International Journal of Database Theory and Application
Vol.7, No.6 (2014), pp.21-28
http://dx.doi.org/10.14257/ijdta.2014.7.6.03

A Weibo Topic Tracking System based on K-means

Yun Liu*, Kun-Peng Xia and Jian-Xun Zhao

School of Electronic and Information Engineering
Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education
Beijing Jiaotong University, Beijing, 100044, China

bshen@bjtu.edu.cn, 610092885@qq.com, zhaojianxun124@gmail.com

Abstract

This article studied weibo text representation. For the weibo features such as short, real-time, colloquialism and originality, in the original vector space model, we propose a suitable method for weibo text representation. Make all the content words as feature words after participation. And we proposed T-TFIDF weight calculation method according to the features of weibo. According to the vector space model, we proposed a weibo adaptive topic tracking methods based on K-means clustering. Simulation analysis shows that, the method can by comparing the similarity micro-blog and sub topic vector set, determine whether weibo belonging to the topic.

Keywords: *Weibo; Topic tracking; K-means; Adaptive*

1. Introduction

The purpose of this article is to be able to get the message which the user interested in among the huge amounts of weibo. In other words, designing a weibo topic tracking system which can Detect the information flow in the related information by analyzing one or several weibo. Automatic track the topic of the follow-up reports, and get the evolution process of the subject. In this way, we can get the dispersed weibos together which can help the user know the outline of the event.

2. Related Work

2.1. Feature Weighting Algorithm

Weight value represents the important degree of the term in the document. First, I will introduce some of the feature weighting algorithm's basic concept. And proposed an algorithm suitable for weibo. At present, the most commonly used method is the TF - IDF algorithm. Then I will introduce the algorithm.

For the item of words in a document t_k we give the definition of the term frequency at first.

$$w\ (t_k) = tf\ (t_k) \times idf\ (t_k) \qquad (2\text{-}1)$$

Nt is the tk lexical item occurrences time in the document and N is the total word count of the document. If we only consider the term frequency calculating the weight of the word, a serious problem will come up. For example, if we talk about

ISSN: 2005-4270 IJDTA

the environment protect (in Chinese 环保), the word environment(in Chinese 环境) appearing in other articles which don't have any connect with the environment protect. So, there is no ability to distinguish the word environment in this document collection. Therefore, we need to weaken the weights of high frequency words within a document collection value. A simple method is giving the have frequency word in the document a lower weight. If a word in the document set appears in n document, its document frequency(Hereinafter referred to as df) will be defined as n. Because of df values are often larger so df usually mapped to a smaller space, and we define the inverse document frequency(Hereinafter referred to as idf) as:

$$\text{idf}（\text{t}_\text{k}）= \log \frac{\text{N}}{\text{df}（\text{t}_\text{k}）} \qquad (2\text{-}2)$$

Among them, N is the total number of document in the document set. In 1000 randomly grab weibo, pair of df for statistics. Table 1 shows one of the three words of the inverse document frequency.

Table 1. Inverse Document Frequency

word	Df	Idf
人	91	1.04
好	66	1.18
中国	43	1.36

For the tk of words in a document, tf and idf values can be grouped to use. Define the term tk weights as follows:

(2-3)

2.2. K-means Clustering Algorithm

K-means clustering algorithm [2] is one of the most important flat clustering. It has the advantage of simple concept and high speed. Its purpose is to make the document and its class center's Euclidean distance minimum after clustering. Class centers calculation formula as follows:

$$\overline{\mu}(\omega) = \frac{1}{|\omega|}\sum_{\overline{\chi}\in\omega}\overline{\chi} \qquad (2\text{-}4)$$

In the formula above each document can be represented as the normalized vector which length is 1. Each kind of center of mass is ideally a ball, and there is no intersection between the balls. In order to measure the mass center's representative for the class, we use formula (2-5) and (2-6):

$$RSS = \sum_{n=1}^{k} RSS_k \qquad (2\text{-}5)$$

$$RSS_k = \sum_{\chi\in\omega_k} |\vec{\chi} - \vec{\mu}(\omega_k)|^2 \qquad (2\text{-}6)$$

The steps of the K-means clustering algorithm is: First, randomly selected k documents from the document collection as the initial clustering center, said that the vector K vector for the seeds. Calculate each document and seeds vector's distance in the collection and classify them to the nearest seeds vector. According to the classification, calculate the center of each vector, and make k center vectors as the new clustering centers. Again to calculate each document and the clustering center distance, and adjust the clustering center. Repeat this process until reaching termination conditions.

