A Public Opinions Monitoring System based on Similarity of Microblogs

Jiajia Miao[1,2]
1) Institute of Command Automation, PLA University of Science and Technology, Nanjing, China
2) Key Laboratory of C4ISR Technology, National University of Defense Technology, Changsha, China
jjmiao@ieee.org

Xuelin Fang
613#, School of Computer, National University of Defense Technology, Changsha, China
fang4210@gmail.com

Guoyou Chen
Institute of Command Automation, PLA University of Science and Technology, Nanjing, China
goyalchen@163.com

Handong Mao
Key Laboratory of C4ISR Technology, National University of Defense Technology, Changsha, China
handmao2005@126.com

Le Wang
Key Laboratory of C4ISR Technology, National University of Defense Technology, Changsha, China
cc_alan@163.com

***Abstract*—Public opinion is the development and changes in certain social space around the intermediary social events, public social managers and held social and political attitudes. In order to improve the efficiency of text clustering, this paper proposes a new similarity measure model OFM. In the first stage, OFM proposed a new term relationship search algorithm named CrtURG non-directional graphics integrated relationship. OFM build another called SIM expand ad hoc, said vector algorithm to different instant messaging. Keywords expand overlap probability vector increased and instant messages containing specific semantic information. The experimental results show that, OFM, is superior to the combined effect of surface matching methods and corpus-based approaches.**

Keywords-component; formatting; style; styling; insert.

I. INTRODUCTION

Public opinion is the development and changes in certain social space around the intermediary social events, public social managers and held social and political attitudes. It is more the sum of the masses on the performance of the social phenomenon, the question of the belief, attitudes, opinions and emotions. Network public opinion is an important form of public opinion, it is with the development of the Internet, communications networks, and broadcast networks, human communication network and the rapid formation and development, and is a huge influence on social life. Especially after 9.11, the countries in the world have accelerated to carry out work on the national security and social security, increase of endangering national security and social stability information monitoring, and to strengthen the network public opinion research. Network public opinion as social Vientiane mapping, the most direct, fastest reflect the social situation and development trend of public opinion. Quick access to the network public opinion effectively analyze and keep track of the timely warning, effective regulation, help judgment and boot of factors and colonize problems in a timely manner, to help effective control of the social public opinion.

With the growing popularity of Web 2.0 technologies and related Internet applications, microblogging and other new media has become an important opinion field of network public opinion. Microblogging has a large user base, spread fast, and convenient to set out on the information, the past two years, the formation of the explosive growth in our country, and the outbreak of the main source of public opinion and the media has become. Illegal and unhealthy information on the microblogging are endless, especially various rumors spread, causing a lot of harm to society.

Opinion leaders have a powerful voice. Opinion leaders to play a major role in emergencies generation, fermentation, spread links Internet users in specific areas, they have a powerful voice, subconsciously affect tens of thousands of onlookers.

Public opinion is the aggregate of individual attitudes or beliefs held by the adult population. One of the principal approaches to study public opinion is to dig into the communication media that disseminate the opinions . The contents of instant messages, which are widely popularized among people (e.g., E-mail, MSN, ICQ, Mobile Message, etc.) , are mostly short text segments and deliver the ideas of people. Employing text clustering on the contents, a kind of short text segments, could find the current hot topics or public opinions in society or some organization [1,2].

The POMS systems including of microblogging data search and acquisition restore encrypted data, Intelligent Data Analysis of microblogging, microblogging incidents show, warning, guidance, and the massive microblogging data storage and management. Text clustering can be adopted to detect the public opinions in instant messages, which offer the greatest potential for social applications.

This paper is supported by the Natural Science Foundation of Jiangsu Province BK2010131, and Foundation of PLA University of Science and Technology 20110208.

All existing clustering technologies are built based on four concepts: data representation model, similarity measure, cluster model and clustering algorithm. And the clusters are built by using the data model and the similarity measure [3].

II. RELATED WORKS

We review existing methods to measure the short text of the similarity between the lots of our work. Surface matching methods, these methods are divided into three categories: a corpus-based method and ontology-based method.

Given a text segment (d1, d2), and their words D1 and D2, respectively, the surface of the matching method is defined as formula (1) [4]. These technologies mainly rely on shared, or a mix of words. When a larger text unit for comparison, the overlap may be sufficient detection similar, but the text when the unit is small, simple words and a phrase of the surface matching is difficult to succeed, because the number of matches is small. The similarity measure of the quality of the short text segments is often unreliable. Therefore, there are also some expansion of said text fragment has been proposed [5, 6].

