Set-Similarity Joins
Based Semi-supervised Sentiment Analysis

Xishuang Dong, Qibo Zou, and Yi Guan

Web Intelligence Lab, Research Center of Language Technology,
School of Computer Science and Technology,
Harbin Institute of Technology, Harbin 150001, China
dongxishuang@gmail.com, zouqibo2009@163.com, guanyi@hit.edu.cn

Abstract. A set-similarity joins based semi-supervised approach is presented to mine Chinese sentiment words and sentences. The set-similarity joins is taken to join nodes in unconnected sub-graphs conducted by cutting the flow graph with Ford-Fulkerson algorithm into positive and negative sets to correct wrong polarities predicted by min-cut based semi-supervised methods. Experimental results in digital, entertainment, and finance domains demonstrate the effectiveness of our proposed approach.

Keywords: Set-similarity joins, Min-cut, Semi-supervised learning, Sentiment analysis.

1 Introduction

Min-cut based methods have shown effectiveness of mining sentiment words by sentiment classification [1,2]. The procedure to employ min-cut contains three are connected to T. If these sets are subject to $(V_S \cup V_T) = V$ and $(V_S \cap V_T) = \emptyset$, min-cut on G is unique, which means that nodes are classified into two categories S and T. However, if they are subject to $(V_S \cup V_T) \neq V$ and $(V_S \cap V_T) = \emptyset$, nodes in $V - (V_S \cup V_T)$ can form the set U, where it is subject to $(U \cap V_S) = \emptyset$ and $(U \cap V_T) = \emptyset$. Moreover, each node v' in U is classified randomly because it does not connect to S or T. Therefore, if v' is classified correctly, the performance of the min-cut based method can be improved.

We adopt the set-similarity joins (SSJs) to optimize min-cut based semi-supervised approaches. SSJs can be methods that unify sets by their similarities, that is, let Sim denote a similarity function. Given two values, $Similarity(U, V_S)$ and $Similarity(U, V_T)$, calculated by Sim, if $Similarity(U, V_S)$ is larger than $Similarity(U, V_T)$, then U is joined into V_S. Otherwise, U is joined into V_T. Therefore, nodes in U are not classified randomly but by set similarities. Experimental results on mining Chinese sentiment words and sentences show that F-measure values are improved prominently.

The remainder of this paper is organized as follows. In section two, the related work of sentiment analysis on words and sentences is introduced. Our semi-supervised approach is described in section three. Experimental results are presented and analyzed in section four. Finally, in section five, we draw our conclusions and outline the future work.

2 Related Work

The goal of sentiment analysis on words and sentences is to divide candidates into positive and negative sets. The approaches can be categorized into three types: unsupervised, supervised, and semi-supervised. For mining sentiment words, template based and char based methods, as examples of unsupervised methods, are presented to recognize sentiment adjectives, which computes collocation frequencies to recognize sentiment polarities in search engine and corpus [3]. As an example of supervised methods, Conditional Random Fields (CRFs) is taken to label sentiment polarities with features such as parts of speech, special chars, and sentence patterns [7]. For an example of semi-supervised methods, the min-cut based method is proposed to mine sentiment words on French and Hindi [1,8,9]. For mining sentiment sentences, as examples of unsupervised methods, positive, negative, and neutral membership functions are constructed to calculate sentiment degree of membership to recognize sentiment sentences [10]. Conjunctions are applied to analyze complex sentences [11]. As an example of supervised methods, k-Nearest Neighbors (kNN) is used to analyze tweets with features such as words, n grams, punctuations, and phrases [12]. Cross-domain sentiment analysis is solved through constructing domain lexicons [13]. For an example of semi-supervised methods, semi-supervised recursive auto-encoders is presented, where words in a sentence are combined as nodes to build a tree, and these nodes as features are used to predict polarities of sentences [14]. Semi-supervised methods have two advantages over unsupervised and supervised methods: (1) they are with higher precisions and recalls; (2) they can exploit unlabeled data to alleviate shortage of training corpus. As a typical semi-supervised method, the min-cut based method not only inherits these advantages but also uses relations between samples to improve performances further.

3 Method

We employ SSJs, which will be introduced briefly as follows, to optimize the min-cut based semi-supervised method.