3. Weibo Adaptive Topic Tracking Algorithm based on K-means

3.1. An Improved Algorithm based on Feature Weighting

After getting understand of the TF-IDF algorithm, we propose an improved algorithm combined with weibo's feature. In the case of a weibo analysis:

【短道速滑 500 米李坚柔奇迹夺金 中国短道队完成四连冠】2 月 13 日，赛前并不被看好的中国 27 岁选手李坚柔，在短道速滑女子 500 米决赛中，以 45 秒 263 的成绩夺得金牌，实现了中国短道队在该项目上的冬奥会"四连冠"，也为中国代表团迎来索契冬奥会的首枚金牌。五星红旗我为你骄傲！

As we can see, the weibo is composed of two parts. The words surrounded by 【】 function similar to the topic. And the other words will detailed describe the news. When we browse the home page of the Weibo, a lot of the weibos have a title surrounded by 【】 or ##. We can use this specific to improve the algorithm. The title can be the key sentence of the weibo, So, we can enhance the term weight of the words appeared in the title. Formula is shown below:

$$w(t_k) = tf(t_k) \times idf(t_k) \times title(t_k) \qquad (3\text{-}1)$$

$$\text{title}(t_k) = \begin{cases} 3, \textit{word appears both in} \text{ title and text} \\ 2, \text{ word appears only in the title} \\ 1, \textit{word appears only in} \text{ the text} \end{cases} \qquad (3\text{-}2)$$

We call it the T-TFIDF algorithm. We can do the vector normalization after calculating the term value. Formula is as follows:

$$w(t_k)_N = \frac{w(t_k)}{\sqrt{\sum_{k=1}^{n} w(t_k)^2}} \qquad (3\text{-}3)$$

3.2. Tracking Algorithm

3.2.1. Inverse Document Frequency

In information retrieval, document collection is certain and will not change over time. So when calculating the idf, it won't change over time, either. And this study is in the case of real time, increase gradually as fetching weibo, document collection is increasing as time changes. Sub topic vector weight, meanwhile, also need to update the idf, to ensure the comparison is under the same vector space. Specific process is as follows:

(1)Make the firstly grabbed N weibos and the given four one to four related weibo as the initial background document collection. These N weibos and the first topic vector's idf calculation is based on this background document collection.

(2)Grab N more weibos, when calculating the term weight, the background document collection idf based on is the former collection plus these N micro blogs. Sub topic vector weights, meanwhile, also need to update the idf, to ensure that under the same vector space similarity comparison. Repeat step (2) until the end of the track.

3.2.2. Similarity Calculation

In K-means based topic tracking process, we need to compute weibo vector and sub topic vector set similarity, that is, a vector with multiple vector similarity calculation. We need to consider two cases, one is an expression of the contents of a tweet involves only one topic, then it only with one child topic vector similarity is higher, and with other sub topic vector similarity is lower. Another case is a weibo content may contain multiple child involved in the topic, so it is with any one individual topic vector similarity degree is not high, but overall, the weibo is related with the topic. In both cases, weibo should be judged belongs to this topic. So does the similarity calculation, as long as satisfy any of the following two conditions, determining weibo belongs to the topic.

(1) When weibo vector and any vector of sub topic vector's similarity is greater than the threshold value of Ts, formula is as follows:

$$\mathrm{Sim}\left(T_m, D\right) > Ts \tag{3-4}$$

T_m is m sub child topic vector, $1 \le m \le M$, M is the number of topic vector。 D is the weibo vector。

(2) When weibo vector and sub topic's average similarity is bigger than Ts, formula is as follows:

$$\sum_{m=1}^{M} \mathrm{Sim}\left(T_m, D\right) > Ts \tag{3-5}$$

3.2.3. Dynamic Adaptive Process

Make the given one to four weibo connected to be a document, and said them with space vector. Then, the vector becomes the topic vector. Grab the weibo and compare its similarity with the topic vector, judging whether it belongs to the topic. This process may make some mistake. Overall, the topic is changing with time , so the "topic drift" phenomenon will occur. When topic evolved, a lot of sub topic will generate. If we

neglect updating the topic vector, we will miss a lot of related weibos. So, we proposes the adaptive tracking algorithm. When tracking the topic, we can update sub topic vector at the same time. In order to cover all aspects of the topic with the topic vector, specific algorithm is as follows:

(1)Make the given one to four weibo connected to be a document, and said them with space vector. Make the vector the first sub topic vector of the sub topic.

(2)Each article grabbing N weibo and detailed processing, transfer each weibo to space vector and calculate its similarity with topic vector set. If similarity is less than the threshold Ts, then determine the weibo has nothing to do with the topic and continue processing the next weibo.

