Computer and Information Science
www.ccsenet.org/cis
Vol. 3, No. 1
February 2010

Infer the Semantic Orientation of Words by Optimizing Modularity

Weifu Du (Corresponding author)
School of Computer Science, Harbin Institute of Technology
92 Xi Da Zhi Street, Harbin 150001, China
Tel: 86-10-6260-0941 E-mail: duweifu@software.ict.ac.cn

Songbo Tan
Institute of Computing Technology, Chinese Academy of Sciences
6 Ke Xue Yuan Nan Street, Beijing 100190, China
Tel: 86-10-6260-0941 E-mail: tansongbo@software.ict.ac.cn

This work was mainly supported by two funds, i.e., 0704021000 and 60803085, and one another project, i.e., 2004CB318109)

Abstract

This paper proposes a novel algorithm, which attempts to attack the problem of word semantic orientation computing by optimizing the modularity of the word-to-word graph. Experimental results indicate that proposed method has two main advantages: (1) by spectral optimization of modularity, proposed approach displays a higher accuracy than other methods in inferring semantic orientation. For example, it achieves an accuracy of 88.8% on the HowNet-generated test set; (2) by effective usage of the global information, proposed approach is insensitive to the choice of paradigm words. In our experiment, only one pair of paradigm words is needed.

Keywords: Sentiment analysis, Opinion mining, Information retrieval

1. Introduction

In the Web2.0 era, the Internet turns from a static information media into a platform for dynamic information exchanging, on which people can express their views and show their individualities. More and more people are willing to record their feelings (blog), give voice to public affairs (news review), express their likes or dislikes on products (product review), and so on. In the face of the increasing volume of sentimental information available on the Internet, there is a growing interest in helping people to better find, filter, and manage these resources.

Automatic sentiment analysis could play an important role in a wide variety of flexible and dynamic information management tasks. For example, with the help of sentiment analysis system, in the field of public administration, the administrators could receive the feedback on one policy in a timelier manner; in the field of business, manufacturers could perform more targeted updates on products to improve the consumer's experience.

Sentiment analysis can be considered as texts classification according to different opinions they hold. A frequently used method is to label the sentiment words manually (Turney and Littman, 2003). The essential idea is to manually label the semantic orientation of words that in common use, such as "good" labeled as "positive" and "bad" labeled as "negative". When classifying, count directly the numbers of subjective words, then the text was sentenced to positive if it contained more positive orientation words, or negative otherwise. Therefore, it is a fundamental and important task to infer the semantic orientation of words (i.e., for a list of words, partition it into two disjoint sub-lists with semantic orientation: one is positive and another is negative). This paper aims to automatically construct such sub-lists from glosses in a lexicon, as well as from a corpus.

Most of previous methods use word-to-word similarity and some paradigm words (i.e. some representative words with pre-labeled semantic orientation, positive or negative) to infer the semantic orientation of words. The basic observations underlying these methods are quite different from each other. However, these methods could roughly be classified into two categories in terms of the manner of using the word-to-word similarity.

The first kind of approaches uses local information to infer the semantic orientation of words (Turney and Littman, 2003). When computing, only the relationship between the word and paradigm words is taken into account, while the relationship between the words and other words in the test set is ignored, which makes they are sensitive to the choice of the paradigm words.

The second kind of approaches can avoid this drawback by the usage of global information (the relationship of the word with not only paradigm words, but also other words in the test set) (Hatzivassiloglou and McKeown, 1997; Kobayashi et al., 2001). The graph based approaches are the typical ones (Kamps et al., 2004; Hu and Liu, 2004; Andreevskaia and Bergler, 2006; Esuli and Sebastiani, 2006; Pang et al., 2004; Takamura et al.,2005). The essential idea of these approaches is 'minimum cut' that is to look for divisions of the vertices into two subgroups so as to minimize the number of edges running between the subgroups. However, if subgroup sizes are unconstrained then we are, for instance, at liberty to select the trivial division of the network that puts all of the vertices in one of the two subgroups and none in the other, which guarantees we will have zero intergroup edges. This division is, in a sense, optimal, but clearly it does not tell us anything of any worth.

