A Clustering Algorithm towards Microblogs based on Vector Space Model

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

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

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

Siyu Jiang
The post-doctoral workstation of the 73111 army, Xiamen, China
jsy7528@163.com

***Abstract*—Weibos have become wildly popular in China in recent years, and state media reports that there are more than 300 million registered users. The "Real Name" Policy[1] requires all users on Chinese weibo websites to register with the name that corresponds with their government issued ID card. With the rapid development of the web, the research of consensus encounters new problems and challenges. Is a practical method for large-scale text clustering, instant messaging, text content analysis features, and find or track the social hot topics. Unlike the file, which is not suitable for very common clustering algorithm? A new method is proposed of the named MVSM synthesis microblogging dialogue, but also enriched the words of the vector is not included in the text of the blog, but existing content is closely related. Extended vector space this MVSM perform the dialogue, k-means clustering. Experiments on public datasets show better, MVSM than traditional k-means and k-means algorithm into two.**

Keywords- microblogs clustering, k-means, vector space model

I. INTRODUCTION

With the growing popularity of Web2.0 technology and Internet applications, microblogging and other new media network public opinion has become an important public opinion field. Microblogging has a large user base, spread fast, and convenient to set out on the information, the outbreak of the main source of public opinion and the media has become. How microblogging has massive amounts of unstructured text data, the large number of users and the real-time characteristics of strong, research microblogging public opinion supervision software platform, has become a priority.

The micro-blog (Micro-blogging) is a technology that allows users to update brief text (usually about 140 words), and can be in the form of publicly blog. It allows anyone to read, or can only be read by the user group. Its core characteristics can be linked via SMS, instant messaging tools, e-mail, MP3, web pages. Some micro blog can also publish multimedia, such as pictures or video clips. Microblogging, everyone is a source of information, you can accept someone else's information, there are others to accept their own information; possible to obtain the information forwarded to their fans, thus forming a core of interpersonal rapid spread of network.

Seen our new Internet media monitoring filter wrong, harmful and wrong, even reactionary information to combat cybercrime effectively curb the breeding and spread of harmful information, is the most urgent task. However, due to the size of the one hundred million users of the microblogging platform daily Bowen ten thousand, of public opinion, minutes, seconds, and unstructured text messages Bowen information propagation speed. Therefore rely on artificial passive monitoring can not comply with the requirements of public opinion Media News Forum and other monitoring public opinion, and often do not have the pressure of work under the supervision of public opinion microblogging new specialized public opinion against the microblogging monitoring software .

The most common word processing method is a file with the vector represents. This is the so-called vector space model, in the words of one vector corresponds to a document, and corresponds to the file size. The derivation of high-dimensional vector, leaving the main challenges of document clustering is how to deal with these high-dimensional data. However, microblogging is a very short joint document. Usually only a few key words relevant keywords in a microblogging, microblogging theme, and sometimes lurking. Sparse keyword, the method of word frequency is not the appropriate measure of the similarity between the micro-Bo. Table 1 is an example to illustrate the above problem; the micro-Bo is about sports and another considerable similarity. MB-1 MB -2 MB -3 0.71 degrees and 0.58 degrees, not any similarity between the MB -2 and MB-3, which conflicts with reality. Therefore, based on the measures of the word bag model and the long-term frequency the microblogging mining NA.

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

This paper presents two methods to improve the description of the reaction of the microblogging, sparse the keyword clustering microblogging. First, we note that microblogging is a semi-structured data, timestamp, source and destination addresses. Sent back and forth between the micro-Bo particular person, in a certain time interval, to form a conversation, this group of micro-Bo to a specific topic. Therefore, we as a microblogging dialogue. Obviously microblogging more key projects, a more complete context information than the simple single-message. And clustering to the dialogue, rather than micro-Bo.

TABLE I. EXAMPLE TO ILLUSTRATE BAG-OF-WORD MODEL

	ball	basketball	football	foot	IM1	IM2	IM3
IM-1	0	1	1	0	-	0.71	0.58
IM-2	0	2	0	0		-	0
IM-3	1	0	2	1			-

Second, we have enhanced the content described in the words of the microblogging, which is not in microblogging existing in the microblogging, but there is a close relationship. Fox example, the word "ball" and "football" added to the IM-2, will not appear in the IM-2, but there is significant correlation between the word "basketball" in IM-2.

