Applied Mechanics and Materials Vols. 303-306 (2013) pp 2289-2293
Online available since 2013/Feb/13 at www.scientific.net

doi:10.4028/www.scientific.net/AMM.303-306.2289

Extracting Hot Topics from Microblogging based on Keywords Detection and Text Clustering

Bo Yi[1, a], Yong Wang[2,b] , Xin Chen[3] , Ying Wang[4]

[1,2,3,4]Faculty of Computer Science, Guangdong University of Technology,Guangzhou, China

[a]yibo_grace@163.com, [b]wy_victor@126.com

Keywords: Microblogging,Keyword,Detection, Dynamic Clustering,Hot Topics

Abstract. Following with news and forums, microblogging becomes the third largest source of Internet public opinion. So it is necessary to do research of microblogging topic discovery. Firstly, we detect hot topics through the the keywords detection algorithm. Secondly, elect most popular microblogging text in the massive microblogging text by combining of keywords weigh and textual information entropy. Finally, using the dynamic clustering algorithm, the microblogging text elected, will form into different news topics by clustering polymerization. According to experimental validation of the true microblogging data, hot topic can be effectively detected from the text of a large number through the method.

Introduction

With the rapid development of the Internet on a global scale, network media following newspapers, radio, TV, has increasingly become the main channel for people to obtain and publish information. According to the China Internet Network Information Center (CNNIC) [1]to data released by the end of 2011, the microblogging users scale had reached about 250 million . The face of such vast amounts of data, research microblogging have important significance. For individuals, it is able to be kept informed of the latest hot spot information; For enterprises, it is able to understand the news related fields, and enhance the competitiveness of enterprises; for the country in terms of the relevant departments rumors, timely public opinion to guide and promote the guidance of public opinion, and to promote economic and social development of healthy and stable.

In recent years, automatic discovery of text mining with a hot topic and carried out extensive research. The hot topic of automatic discovery core topic detection and tracking technology (Topic Detection and Tracking TDT) [2], which aims to study the important information in the algorithm and to find a large number of news data streams, TDT technology in actual field important applications. But microblogging is essentially different from the traditional hot topic detection algorithm which does not fully meet the algorithm at the the microblogging hot topics detection. The large amount of data and microblogging there are a large number of the deformation words and new words, and the Chinese short-text feature word frequency and low, Yongheng Wang [3] propose a text clustering algorithm oriented mass phrase information The algorithm is based on frequent word sets, parallel to improve efficiency through the use of semantic information to improve the clustering accuracy. Microblogging data for the the microblogging produce fast, dynamic growth, Chunxia Jin [4] dynamic vector Chinese short text clustering algorithm, this algorithm for short text similarity drift problem, based on HowNet expansion related words set to build dynamic text vector method, dynamic vector calculation of Chinese short text similarity, and then found that the inherent relationship between short text, thus mitigating features of word frequency is too low and the existence of a deformation of the word and a new word clustering the impact of better clustering results.

The purpose of this article is intended to discover new microblogging from a large number of data, how to find hot topics from the vast amounts of data microblogging is the focus of this study, and in order to fit the characteristics of the dynamic growth of the microblogging clustering algorithm which is also the focus of our research.

Microblogging Topic Ideas And Algorithms

Method of thought and the basic framework

This method of thought , detected a time pane Popular keywords, in through the the keyword weight microblogging combining text information entropy popular microblogging elect a time panetext . This method can be taken out of the massive microblogging text most likely to describe a hot topic of microblogging text. Finally, using the dynamic clustering algorithm microblogging text, will be elected by the clustering aggregation into different news topics.

The framework of proposed method is shown in Figure 1.

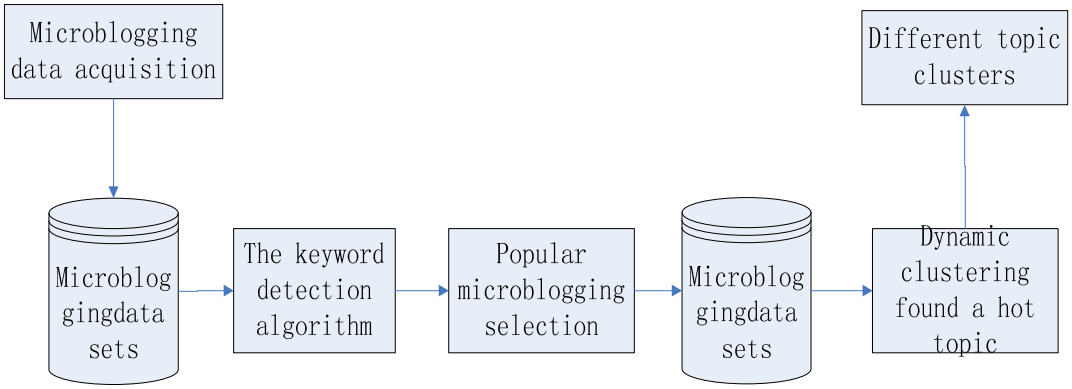

Figure 1 The overall framework

Keyword Detection

ICTCLAS segmentation system [5] a microblogging text segmentation, been a part of speech vector, each word has its own part of speech, such as nouns, verbs and other parts of speech. Word frequency statistics due to the relatively large contribution to the extent of the verb and noun expression of detection with the theme of the theme, we temporarily consider only the verb and noun. Statistically, we will set the unit pane T, T units pane Frequency Statistics, a period of time the word list.