(3)If the similarity is greater than the threshold Ts, determine the weibo and topic, and judge whether exist a boundary point. If the boundary point exist, jumping to step (3), otherwise, circulating step (2).

(4)After the step (3) processing, we get M + 1 clustering center. We make the M + 1 center vector as a new child topic vector set, and returns (2) to continue.

3.3. Abstract Topics

3.3.1. Automatic Abstract

Automatic abstract technology is a technology using computer aided to make text analysis, content induction, abstract generation. Abstract can be classified according to many standard automatically. According to the number of document automatically abstract, it can be divided into single document abstract and multi-document abstract. According to the abstract extraction method can be divided into excerpt type of abstract and Understand the type of abstract.

3.3.2. Weibo Automatic Abstract Algorithm

At the end of the topic tracking, we get a collection of weibo report under the topic, and the topic vector set. On this basis, this paper proposes a weibo automatic abstract algorithm which can get the automatic summarization of the topic. Specific algorithm is as follows:

(1)Make the n sub topic as the initial clustering center and cluster the weibo vector collection. Then, we can get n weibo clustering collection.

(2)Make weight calculation to each collection, chose the weibo which the weight value is biggest as the abstract of the class.

(3)Rank the abstract of each class, the final topic abstract can gotten.

In step (2), weibo weight calculation method is as follows:

$$tf(t_k) = the\ \ frequency\ of\ \ t_k\ \ appears\ in\ the\ class \qquad (3\text{-}6)$$

$$\mathrm{idf}(t_k) = \frac{the\ total\ number\ of\ \ the\ weibo}{the\ number\ of\ \ the\ weibo\ in\ which\ t_k\ \ appears} \qquad (3\text{-}7)$$

tk is the lexical item of the weibo and m is the number of the word in the weibo vector. $tf(t_k)$ is t_k's word frequency in the class. $idf(t_k)$ is the reverse word frequency which is the topic under all weibo number and the ratio of the word t_k appears. We can see from the above threes formulas, we improve the traditional TF-IDF term weight calculation. For the weibo is shorter, if we use the traditional calculation, the value of tf may be very low. Formula (3-6) calculates the word frequency in the class. We can get the weight of the word by multiplying (3-6) and (3-7). And the word appears not frequently will have higher weight.

3.4. Experiment and Result Analysis

According to the TDT usual evaluation index, select the following indicators to test the performance of topic tracking algorithm.

$$\text{Recall}=\frac{A}{A+C} \quad (3\text{-}8)$$

$$\text{Miss}=\frac{C}{A+C} \quad (3\text{-}9)$$

$$Precision=\frac{A}{A+B} \quad (3\text{-}10)$$

$$False\ Alarm=\frac{B}{B+D} \quad (3\text{-}11)$$

A is the report number when we correctly detect the topic. B is the report number when we falsely detect the topic. C is the report number the system missed. D is the report number which don't belongs to the topic.

3.4.1. TF-IDF Algorithm Compared with T-TFIDF Algorithm

In order to test the performance of the T-TFIDF, we take two groups of experiment. One group use the TF-IDF, while the other use the T-TFIDF. The result is shown in the Table 2.

Table 2. Experiment Result

	TF-IDF	T-TFIDF
Recall	77.63%	78.51%
Miss	22.37%	21.49%
Precision	69.56%	70.42%
False Alarm	1.30%	1.23%

3.4.2. Tracking Algorithm Comparison

This section will test the K-means method. we proposed in the paper. We compare the K-means method with the query vector method. The result is shown in the Table 3.

Table 3. Experiment Result

	Query vector method	K-means method
Recall	78.51%	79.52%
Miss	21.49%	20.48%
Precision	70.42%	70.43%
False Alarm	1.23%	1.15%

4. The Design and Realize of the Weibo Tracking System

4.1. System Functions Overview

Our system includes the following functions:

(1) weibo test collection: Use Weibo's API to collect weibo. Make permanent storage of the weibo in the database. The collected weibo can be the corpus of topic tracking module.

(2) weibo to quantify: Transfer the weibo text to the space vector by vector space model.

(3) weibo topic tracking: User can input one to four weibo under one topic. The system can analysis the weibo and track the related weibo under the topic.

(4) User and management: This part can add new users, view users' information, delete users and modify users' information.

4.2. System Overall Design

Figure 1 is the overall system design. The system is constituted by four modules. Each module represent a function.