Graphics-based approach to explore the classification of knowledge and / or information construction in the training corpus word similarity diagram [7]. The degree of similarity between the documents, according to the diagram deduced. Maguitman produced a graph-based method to measure the similarity between the webs [8]. They hired the original ODP ontology, and explore its hierarchical and non-hierarchical (such as graphics processing) relationship to define the relationship between the words of the score. This method to take full advantage of the ODP Web Links (marked as "see also"). However, there is no hyperlink in this instant messaging. Therefore, this method cannot accommodate the instant message.

Rada Mihalcea some experimental corpus-based, knowledge-based measures of text similarity [9]. He found that the corpus-based approach is clearly better performance than a knowledge-based approach to deal with a short paragraph of text; the hybrid approach has the best results. In this paper, we propose a hybrid approach short text part and use it to look for instant messaging, text clustering public opinion.

III. SIMILARITY MEASURE MODEL

Relationship graphs and two algorithms word similarity measure model, known as the face of public opinion similarity measurement model (OFM). In this section, we first introduce the algorithm CrtURG, which is then used to build vocabulary classification, HowNet diagram based on the training corpus. Then, we came to an algorithm, called a SIM card extend traditional TF-IDF said to combine more contextual information. Finally, our method of instant messages between any two points calculated using the cosine measures, based on the similarity of the SIM score.

A. Define the Words' Relationship Degree

- The Word Relationship Graph

URG is a triple of (W, E, V), wherein W is a word set, E denotes a nondirectional edge set on W and V is a set of weights for edges in E. Given the relationship degree function f, if and only if f(w1,w2)>α, there is an edge e$\in$E whose weight is defined as f(w1,w2).

α is a threshold that controls the edge between two words in W. And the edge weight is defined by function f. A demonstration of URG structure is showed in Figure 2.

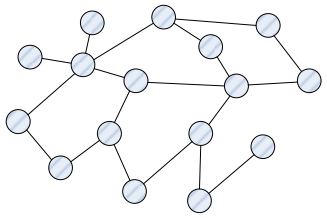

Figure 1. The structure of URG.

URG is a nondirectional graph. Only the words which are obviously worth focusing about some public opinions are included as vertices in this graph. The edge between two words in URG defines the relationship degree of these two words. Only the edges whose relationship degrees are large enough (f(w1,w2)>α) are contained in URG.

- CrtURG Algorithm

To construct URG, a critical problem is how to define the function f. As we know, the ontology-based method faces the problem caused by the new coming words and the corpus-based method faces the challenge of dealing with short-length instant message. We define the function f by formula (2), which combines the knowledge of taxonomy, HowNet, and the information in one training corpus.

$$f(w_1, w_2) = \begin{cases} D(w_1, w_2), & w_1, w_2 \in HowNet \\ L(w_1, w_2), & w_1 \text{ or } w_2 \notin HowNet \end{cases} \quad (2)$$

D(w1,w2) denotes the similarity score between w1 and w2 based on HowNet. If both w1 and w2 are included in HowNet, we compute D(w1,w2) according to the method in [21]. Some more sophisticated forms of similarity measurement based on HowNet can be leveraged here, which will be studied in our future work.

L(w1,w2) in formula (2) is the corpus-based part of total similarity score. It is calculated by formula (3). p is the probability function of words in training corpus. Formula (3) produces the relationship degree in terms of corpus according to word co-occurrence.

$$L(w_1, w_2) = \begin{cases} log_2\left[p(w_1, w_2) \Big/ p(w_1) \cdot p(w_2 \cdot) \right] \\ 0, \quad if \ p(w_1, w_2) = 0 \end{cases} \quad (3)$$

URG word of three types: A, B, and C, A refers to the case (with labels), an impressive training corpus. B shows, this is clearly the classification with the A's (tab and B). (Label C) is not included in the corpus, but really it comes to some public opinion, can also be added manually to the URG. The lighting

formula (2), these words can be obtained from the taxonomic weight, although they did not get the contribution from the corpus.

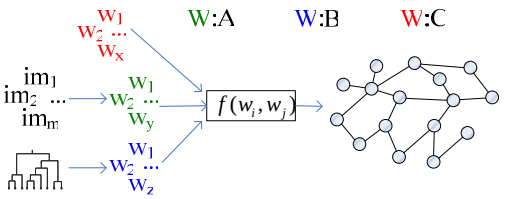

Figure 2. Composition of URG.

The algorithm created CrtURG (URG) shows a diagram of how to build the word. Beforehand, URG can build. Only training of new data or new Focus key words should be included in the URG case, it needs to be updated. How to update URG is an open question. This topic is beyond the scope of this article.