Several approaches have been proposed to get around this problem. For instance, the ratio cut method (Wei and Cheng, 1989) minimizes not the simple cut size but the cut ratio. The ratio cut method does allow some leeway for the sizes of subgroups to vary around their specified values, which makes it more flexible than the simple minimum cut method, but at its core it still suffers from the same drawbacks that they require in advance the sizes of subgroups, which would be determined after the computing.

The study of community finding is the extension of the graph partition and many algorithms had been proposed to find more nature community, such as edges density (Palla et al., 2005), betweeness (Newman and Girvan, 2004), information centrality (Fortunato et al., 2004), random walk (Pons and Latapy, 2005), spectral analysis (Newman, 2006) etc. Modularity optimization based methods (New-man, 2004; Reichardt and Bornholdt, 2006) are the typical ones. Different from conventional graph-partitioning based algorithms' manner of counting edges directly, modularity based algorithm takes the hypothesis that a good division of a network into subgroups is not merely one in which there are few edges between subgroups; it is one in which there are fewer than expected edges between subgroups. This makes the algorithm successful in finding the communities that the sizes of which are unknown in advance.

The efficiency of modularity based algorithm motivates us to take it as the fundamental framework of our method, as far as we know, this approach has not been employed in inferring the term semantic orientation yet.

As a result, we propose an algorithm based on spectral optimization of modularity matrix to infer the semantic orientation of terms. Modularity (Newman, 2004) is a measure to evaluate the goodness of a partition of matrix. Therefore we construct a quality function in the use of modularity (*Q* value) and then partition the modularity matrix by the manner of making the objective function (*Q* value) have the maximum value.

2. Proposed algorithm

The proposed word semantic orientation inferring method consists of two steps: (1) the modularity matrix is built to reflect the semantic relationship between words; (2) based on the modularity matrix, an spectral optimization based algorithm is imposed to obtain the semantic orientation of every word.

2.1 Matrix building

Given the word collection $T = \{t_j \mid 1 \le j \le m\}$ of a document, the semantic similarity between any two words t_i and t_j can be computed using approaches that are either knowledge-based or corpus-based.

In this study, we simply choose the mutual information to compute the semantic similarity between word t_i and t_j as follows:

$$sim(t_i, t_j) = \log \frac{N \times p(t_i, t_j)}{p(t_i) \times p(t_j)}$$

which indicates the degree of statistical dependence between t_i and t_j. Here, N is the total number of words in the corpus and $p(t_i)$ and $p(t_j)$ are respectively the probabilities of the occurrences of t_i and t_j, i.e. $count(t_i) / N$ and $count(t_j) / N$, where $count(t_i)$ and $count(t_j)$ are the frequencies of t_i and t_j. $p(t_i, t_j)$ is the probability of the co-occurrence of t_i and t_j within a window with a predefined size k, i.e. $count(t_i, t_j) / N$, where $count(t_i, t_j)$ is the number of the times t_i and t_j co-occur within the window.

We use an adjacency matrix $A=[A_{ij}]_{m\times m}$ to describe the initial word-to-word relationship, where $A_{ij}=sim(t_i, t_j)$, if $i \neq j$ and $A_{ij}=0$ if $i=j$. Then A is normalized to make the sum of each row equal to 1.

Let k_i be the degree of the vertex i and $m = 1/2\sum_i k_i$, which denotes the total number of the edges in the matrix. Then, we can build the modularity matrix $B=[B_{ij}]_{m\times m}$, where $B_{ij}= A_{ij} - k_ik_j/2m$.

2.2 Semantic orientation inferring

Based on the modularity matrix *B*, we take a spectral optimization based algorithm to infer the semantic orientation of words.

In computing, we find the single eigenvector of the modularity matrix *B* corresponding to the most positive eigenvalue firstly. This is most efficiently achieved by the direct multiplication or power method. Starting with a trial vector, we repeatedly multiply by the modularity matrix and—unless we are unlucky enough to have chosen another eigenvector as our trial vector—the result will converge to the eigenvector of the matrix having the eigenvalue of largest magnitude. In some cases this eigenvalue will be the most positive one, in which case our calculation ends at this point. In other cases the eigenvalue of largest magnitude may be negative. If this happens then, denoting this eigenvalue by β_n, we calculate the shifted matrix $B\text{-}\beta_n I$, which has eigenvalues $\beta_i\text{-}\beta_n$ (necessarily all nonnegative) and the same eigenvectors as the modularity matrix itself. Then we repeat the power-method calculation for this new matrix and this time the eigenvalue of largest magnitude must be $\beta_1\text{-}\beta_n$ and the corresponding eigenvector is the one we are looking for.