3.1 Set-Similarity Joins (SSJs)

Considering two set collections as inputs, A and B, SSJs can merge all pairs (a, b), $a \in A$ and $b \in B$, if their similarity is larger than some predefined threshold t, into a union set. The similarity function can be edit distance, cosine similarity, or jaccard similarity [15,16]. In this paper, cosine similarity is taken as the similarity function because of comparatively high accuracy [17].

3.2 SSJs Based Semi-supervised Approach

The first step of our method is to construct a flow graph with candidates. A example is shown in Fig.1.

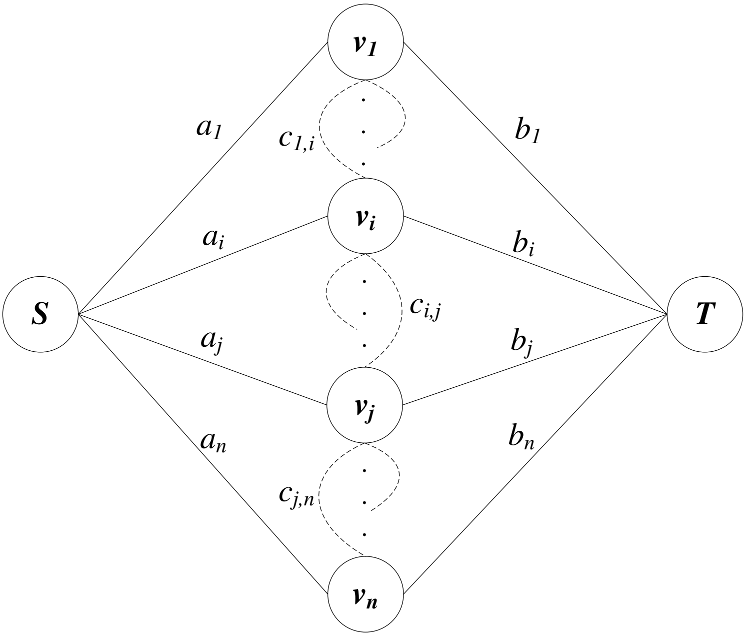

Fig. 1. A flow graph

where S and T are two end-points denoting positive and negative polarities respectively. a_i and b_i are possibilities predicted by sentiment classifiers, which is subject to $a_i + b_i = 1$ for $1 < i < n$, $i \neq j$, and $1 < j < n$. $c_{i,j}$ is the strength of semantic relations. v_i is the candidate.

Then, the graph may be cut into many sub-graphs by FFA as Fig.2.