Then, the two words in the relationship between the degree of the URG, can be represented by the formula (4) are as defined. vi is the weight of the connecting side of w1 and w2 in the shortest path, k is the length of the path. In fact, some of the more complex the distance measurement can take advantage of the study, which will work in the future. Can be effectively used for the shortest path algorithm [10].

$$r(w_1, w_2) = \frac{1}{k} \cdot \sum_{i=1}^{k} v_i \qquad (4)$$

B. *Measure Similarity of Instant Messages*

- Expanded Ad Hoc Representation

In this section, we derive the degree of association between two specific words URG, and calculate their similarity score, the proposed extension of the project on behalf of instant messaging.

In the conventional TF-IDF architecture, the similarity score between the text segment, that is, they expressed the inner product of vectors is effective because the short length of the instant message from Figure 3 (a) is lit.

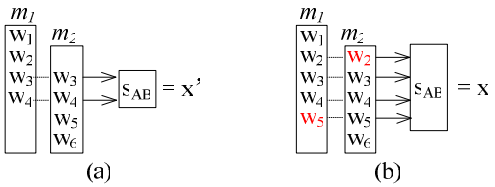

Figure 3. Low efficiency of traditional TF-IDF schema and its expansion.

When measuring two instant messaging, for example, the similarity between the M1 and M2, expand its representation vector difference does not appear in the text, but the word that appears URG significantly related. Furthermore, the additional words in a message must appear in another. Figure 4 (b), this can be irradiated. For example, W5 is not in the money supply, M1, and W4 are included in M2. Therefore, we will W5 is m1 calculate the similarity between the M1 and M2 money supply. The weight of the additional words, as in Figure 4 (b), M1 W5, is determined by the following formula (5). Enlarged only when measuring the similarity between them instant messages, they should be an ad hoc epitaxial. Their representatives may also be different for different pairs of the instant message, because the added words and corresponding weight can be changed.

The new value of the word in the expanded representation of the carrier in an instant message can be defined by the formula (5). If in the Figure 4 (b) in the example in the figure, the word W5 M1 calculated value increases according W4 value m1 and their degree of relationship between the r (W4, W5).

- SiM Algorithm

The expanded representation vector of instant message combines the knowledge of lexical taxonomy and interested information in training corpus which is provided by the users. Similarity is calculated based on the expanded representation using surface matching method, e.g., cosine. We use the formula (6) to define the similarity score.

$$sim(m_1, m_2) = (m_1 \cdot m_2) / (\|m_1\| \times \|m_2\|) \qquad (6)$$

The algorithm SiM firstly constructs the expanded ad hoc representations for two instant messages when comparing their similarity, and then returns the final similarity score.

IV. EXPERIMENTAL EVALUATIONS

Because of the similarity measure using instant messaging clustering method discussed in this article, we assess the our model cluster case, using the traditional k-means algorithm [11], the instrument. Traditional TF-IDF methods and corpus-based method was introduced as a baseline. We use the vocabulary Classification the HowNet System 2000 Edition. The HP unit 4 the Itanium II1.6G processor and 48 GB of memory used by the hardware platform.

Holding 1027 words, 889 words, they are included in Text Text 138, and other word does not exist. First, we calculate the similarity score of 889 words Text, and then tap the relationship degree between the other 138 word as well as from the two different groups, by the extent of co-occurrence of words between.

Expanding Ad Hoc represented compared to the instant message is different, the different messages. After the expansion, which is difficult to determine the exact length of the instant message. The news said that in order to evaluate the effect of the expansion, we have chosen a representative of the news, and collect the length of the expansion of its statistical significance. Figure 8 is a diagram showing different representations of the length of this message, when measured similarities with other different message. We can see from Figure 4, a large extent of the changes in the length of this message, that is, some relevant terms are added to the representation of the message. Therefore, this message and other messages between shared words of more than the original situation.

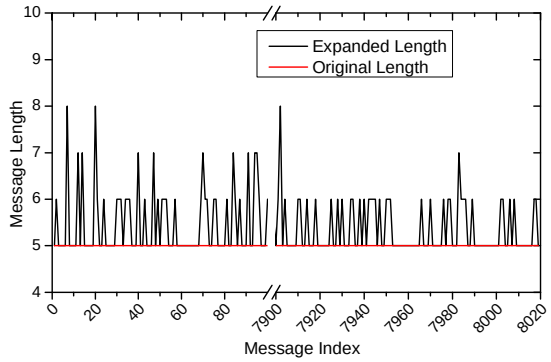

Figure 4. Message Length after Expansion.

Because there is no standard data set to assess the clustering results of our human study. If an instant messaging with others in the same cluster is consistent, then we believe that this result is correct, otherwise false. Figure 5 (a) shows the 15 iterations for different k-means clustering performance and (b) the production of 15 clusters for different iterations.