The microblogging keyword detection, has a strong time-domain characteristics to represent word frequency of words in the time period of the growth rate of the introduction of growth coefficient based on this feature here. Formula is as follows:

$$G_{it} = \frac{F_{it}}{F_i} = \frac{F_{it}}{1/K * \sum_{j}^{K} F_{ij}}$$

The formula, F_{it} represents the frequency of the word appears in the time period , and then calculate the average of the time periods the words. The larger the value of G_{it}, the faster growth of word frequency, the more the term may be popular phrase.

In order to be able to elect real popular keyword, the use of word frequency and growth rate on the value of the composite sum, a subject headings weight:

$$w_{ij} = \log G_{ij} + \lambda \log \frac{F_{ij}}{F_{\max}}$$

In the formula , $F_{\max}$ that represents the frequency of the words in the maximum value in the time segment. λ can be adjusted by adjusting the value of the parameter, the ratio between the growth rate and word frequency: we can adjust the proportion of the relationship, depending on the environment, when word frequency plays a major role, the adjustment value is larger; When the growth rate plays a major role, the adjustment value small. When the larger the value, the more popular the word is likely that we need keyword.

Popular Microblogging Text Filtering

In massive microblogging information, how to choose the best performance of a period of the most popular microblogging text, here we use the microblogging two features to elect a certain time period that best represents the hot topic of microblogging text: information entropy the keyword weight.

Combination of the calculated values by weight of the the information entropy and keyword weight,Formula is as follows: $w = Entropy + \partial \sum_{i}^{k} w_{it}$

Its value is calculated each time period microblogging text, larger microblogging text selected in according with the threshold w, that is, a period of time microblogging text list. Microblogging text list can be a good representative of a certain period of time microblogging hot topic material text, thus completing the initial detection of microblogging text do data preparation work for the topic.

Text Representation Model

Set Feature Items And Calculation Feature Weight

In the text, the contribution of each word of the text. Important in order to reflect the different words in the text, in the vector space model for different words given different weights. Then construct the vector space model when calculating the feature item weights will be our focus. $TF - IDF$ (term frequency–inverse document frequency),The main idea is that if the the TF higher frequency of a word in a text, but rarely occurs in the other text, then this word has a good ability to distinguish between, and should have a larger weight.Formula is as follows:

$$a_{ik} = \frac{f_{ik} \log(\frac{N}{n_i})}{\sqrt{\sum_{j=1}^{n} \left[f_{jk} \log(\frac{N}{n_j}) \right]^2}}$$

Build VSM

Construction of vector space model (VSM) is through a text into a space vector language processing problems into easy to calculate mathematical problems. After word vocabulary in the text, after the calculation of the right of individual words heavy vocabulary weights constitute the dimensions of space, Thus, a text from the can into vector,Formula is as follows:

$$\vec{D}_i = (t_{i1}, w_{i1}; t_{i2}, w_{i2}; .....; t_{ij}, w_{ij}) \text{ , and } i \le j \le M$$

Similarity Algorithm

Because the microblogging is often different words to express the same semantics.In this paper, based on word semantic similarity algorithm, even if the same semantics to express different words associated with it. Formula is as follows: $sim(\vec{D}_a, \vec{D}_b) = \frac{\sum_{i=1}^{M_a} (t_{ai} t_{b1} s_{i1} + t_{ai} t_{b2} s_{i2} + ... + t_{ai} t_{bM_b} s_{iM_b})}{\sqrt{\sum_{i=1}^{M_a} t_{ai}^2} \times \sqrt{\sum_{i=1}^{M_b} t_{bi}^2}}$

Dynamic Clustering Algorithm

Dynamic clustering algorithm is simple, fast operation, the corpus of the input sequence is sensitive, but in the discovery process in the microblogging topic chronologically organization, the order of the input is determined, the input corpus is not the order of the algorithmwill have an impact, and the computing speed of the algorithm for computing time. Dynamic clustering algorithm calculation steps:

Step1, ready to be input text vector collection $\vec{D} = \{\vec{D_1}, \vec{D_2}, ..., \vec{D_N}\}$, similarity threshold based on experience with the training set, usually 0.6 to 0.9.