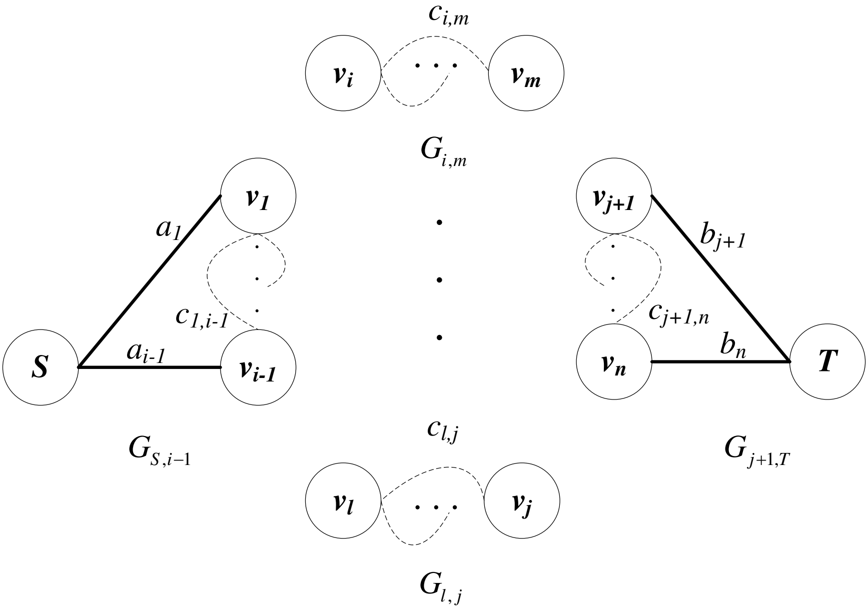

Fig. 2. Sub-graphs conducted by cutting the flow graph with FFA

where $1 < i < m < l < j < n$. There will be a unique min-cut for $i - 1 = j$, which means that all nodes are cut into two sets. Nodes in $G_{S,i-1}$ are classified as the positive polarity, and nodes in $G_{j+1,T}$ are classified as the negative polarity. Otherwise, there would be many min-cuts because nodes in $G_{i,m}$ to $G_{l,j}$ are classified randomly because these nodes are not connected to S or T. Node sets of graphs $G_{S,i-1}$ and $G_{j+1,T}$ can be denoted as sets $V_{S,i-1}$ and $V_{j+1,T}$ without considering edges in graphs $G_{S,i-1}$ and $G_{j+1,T}$. Therefore, sets $V_{i,m}$ to $V_{l,j}$ can be obtained from graphs $G_{i,m}$ to $G_{l,j}$ in this manner. To avoid random classification, which is the main deficiency of the current method, our method introduces the third step: nodes from sets $V_{i,m}$ to $V_{l,j}$ are joined into the set $V_{S,i-1}$ or $V_{j+1,T}$ with SSJs. For example, firstly, nodes in $V_{i,m}$, $V_{S,i-1}$, and $V_{j+1,T}$ are constructed as vectors $vt_{i,m}$, $vt_{S,i-1}$, and $vt_{j+1,T}$, where elements in these vectors are nodes in $V_{i,m}$, $V_{S,i-1}$, and $V_{j+1,T}$. Then, Cosine Similarity (CS) is taken to measure similarities $CS(vt_{i,m}, vt_{S,i-1})$ and $CS(vt_{i,m}, vt_{j+1,T})$. If $CS(vt_{i,m}, vt_{S,i-1})$ is larger than $CS(vt_{i,m}, vt_{j+1,T})$, then nodes in $V_{i,m}$ are joined into the set $V_{S,i-1}$. Otherwise, nodes in the set $V_{i,m}$ are joined into the set $V_{j+1,T}$. Therefore, nodes from sets $V_{i,m}$ to $V_{l,j}$ are joined into the set $V_{S,i-1}$ or $V_{j+1,T}$ according to similarities. To improve performances further, we adopt self-training to extend the training set by selecting samples of high confidences with a voting method. That is, for a unlabeled sample, its polarity is predicted by both the sentiment classifier and our method. If these two polarities are the same, the sample is added into the training set. When the number of selected samples is over a predefined threshold, the enlarged training set is used to retrain the model.

3.3 Sentiment Analysis on Words and Sentences

We adopt the SSJs based semi-supervised approach to mine sentiment words and sentences. First step is to extract candidate words and sentences and to construct flow graphs. For mining sentiment words, nodes are words, and candidate words are adjectives and adverbs. Semantic relations between words are extracted by the similarity computation module in HowNet. Maximum Entropy (ME) based sentiment word classifier is used to predict the primary polarities and to construct polarity relations between candidates and end-points. Then, we construct a flow graph of words with these candidate words and relations. For mining sentiment sentences, nodes are sentences, and candidate sentences are sentences containing sentiment words. Semantic relations between sentences are built by calculating cosine similarities between sentences. ME based sentiment sentence classifier is used to predict the primary polarities and to construct polarity relations. Then, we construct a flow graph of sentences with these candidate sentences and relations. Second step is to classify candidates into positive and negative polarities with our proposed methods introduced in section 3.2. Finally, samples with high confidences are selected to improve performances.

4 Experiment and Analysis

4.1 Data and Evaluation

Data sets for training sentiment classifiers are collected from Sogou, Myspace, and Alibaba in which each item is consist of a sentiment word or sentence with its corresponding sentiment polarities. Sentiment word features are synonyms while sentiment sentence features are sentiment words, sentence patterns, and sequences of words [2]. Testing sets, which are collected from digital, entertainment, and finance domains, are provided by task one and task two in Chinese Opinion Analysis Evaluation 2011 (COAE 2011)[1]. Its data distribution is shown in Table 1.