When come here, we only get an approximate division, and there is room for improvement of the solution. Then we move single vertices between the subgroups so as to increase the value of the modularity as much as possible, with the constraint that each vertex can be moved only once. Repeating the whole process until no further improvement is obtained, we find a final value of the modularity, and get the semantic orientation of every word in the test set.

3. Datasets and experimental setup

We download texts from the Internet, which including comments on education (from http://blog.sohu.com/learning/), electronics (from http://detail.zol.com.cn/) and stock (from http://blog.sohu.com/stock/). The detail information is illustrated in table 1.

We use ICTCLAS, a Chinese word segmentation software, to extract words (including adjectives, adverbs, nouns, and verbs) from these texts. For each word, if it also occurs in *HowNet*, it is inserted into termset1.

After scanning the words in this set, we find many words either have no sentiment at all or will show distinct orientation in different context, so we ask three people to select words that are considered full of sentiment and as definite as possible by them. Finally, we label each word with the semantic orientation that agreed by the most of people and use the similarity provided by *HowNet* to build the word-to-word matrix. Termset2 and termset3 are constructed in this process. The detail information of individual test set generated by *HowNet* is illustrated in table 2.

There are other three test sets generated by the means of co-occurrence based term similarity computing method. In general, the scale of the corpus and the size of the co-occurrence window will affect the result of term similarity. To test the impact of the variation of corpus's scale, we use the whole corpus to generate termset4; and then followed by decreasing the corpus size by 10%, we generate termset5 and termset6.

For the sparsity of the corpus (nearly almost all the terms occur in one text only one time), we set the co-occurrence window with the range of a whole text. After removing the isolate terms (not co-occur with any other terms at all), we get these three term sets. The detail information of individual test set is illustrated in table 3.

We later compare our approach with Turney and Littman's PMI approach.

4. Experimental results and conclusions

4.1 Performance comparison

The selection of the paradigm words is a pivotal step in the PMI method, and the accuracy of solutions is affected greatly by the choice of the paradigm words. To illustrate this, we ask four people to select some pairs of representative words respectively as the candidate paradigm words. Then put these words into the search engine, google, and sort these words by the related page count returned by the search engine, finally, we take the top 20 pairs of words as the paradigm words.

Table 4 shows the detail information of the paradigm words. Column 1 is the word pair ID, then for positive words and negative words, the word (in English) and related page count (unit is million) are listed respectively.

Then we conduct experiment to observe the fluctuation of solution in the PMI method and the proposed method. As we mentioned, the paradigm words were removed from the testing words for our experiments. The detail information is illustrated in figure 1.

Six curves are plotted in figure 1, one for each of the performance of the two approaches on the three term sets generated by the *HowNet* similarity. The three blue curves are PMI approaches, and the three red curves are our proposed approaches.

Seeing the figure1, we can find that the choice of paradigm words affect the accuracy of the PMI solutions very much, though the rise in accuracy correlates with the increase of the paradigm word set, which is one of the motivations of us to find a novel method to solve the problem of identifying the semantic orientation to reduce the dependence on the paradigm words. From this figure, we can find the proposed approach is insensitive to the choice of paradigm words.

We evaluate the performance of our method in terms of the comparison with PMI method with 20 pairs of paradigm words which is mentioned in table 4. These comparisons indicate the validity of the proposed method. Table 5 shows the performance comparison on test sets generated by *HowNet*, table 6 shows the performance comparison on test sets generated by co-occurrence.

From table 5, we can find that in different term sets, with the exclusion of noise, the three approaches all have the enhancement in the accuracy. In the three test sets, the proposed approach outperforms the PMI algorithm. Our proposed method outperforms the baseline method in termset2 and termset3 while is exceeded by the baseline method in the termset1. For the reason that the test set2 and test set3 are refined by people from the test set1, we consider that the words in them display more strong semantic orientation, and therefore they show community structure more evidently, which make the proposed method work efficiently.

From table 6, we can find that the proposed method outperform the baseline approach in all term sets generated by co-occurrence. In this experiment, we find the accuracy of proposed method is stable, which indicate that the proposed approach is relatively insensitive to small scale of corpus.