Step2, input text vector d, d individually prior Discussions of various reported similarity calculation $sim(\vec{D_a}, \vec{d_k})$, draw the largest similarity value ($S_a = \max_{k=1 to T} (sim(\vec{D_a}, \vec{d_k}))$), where in the total number of topics;

Step3, if the maximum degree of similarity is greater than or equal to the preset threshold, the microblogging text d assigned to text in class cluster, jump to Step5;

Step4, the greatest similarity is less than the preset threshold, the microblogging text d is not part of any existing topic cluster, create a new topic clusters based microblogging text d;

Step5, the end of the cluster, to wait for the next text input.

Empirical researches

From Sina or Tencent Weibo provide an open platform for data acquisition, The platform used herein Sina open API interface (http://open.weibo.com/) a microblogging data reptiles. The interface can get to the latest public microblogging, and ultimately the reptiles crawling from August 25 2012 to August 30, 2012, part of the Sina microblogging data. These microblogging about 3 million from about 200 million users to publish. Total it up, although the data from these micro-Bo is not Sina Bo Full Site data, but starting from the experimental point of view, the data amount is also relatively more, sufficient for the completion of the text of the experiment.

keyword detection evaluation

Table 1 The keyword detection fragment

Theme time	keyword	news events
2012 / 08 / 25	Harbin | Collapse |shabby	5 pm on August 24, 2012, Harbin Yangming Beach Bridge opened less than a year of fracture, shabby?
2012 / 08 / 25	Daimo Li | Out | China's voice | Inside story	2012/08/24 Lee on behalf of China's voice player to be eliminated off hot, many users questioned said: he's out is not a problem because the vocals, but otherwise Insider.
2012 / 08 / 28	Yan'an | Bus | Fire	Shaanxi Province, Yan'an City in serious traffic accident occurred in the early morning of the 26th, confirmed that a total of 36 people were killed.

As can be seen from Table 1, using keywords detection algorithm can extract out the hot topic of keywords.

Clustering algorithm evaluation

The meeting the evaluation established norms for the topic detection task, according to the characteristics of this article, we have chosen the recall rate, accuracy rate, undetected rate and false detection rate of detection as a topic of microblogging evaluation criteria.It is shown as Table 2:

Table 1 Topic detection evaluation results

	keyword dynamic clustering algorithm	Dynamic clustering algorithm	the reference range
recall rate	78.23%	59.93%	65%-88%
accuracy rate	74.33%	60.88%	65%-88%
undetected rate	21.27%	40.07%	65%-88%
false detection	1.78%	5.44%	0.1%-5%

By the test results show that the simple dynamic clustering algorithm, recall and precision rates low missing rate, overall performance deviation. Using the dynamic clustering algorithm has MeSH complex rate increase, but the recall rate and the accuracy rate has significantly improved, undetected rate and false detection rate is low, the overall performance of the improved algorithm better for the topic of microblogging is feasible and effective.

Conclusion

In this paper, the text data from the massive microblogging concentration detected news topic to do research. Firstly, combined with the keyword detection algorithm microblogging text information entropy algorithm to select the most likely to describe the topic of the press microblogging text

effectively solve the problem of massive data text, using the dynamic clustering algorithm to polymerize into different news topic, dynamic discovery news topics. For this work, there is still room for improvement. First, the accuracy of the experimental results obtained with the false detection rate considerable improvement in space, and detecting the effect will be limited by the calculated parameter value. Secondly, in order to improve the accuracy, sacrifice some algorithm speed. Therefore, further improve the algorithm speed and precision, the parameters of the adaptive selection, will focus on the direction of future research work.

References

[1] China Internet Network Information Center, China's Internet Development Survey Report. 2011

[2] J.Allan,J.Carbonell,G.Doddington, J.Yamron and Y.Yang． Topic detection and tracking pilot study: Finalreport [A] ． In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop,Virginia: Lansdowne,February 1998：194-218P

[3]Yongheng Wang , Yan Jia , massive short message text clustering technology . Computer Engineering, 2007,33 (14): 38-40

[4]Chunxia Jin , Haiyan Zhou , the dynamics of Chinese short text clustering . computer engineering and applications, 2011,47 (33): 156-158

[5]Zhang H P, Yu H K, Xiong D Y ,et al. HHMM-based Chinese lexical analyzer ICTCLAS[A] // Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17,2003[C]. Sapporo,Japan:Association for Computational Linguistics,2003:184-187

[6]M. Sahami and T.D. Heilman. A web-based kernel function for measuring the similarity of short text snippets[C]. In Proc of WWW`06,2006,pp 377-386

[7] Tao He ,Xianbin Cao , based on Chinese network short text clustering algorithm. Journal of automation, 2009,35 (7): 896-902

Sensors, Measurement and Intelligent Materials
10.4028/www.scientific.net/AMM.303-306

Extracting Hot Topics from Microblogging Based on Keywords Detection and Text Clustering
10.4028/www.scientific.net/AMM.303-306.2289