We employ ME models as baselines. Min-cut based semi-supervised methods, which are called McME because they use ME models to recognize primary polarities, are taken to demonstrate effectiveness of improvement. In addition, we select two sentiment analysis models Mc [2] and MKIMCV [18] in COAE 2011 to compare with our method to demonstrate effectiveness on mining sentiment

[1] http://ir-china.org.cn/coae2011.html

words and sentences because they are the state-of-art of recognizing Chinese sentiment words and sentences in COAE 2011 [19]. F-measure is used to measure the performance.

Table 1. Data distribution in three domains

Domains	Positive Words	Negative Words	Positive Sentences	Negative Sentences
Digital	2134	893	4920	755
Entertainment	1271	729	961	268
Finance	471	461	308	209

4.2 Results and Analysis

The performance comparison of different models for recognizing sentiment words and sentences is shown in Fig.3 and Fig.4 respectively. From Fig.3., we can gain the following results: (1) our method outperforms ME model significantly on mining sentiment words. F-measure values of our method are about 13 to 16 percents higher than those of ME models; (2) compared with McME, F-measure values of our method are improved by about 7 to 9 percents in three domains. On average, F-measure values are improved by about 8 percents.

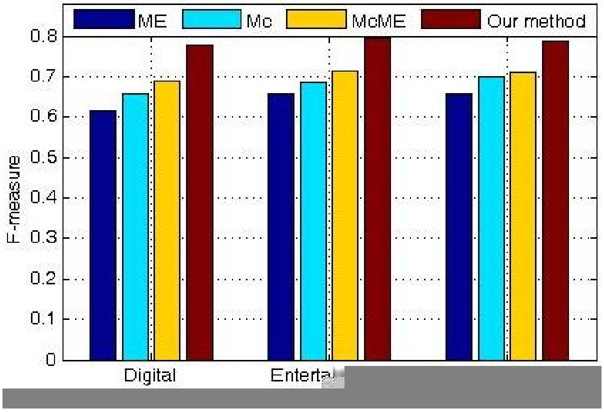

Fig. 3. Performance of mining sentiment words

From Fig.4., we can gain similar results: (1) our method outperforms ME model on mining sentiment sentences. F-measure values of our method are improved by 22 to 29 percents in three domains; (2) compared with McME, F-measure values of our method are improved by about 8 to 17 percents; (3) compared with MKIMCV, F-measure values of our method are improved by about 3 to 5 percents. On average, F-measure values are improved by about 4 percents. From the results, we can believe that our method can recognize Chinese sentiment words and sentences more efficiently.

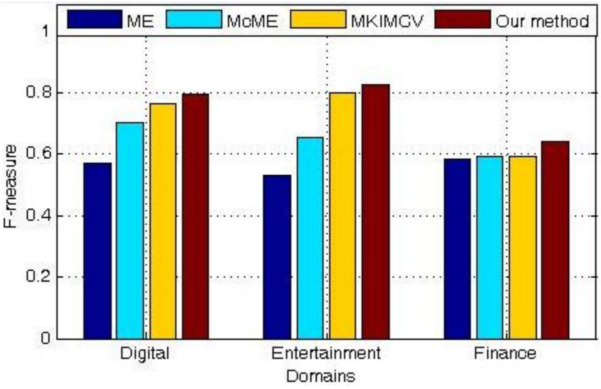

Fig. 4. Performance of mining sentiment sentences

5 Conclusion and Future Work

We present a SSJs based semi-supervised approach that fully utilizes relations between samples to mine Chinese sentiment words and sentences. Experimental results in digital, entertainment, and finance domains show that F-measure values are improved by about 8 percents and 4 percents respectively, which means that our method can be taken to mine Chinese sentiment words and sentences more effectively. Moreover, we believe that our approach is adequate for solving binary classification such as sentiment classification when relations between samples can be gained. Future work includes: (1) to construct the optimal SSJs to calculate similarities of sets of words and sentences; (2) to implement sentiment analysis on texts with our proposed approach.

Acknowledgement. This work was supported by National Natural Science Foundation of China (Project No. 60975077) and National Natural Science Foundation of China (Project No. 90924015).