In this paper, we propose the the microblogging clustering called MVSM the method, it can automatically scan instant messaging corpus, to build dialogue and strengthen the traditional TF-IDF model, then by increasing in conversation. Dialogue MVSM clustering model, such as k-means development [2]. MVSM with two other well-known text clustering methods is based on the traditional TF-IDF method for calculations and comparisons. HowNet knowledge base is used to quantify the intensity of the relationship between the microblogging words during the experiment. Experimental evidence shows that MVSM significantly outperform the market. , HowNet bilingual linguistics, so MVSM smoothly converted to handle China [3].

II. RELATED WORKS

Various methods can be used for different types of micro-Bo recognized boundary conversation. Microblogging can easily capture the thread posted on Usenet thread because of its inherent characteristics [4]. Faisal M. Khan mode [5], in order to determine the chat thread starts to flow in a chat room. These modes by a few sentences, such as "Hi, Hi" or "How are you", which is observed by the experts through chat conversations. The medium has a very strong interaction, such as chat rooms, it is an effective method. Marty A., the the Hearst advantage TextTiling algorithm to locate the subject within the boundaries described in [6]. The purpose of the algorithm is the starting point of the text separated into segments, and lexical analysis using the TF-IDF model based on the identification of themes. The above-mentioned method, the basis of the main content of corpus.

The relationship between the words has been one of the extensive research areas of natural language processing, text mining and information retrieval methods of latent semantic indexing (LSI) [7], by the singular vector decomposition, it can automatically discover the relationship between potential corpus. However, this method is very time-consuming when applied to a large corpus. Kenneth Ward Church the Association than mutual information estimated Lenovo word specification, on the basis of their co-occurrence probability concepts [8]. This criteria is not appropriate, then the symbiotic because the sparse instant message. Satoru Ikehara semantic attributes, it uses semantic properties of the system [9] on the basis of the vector space model. The purpose of this method, in order to reduce the use - a lower semantic attributes between the words, and to achieve a good efficiency, the dimension of the vector in the processing in Japanese.

III. MVSM METHOD

A. *Synthesizing Conversation*

Mail database (MDB) is a set of information, the message is stored in the form of an easy access to a set of messages. T1 and T2 are used to indicate the start and end time of a specific period, respectively. SI and DI source and destination addresses. CI is the text content of the instant message.

Further observed from the the reality mining projects MIT microblogging data set allows a dialogue before and after the instant message frequency is usually lower than that during a session, which is shown in Figure 1, we found that. This can be explained as follows. From the view of point of time, the people tend to intensive communicate with each other and on the same theme. In other words, this is one pair is about the same topics microblogging together approximate communication SCE generation time.

We define the following rules for synchronizing instant messages into conversations on the basis of above analysis. $V_{i,i+1}$ is used to denotes the time interval between two adjoint instant messages, m_i and m_{i+1}. We assume that if $V_{i,i+1} < \alpha$ and $V_{i,i+1} < V_{i+1,i+2}$, then m_{i+1} and m_i belongs to the same conversation; otherwise, m_{i+1} is the starting of next conversation. Where α is a statistic constant which describes the biggest interval between two adjoint instant messages that belong to the same conversation. MVSM orderly compares the intervals of adjoint IMs between two specific persons and synthesizes conversations for each pair of all persons.

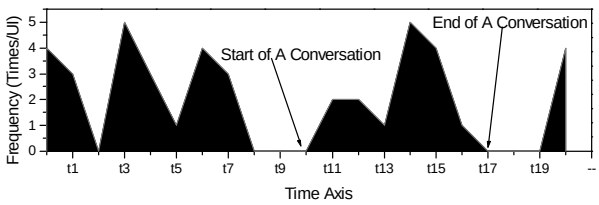

Figure 1. Frequency Change between two persons

B. *Enhancing the Representation of Conversation through Relevant Words*

Assume that there are *m* conversations, which consist of *n* different words totally. We calculate the relevant strength of

each pair of words according to HowNet. Given a conversation C_l, the word, t_j, which is not in conversation C_l, is used to enhance the vector representation of C_l if the relevant strength ($\delta_{i,j}$) between t_j and t_i, which is originally in C_l, is beyond one threshold of relevant strength.

$$\beta_i = \frac{E_i}{\sum_{k=1}^{num(C_l)} E_k}, (0 < i \le num(C_l)) \quad (1)$$

Then the value in vector delegating word t_j, which is added to enhance the vector representation of the conversation, can be determined according to formula (2) where t_k is the value of k-th word in the vector of C_l according to TF-IDF model.