4.2 Discussions

Seen from experiments above, in termset2 and termset3, which are generated by *HowNet* similarity (see in table 2), the proposed method outperforms both the baseline approach and graph partitioning approach. The high performance achieved by our method benefits from the effective utilization of the global information in the term graph.

In the test sets generated by the co-occurrence information of words in corpus (see in table 3), the performance of the three approaches decline sharp, we consider it is because that the co-occurrence relation of words is more the ‘relatedness’ than the ‘similarity’. Incorporating more information, the similarity provided by lexicons is more reasonable than the co-occurrence information of words in relative small corpus.

Because of the consideration of the expected edges within a group, modularity optimization based method can find more nature community, which contributes to the performance of our approach greatly. In the proposed approach, the expected edges were quantified by the probability manner, if it is computed with a more exact manner, the better solution will be attained.

5. Conclusions and future work

Term semantic orientation computing aims to identify the semantic orientation, commendatory or derogatory, of terms; it is the foundation of text sentiment analysis. In this paper, we present a novel modularity optimization based method to identify the term semantic orientation. The proposed approach attains an accuracy of 88.8% on the *HowNet*-generated test set. The experimental results suggest that our algorithm is effective.

Our proposed method has two main advantages: (1) by spectral optimization of modularity, proposed approach displays a higher accuracy in identifying the term semantic orientation. (2) by effective usage of the global information, only one pair of paradigm words is needed in proposed method.

As future work, for modularity computing, we plan to explore in how to find more exact quantity of the expected edges within a group to substitute the current probability manner. Furthermore, we use the co-occurrence information of words in corpus and similarity function provided by *HowNet* in this paper, if more essential relation between words can be detected, our algorithm could be further improved. Thus, in the further work, we will study how to find more essential relation between words.

References

Alina Andreevskaia and Sabine Bergler. (2006). *Mining WordNet For a Fuzzy Sentiment: Sentiment Tag Extraction From WordNet Glosses*. In Proc. EACL-06, Trento, Italy.

Andrea Esuli and Fabrizio Sebastiani. (2005). *Determining the semantic orientation of terms through gloss classification*. In Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management, Bremen, DE, 2005.

Andrea Esuli and Fabrizio Sebastiani. (2006). SentiWordNet: *A Publicly Available Lexical Resource for Opinion Mining*. In Proceedings of LREC 2006, 5th Conference on Language Resources and Evaluation, Genova May 2006.

Andrea Esuli and Fabrizio Sebastiani. (2006). *Determining term subjectivity and term orientation for opinion mining*. In Proceedings of EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, 2006.

Bo Pang and Lillian Lee. (2004). *A sentimental education: sentiment analysis using subjectivity summarization based* *on minimum cuts*. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, July 21-26, 2004, Barcelona, Spain.

Fortunato S, Latora V, Marchiori M. (2004). *A method to find community structures based on information centrality*. Phys. Rev. E, 2004, 70: 056104.

H Takamura, T Inui, and M Okumura. (2005). *Extracting Semantic Orientations of Words using Spin Model*. In Proceedings of the 43rd Annual Meeting of the ACL, Ann Arbor. 2005: 133-140.

Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten de Rijke. (2004). *Using WordNet to measure semantic orientation of adjectives*. In Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, volume IV. 2004: 1115-1118.

Michael Gamon and Anthony Aue. (2005). *Automatic identification of sentiment vocabulary exploiting low association with known sentiment terms*. In Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in NLP. 2005: 57-64.

Newman M E J, Girvan M. (2004). *Finding and evaluating community structure in networks*. Phys. Rev. E, 2004, 69: 026113.

Newman M E J. (2006). *Finding community structure in networks using the eigenvectors of matrices*. Phys. Rev. E, 2006, 74: 036104.

Newman M E J. (2004). *Fast algorithm for detecting community structure in networks*. Phys. Rev. E, 2004, 69: 066133.

Nozomi Kobayashi, Takashi Inui, and Kentaro Inui. (2001). *Dictionary-based acquisition of the lexical knowledge for p/n analysis* (in Japanese). In Proceedings of Japanese Society for Artificial Intelligence,SLUD-33. 2001: 45-50.