References

1. Rao, D., Ravichandran, D.: Semi-supervised Polarity Lexicon Induction. In: The 12th Conference of the European Chapter of the Association For Computational Linguistics, pp. 675–682 (2009)
2. Dong, X., Zou, Q., Guan, Y., Gao, X., Yan, M.: Positive And Negative Polarity Analysis on Chinese Words And Sentences Based on Maximum Entropy Model And Min-Cut Model. In: The 3rd Chinese Opinion Analysis Evaluation, pp. 97–105 (2011)

3. Wu, Y., Wen, M.: Disambiguating Dynamic Sentiment Ambiguous Adjectives. In: The 23rd International Conference on Computational Linguistics, pp. 1191–1199 (2010)
4. He, H., Li, S., Xiao, F., Xu, W., Guo, J.: PRIS Sentiment Analysis System Techical Report. In: The 1st Chinese Opinion Analysis Evaluation, pp. 46–55 (2008)
5. Ford, L., Fulkerson, D.: Maximal Flow Through a Network. Canadian Journal of Mathematics 8, 399–404 (1954)
6. Cormen, T., Leiserson, C., Rivest, R., Stein, C.: Introduction to Algorithms. The MIT Press and McGraw-Hill Book Company (2001)
7. Das, D., Bandyopadhyay, S.: Word to Sentence Level Emotion Tagging For Bengali Blogs. In: The 47th Annual Meeting of The Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, pp. 149–152 (2009)
8. Blum, A., Lafferty, J., Rwebangira, M., Reddy, R.: Semi-supervised Learning Using Randomized Mincuts. In: The 21st International Conference on Machine Learning, pp. 97–104 (2004)
9. Zhu, X., Ghahramani, Z.: Learning From Labeled And Unlabeled Data With Label Propagation. Technical report CMU-CALD-02-107. Carnegie Mellon University (2002)
10. Fu, G., Wang, X.: Chinese Sentence-level Sentiment Classification Based on Fuzzy Sets. In: The 23rd International Conference on Computational Linguistics, pp. 312–319 (2010)
11. Meena, A., Prabhakar, T.V.: Sentence Level Sentiment Analysis in the Presence of Conjuncts Using Linguistic Analysis. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECiR 2007. LNCS, vol. 4425, pp. 573–580. Springer, Heidelberg (2007)
12. Davidov, D., Tsur, O., Rappoport, A.: Enhanced Sentiment Learning Using Twitter Hashtags And Smileys. In: The 23rd International Conference on Computational Linguistics, pp. 241–249 (2010)
13. Guo, H., Zhu, H., Guo, Z., Su, Z.: Domain Customization For Aspect-oriented Opinion Analysis With Multi-level Latent Sentiment Clues. In: The 20th ACM International Conference on Information And Knowledge Management, pp. 2493–2496 (2011)
14. Socher, R., Pennington, J., Huang, E., Ng, A., Manning, C.: Semi-supervised Recursive Autoencoders For Predicting Sentiment Distributions. In: The 16th Conference on Empirical Methods in Natural Language Processing, pp. 151–161 (2011)
15. Chaudhuri, S., Ganjam, K., Ganti, V., Motwani, R.: Robust And Efficient Fuzzy Match For Online Data Cleaning. In: The 22nd ACM SIGMOD International Conference on Management of Data, pp. 313–324 (2003)
16. Arasu, A., Ganti, V., Kaushik, R.: Efficient Exact Set-similarity Joins. In: The 32nd International Conference on Very Large Data Bases, pp. 918–929 (2006)
17. Chandel, A., Hassanzadeh, O., Koudas, N., Sadoghi, M., Srivastava, D.: Benchmarking Declarative Approximate Selection Predicates. In: The 26th ACM SIGMOD International Conference on Management of Data, pp. 353–364 (2007)
18. Xu, R., Wang, Y., Xu, J., Zhang, Y., Zheng, H., Gui, L., Ye, L.: Chinese Opinion Analysis Based on Multi Knowledge Integration And Multi Classifier Voting. In: The 3rd Chinese Opinion Analysis Evaluation, pp. 77–87 (2011)
19. Xu, H., Sun, L., Yao, T., Liao, X.: Overview of The Third Chinese Opinion Analysis Evaluation. In: The 3rd Chinese Opinion Analysis Evaluation, pp. 1–24 (2011)