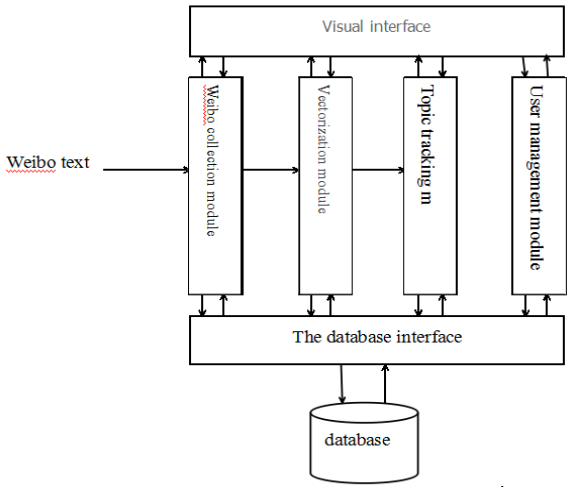

Figure 1. System Overall Design

This system implementation using Java code, the database using MySQL5.5 database. Visual interface is edited by JSP implementation. Use MyEclipse10.0 development tools, server using Tomcat7.0.

5. Research Prospect

Current research topic tracking is mainly from the methods such as information retrieval and natural language processing. This process ignores the real-time news, sudden, *etc.* So some researchers consider the time factor when calculating the impact on the topic, some researchers using named entities to characterization of news reporting. But these are all on the basis of the traditional methods of improvement. The future research direction is embodied in the following respects. Firstly, building new models. Secondly, design new algorithm which take the time factor into consideration. Thirdly, propose new information extraction method.

References

[1] Z. Liu, W. Yu, and W. Chen, "Short text feature selection for micro-blog mining", Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on. IEEE, **(2010)**.
[2] Y. Watanabe, Y. Okada, K. Kaneji, "Multimedia database system for TV newscasts and newspapers", Advanced Multimedia Content Processing, **(1999)**, pp. 208-220.
[3] B. Masand, G. Linoff and D. Waltz, "Classifying news stories using memory based reasoning", Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval.ACM. **(1992)**, pp. 59-65.
[4] J. Carbonell, Y. Yang and J. Lafferty, "CMU report on TDT-2: Segmentation, detection and tracking. Proceedings of the DARPA Broadcast News Workshop, **(1999)**, pp. 117-120.
[5] J. Kupiec, J. Pedersen and F. Chen, "A trainable document summarizer", Proc. of the 18th annual Int'l ACM SIGIR conference on Research and Dev't in information retrieval, ACM, **(1995)**, pp. 68-73.
[6] T. Strzalkowski, G. C. Stein and G B. Wise GE, "Tracker: A Robust, Lightweight Topic Tracking System", Proc. of the DARPA Workshop, **(1999)**.
[7] J. Yamron, S. Knecht, P. Van Mulbregt, "Dragon's tracking and detection systems for the TDT2000 evaluation", Proceedings of Topic Detection and Tracking Workshop, **(2000)**, pp.75-80.
[8] J. Allan, V. Lavrenko and D. Frey "UMass at TDT 2000", Proceedings of Topic Detection and Tracking Workshop. USA: National Institute of Standar and Technology, **(2000)**, pp.109-115.
[9] W. Lam, S. Mukhopadhyay and J. Mostafa, "Detection of shifts in user interests for personalized information filtering", Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, **(1996)**, pp. 317-325.
[10] Y. Lo and J. L. Gauvain, "The LIMSI Topic Tracking System for TDT 2002", Proceedings of Topic Detection and Tracking Workshop, **(2002)**, Gaithersburg.
[11] T. Dunning, "Accurate methods for the statistics of surprise and coincidence", Computational linguistics, vol. 1, no. 19, **(1993)**, pp. 61-74.
[12] J. Allan, "Topic detection and tracking: event-based information organization", **(2002)**, p. 19.
[13] H. P. Luhn, "The automatic creation of literature abstracts", IBM Journal of research and development, vol. 2, no. 2, **(1958)**, pp.159-165.

Authors

Yun Liu is a Professor in School of Electronic and Information Engineering in Beijing Jiaotong University where she received her PhD degree in the Communication and Information System. She is interested in Opinion Dyanamics, Network/Information Security, Computer Communication, and the Intelligent Transportation System.

Kun-Peng Xia received his B.E. degree in Communication Engineering at the PLA Information Engineering University in 2011. Currently, he is a graduate student in the Department of Electronic and Information Engineering, majoring in Communication and Information Systems.

Jian-Xun Zhao is a freshman in the Department of Electronic and Information Engineering at Beijing Jiaotong University.