$$t_j = \sum_{k=1}^{k=num(c_l)} \beta_k \cdot t_k \cdot \delta_{k,j} \quad (2)$$

For example, in table 2, a word-by-conversation matrix is constructed from 3 conversations (C1, C2, and C3) and 7 words. Only relevant strengths that are beyond 0.4 are considered and set $\delta_{i,j}$ equal to $\delta_{j,i}$. For T1 in C1, the value in TF-IDF is $2*\log_{10}(3/1)=0.9542$. The word, T4, which is not in C1, has a relevant strength beyond 0.4 with T1. So T4 should be added into the vector of C1. The value of T4 in vector is $0.9542*0.5283*0.62=0.3126$, where $-1/3*\log_2(1/3)=0.5283$ is the weight of T1.

MVSM as a microblogging clustering method, which is a variant of the standard k-means algorithm. The algorithm pretreatment instant message conversations and extendes carrier conversation before enhanced TF-IDF model clustering. MVSM measuring the similarity between the dialogue based on the cosine measures. And bound by the terms in a limited, pre-processing involves the amount of instant message set. In Therefore, MVSM is a scalable way.

TABLE II. EXAMPLE OF EXTENDING WORD-BY-CONVERSATION

	Original word Frequency			word relevant strengths						
	C1	C2	C3	T1	T2	T3	T4	T5	T6	T7
T1	2	0	0	N/A	-	-	0.62	-	-	-
T2	1	1	0		N/A	-	-	0.43	-	-
T3	0	1	1			N/A	-	-	-	0.42
T4	0	3	2				N/A	-	-	-
T5	0	0	4					N/A	-	-
T6	0	2	0						N/A	-
T7	2	0	0							N/A

IV. EXPERIMENTAL EVALUATIONS

Three different algorithms, MVSM, k-means, bisecting k-means and standard implementation and comparison. All these experiments manually predefined classification in two public data sets.

A. Evaluation Criteria

Both cluster validation methods, outline coefficient (SC) [11] and normalized mutual information (NMI) [12] are used to assess, because both of them are independent digital trunking, K clustering performance.

Contour coefficient (SC). Its value is typically between 0 and 1. Value of more than 0.5 indicates that the clustering result is divisible clear. If it is less than 0.25, it becomes very difficult to find practical significant cluster.

Normalized mutual information (NMI), the NMI measures the clustering results between the data sets and the original classification level, which is usually provided by a human expert. The greater value NMI is the ideal human; it illuminates a perfect clustering method.

B. Experimental Setting

The two sets of data: (1) in the Reuters -21,578 corpus, (2) 20 newsgroups data. Two data sets, including a priori classification file, as well as their field is extensive enough to be a real conversation. Preprocessing the original data set, the above-mentioned toolkit used bow and Porter function [13].

HowNet System 2000 Edition, the free version of this software is used to quantify mutual terms. The HP the unit 4 the Itanium II1.6G processor and 48 GB of memory as the hardware platform.

The conversation word matrix of two sets of data preprocessing, according to the TF-IDF model (standard k-means peace points k-means algorithm) and the enhancement of the TF-IDF model (WR-KMEANS method).

C. Experimental Results

We use the number of iterations of all three algorithms for the maximum number of 20 (in order to make a fair comparison). Each experiment running tenfold. We set the strength of the association between the two words threshold to 0.4.

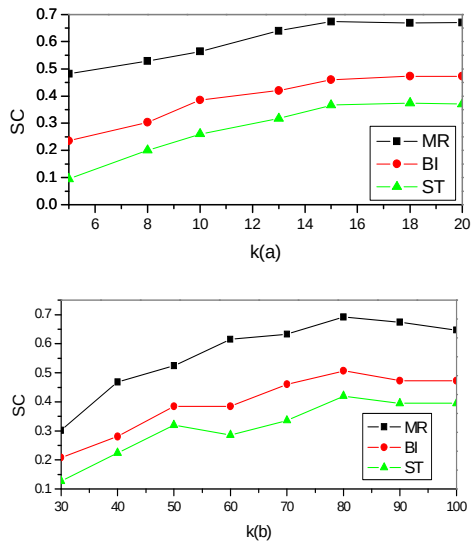

Figure 2. Comparing the best SC results on NG20(a) and on Reuters-21578(b)

The reason is that the extended vector space model richer semantic information than traditional TF-IDF model, and to strengthen the vector representation of the text content, the real theme. The carrier used, adding, in which only the relevant terms to calculate the similarity. This method brilliant avoiding the warp yarns from the sparse keywords, measuring the similarity of the text, thereby achieving a better efficiency than the original TF-IDF model method.