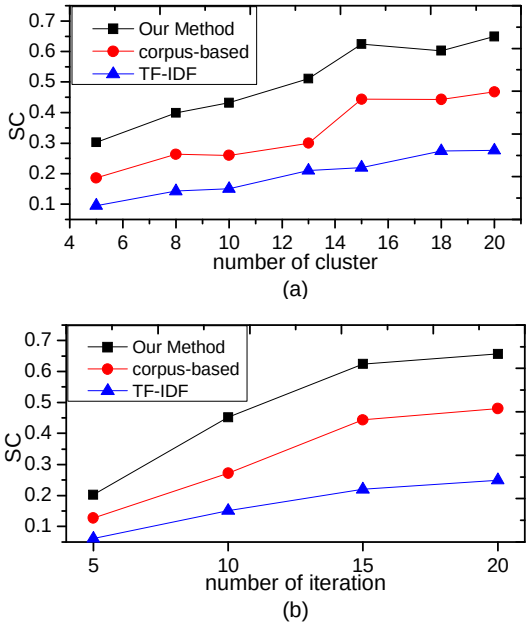

Figure 5. The SC of clustering results

SC said divisible cluster. From Figure 5, we can find that our model can be relatively clear partition corpus, which is illuminated by SC in FIG 5 (a) (SC = 0.65, K = 20). This is a reasonable result. We can find instant messaging with a poor performance, due to lack of sufficient context, the traditional TF-IDF method. In addition, TF-IDF method cannot benefit from more iteration, because the line is relatively flat, as in Figure 5 (b) below. Corpus-based approach has a better performance than the TF-IDF.

V. CONCLUSION AND FUTRUE WORKS

In this paper, we propose a new mixed-mode similarity measure instant messaging clustering found that the views of the public. The model combines semantic knowledge corpus callosum the vocabulary classification and statistical information, and integrated graphics structural relationship information. By logical extension, in the drawings, the relationship is added to the instant message of the carrier. After that, the message can be clustered based on such an expansion, said. Experimental results show that our model, SC, about 10%, and nearly 40% SC than corpus-based approach than the traditional TF-IDF method of instant messaging in the cluster effect.

REFERENCES

[1] Yongheng Wang, Yan Jia, and Shuqiang Yang. Ontology-based Short Conversations Clustering in Very Large Text Database. In The 2nd VLDB Workshop on Ontologies-based techniques for DataBases and Information Systems (ODBIS'06)(Seoul, Korea, September 11, 2006). Springer Press, Berlin, LNCS 4256, 2006, 83-93.

[2] Le Wang, Yan Jia, and Weihong Han. Instant Message Clustering Based on Extended Vector Space Model. In Proceeding of The 2nd International Symposium on Intelligence Computation and Applications (ISICA'07)(Wuhan, China, Sept. 21-23, 2007). Springer Press, Berlin, LNCS 4683, 2007, 435-443.

[3] Khaled M. Hammouda, and Mohamed S. Kamel. Document Similarity Using a Phrase Indexing Graph Model. Springer Knowledge and Information Systems, 6, 6 (Dec, 2004), 710-727.

[4] Christopher D. Manning, and Hinrich Schuetze. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA, 1999.

[5] Sahami, M., and Heilman, T. A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. In Proceedings of the 15th International Conference on World Wide Web (WWW'06) (Edinburgh, Scotland, May 23 - 26, 2006). ACM Press, New York, NY, 2006, 377-386.

[6] Metzler, D., Dumais, S., and Meek, C. Similarity measures for short segments of text. In Proceeding of 29th European Conference on Information Retrieval (ECIR'07)(Rome, April 2-5, 2007). Springer Press, Berlin, LNCS 4425, 2007, 16–27.

[7] Xiaojun Wan and Yuxin Peng. A Measure Based on Optimal Matching in Graph Theory for Document Similarity. In Proceeding of Asia Information Retrieval Symposium (AIRS'04)(Beijing, China, October 14-16, 2004). Springer Press, Berlin, LNCS 3411, 227–238

[8] Ana G. Maguitman, Filippo Menczer. Algorithmic detection of semantic similarity. In Proceedings of the 14th international conference on World Wide Web (WWW'05) (Chiba, Japan May 10 - 14, 2005). ACM Press, New York, NY, 2005, 107–116.

[9] Mihalcea R, Corley C, and Strapparava C. Corpus-Based and knowledge-based measures of text semantic similarity. In Proceeding of the 21st American Association for Artificial Intelligence (AAAI'06)(Boston, USA, July 16–20, 2006). AAAI Press, Menlo Park, California, 2006.

[10] T. H. Cormen, et al. Introduction to Algorithms (Second Edition). The MIT Press, Cambridge, MA, 2001..

[11] J. MacQueen. Some methods for classification and analysis of multivariate observations. In Proceeding of the 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, 1967, 281-297.