1] Xie, H., Li, X., Wang, T., Lau, R. Y. K., Wong, T. L., & Chen, L., et al. (2016). Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy. Information Processing & Management, 52(1): 61-72 https://doi.org/10.1016/j.ipm.2015.03.001

[2] Chou, P. H., Tsai, T. H., & Hsu, Y. J. (2017). Context-aware sentiment propagation us-ing lda topic modeling on chinese conceptnet. Soft Computing, 21(11): 1-11 https://doi.org/10.1007/s00500-016-2273-0

[3] Jung, H., Park, H. A., & Song, T. M. (2017). Ontology-based approach to social data sentiment analysis: detection of adolescent depression signals: Journal of Medical In-ternet Research, 19(7): e259 https://doi.org/10.2196/jmir.7452

[4] Zadeh, A., Zellers, R., Pincus, E., & Morency, L. P. (2016). Multimodal sentiment in-tensity analysis in videos: facial gestures and verbal messages. IEEE Intelligent Sys-tems, 31(6): 82-88 https://doi.org/10.1109/MIS.2016.94

[5] Phu, V. N., Chau, V. T. N., Tran, V. T. N., & Dat, N. D. (2017). A vietnamese adjective emotion dictionary based on exploitation of vietnamese language characteristics. Ar-tificial Intelligence Review, 21(4): 1-67

[6] Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. Quality & Quantity, 51(6): 2623-2646 https://doi.org/10.1007/s11135-016-0412-4

[7] Song, W., Deng, Z., Wang, S., & Feng, X. (2016). A simple model predictive power control strategy for single-phase pwm converters with modulation function optimization. IEEE Transactions on Power Electronics, 31(7): 5279-5289

[8] Tsoulos, I. G., Tzallas, A., & Tsalikakis, D. (2016). Pdoublepop: an implementation of parallel genetic algorithm for function optimization. Computer Physics Communica-tions, 209: 183-189 https://doi.org/10.1016/j.cpc.2016.09.006

7 Authors

Weiwen Shen and **Shuping Zhang** are with Wuhan Technical College of Communications, Wuhan, Hubei 430065, China (weiwenshen1001@21cn.com).