Experimental study of K NG20 Reuters-21578 on the effect of the SC, the results shown in Figure 1. SC Clusters divisible. WR-KMEANS point of the original class included in the data set, you can get a clear partition corpus can induce SC Figure 1 (NG200.67 when K = 20, Reuters -215780.69, K = 80). This is a reasonable result.

Extended vector space model, combined with the long-term mutual information, more knowledge of the language than the TF-IDF model, we can conclude that: From the above results. Context information needed to distinguish classes of documents.

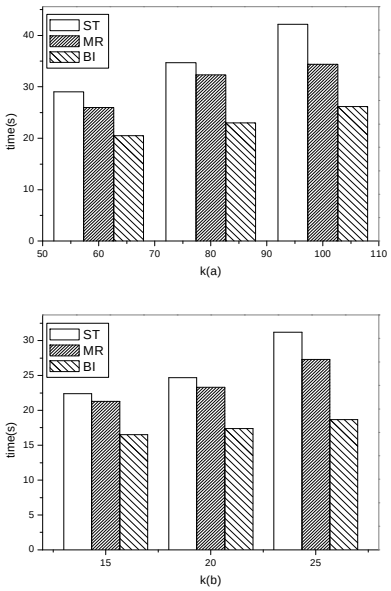

Figure 3. Comparing the best SC results on NG20(a) and on Reuters-21578(b)

Figure 3 illustrates the running time of the two data sets three algorithms. We can see, WR-KMEANS relative need more time than bisecting k-means, but a little faster than the standard k-means. The reason is, WR-kmeans is optimized in the pretreatment and text only, said not included in the clustering process. WR-KMEANS better benefits than the other two algorithms, efficiency are not much advantage.

V. CONCLUSION AND FUTURE WORKS

In this article, we focus on instant messaging cluster, and put forward the average WR-K method to solve sparse keywords, followed by the. WR-KMEANS automatic synthesis of instant messaging conversations, it has more keywords and more complete context information than simple message, and expand the traditional TF-IDF model dialogue assistance HowNet words. Experimental evidence indicates, WR-kmeans significantly outperformed traditional TF-IDF model based on the other two methods.

We plan to WR-K mean clustering, the initial partition optimization to improve speed. In addition, we have to analyze the IM networks, social network analysis, in future work.

REFERENCES

[1] J. Resig and A. Teredesai.: A framework for mining instant messaging services. In Proceedings of the 2004 SIAM Lake Buena Vista, Florida (2004)

[2] J. MacQueen.: Some methods for classification and analysis of multivariate observations. In proceedings of 5th berkeley SMSP, pp. (1967) 281–297

[3] Yi Guan, etc.: Quantifying Semantic Similarity of Chinese Words from Hownet. IEEE Proceedings of ICMLC02, Volumn 1. Beijing (2002) 234-239

[4] Sack, etc.: A Content-Based Usenet Newsgroup Browser. Proceedings of the international conference on Intelligent user interfaces. 233 –240. New Orleans, Louisianna, 2000.

[5] Faisal M. Khan, Todd A. Fisher, Lori Shuler, Tianhao Wu, and William M. Pottenger.: Mining chat-room conversations for social and semantic interactions (2002)

[6] Hearst, Marti A. TextTiling: A Quantitative Approach to Discourse Segmentation, Technical Report UCB: S2K-93-24, 1993

[7] Scott Deerwester, etc. Indexing by latent semantic analysis. Journal of the American Society of Information Science, vol. 41, issue 6, 1990, 391-407.

[8] Ding, C. H. Q. A probabilistic model for dimensionality reduction in information retrieval and filtering. In Proc. of the 1st SIAM, Raleigh, NC, 2000.

[9] Ikehara, S., etc. Vector space model based on semantic attributes of words. In Proc. of the Pacific Association for Computational Linguistics (PACLING), Kitakyushu, Japan, 2001.

[10] A.Daemi, etc. From Ontologies to Trust through Entropy, Proceedings of the International Conference on Advances in Intelligent System, Luxembourg (2004)

[11] Andreas Hotho, etc.: Ontology-based Text Document Clustering. KI 16(4) (2002) 48-54

[12] Strehl, A., & Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining partitions. Journal of Machine Learning Research, 3, 583–617.

[13] M. F. Porter. An algorithm for suffix stripping. Program, 14(3):130–137, 1980.