Palla G, Derényi I, Farkas I, and T. Vicsek T. (2005). *Uncovering the overlapping community structure of complex networks in nature and society*. Nature. 2005, 435: 814-818.

Peter D. Turney and Littman M.L. (2003). *Measuring Praise and Criticism: Inference of Semantic Orientation from Association*. ACM Transactions on Information Systems, 2003, 21(4): 315-346.

Pons P, Latapy M. (2005). *Computing communities in large networks using random walks*. Proceedings of the 20th International Symposium on Computer and Information Sciences, volume 3733 of Lecture Notes in Computer Science, Springer, New York. 2005: 284-293.

Reichardt J, Bornholdt S. (2006). *Statistical mechanics of community detection*. Phys. Rev. E, 2006, 74: 016110.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. (1997). *Predicting the semantic orientation of adjectives*. In Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and the Eighth Conference of the European Chapter of the Association for Computational Linguistics. 1997: 174-181.

Y.-C. Wei and C.-K. Cheng. (1989). *Toward efficient hierarchical designs by ratio cut partitioning*. In Proceedings of the IEEE International Conference on Computer Aided Design, pp. 298–301, Institute of Electrical and Electronics Engineers, New York (1989).

Table 1. The detail information of the text sets

TextSet ID	Positive	Negative	Total
Education	254	1012	1266
Stock	364	683	1047
Electronics	1054	554	1608

Table 2. The detail information of the term sets generated by *HowNet*

TermSet ID	Positive	Negative	Total
1	2365	2923	5288
2	1881	2481	4362
3	1098	2039	3137

Table 3. The detail information of the term sets

TermSet ID	Percent of full corpus	Positive	Negative	Total
4	100%	512	664	1176
5	90%	498	647	1145
6	80%	474	641	1115

Table 4. Paradigm words used in the PMI method

Positive		Negative	
Word	Freq (mil.)	Word	Freq (mil.)
Good	2,400	Mistake	214
Active	220	Badness	190
Excellent	219	Agony	96.4
Beautiful	203	Depressed	68.8
Proficient	142	Conservative	44
Mature	127	Worry	43.1
Nice	114	Falsity	41.7
Harmonious	113	Lousy	37.5
good luck	79.7	Terrific	36.5
Peace	78.2	Collapse	34.5
Energy	77.2	Maze	29.4
Comfortable	69.6	Shortcoming	26.2
Fineness	53.7	Misery	24.4
Grateful	50.5	Contort	18.1
Summit	48.6	Phony	15.1
Goodness	43.6	Freaky	12.8
Honest	27	oafish	10.5
Allowance	23.8	Sad	10.4
Glary	23.4	Shame	9.99
Decency	21.5	Asperity	8.50

Table 5. the performance of the two methods on test sets generated by *HowNet*

Approach	Testset ID	Accuracy	Average Accuracy
PMI	1	0.642	0.726
	2	0.694	
	3	0.843	
Proposed Approach	1	0.617	0.741
	2	0.718	
	3	0.888	

Table 6. the performance of the two methods on test sets generated by co-occurrence

Approach	Testset ID	Accuracy	Average Accuracy
PMI	4	0.457	0.453
	5	0.456	
	6	0.446	
Proposed Approach	4	0.618	0.604
	5	0.598	
	6	0.596	

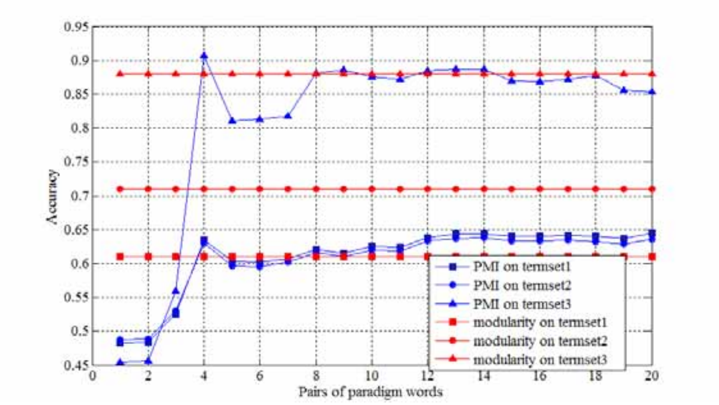

Figure 1. the accuracy of the PMI method and the proposed method on different term sets with the variety in the selection of paradigm words