Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(3):869-873

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Extraction technologies of blog topic detection

Tao Kuang[1] and Shanhong Zhu[1,2]

[1]School of Computer and Information Engineering, Xinxiang University, Henan, China
[2]International School of Software, Wuhan University, Wuhan, China

ABSTRACT

The emergence of blog hot topic means that the user's interest, participation behavior and various media report coverage reach to its climax，a detecting method of topics on blog based on blog bursty words is proposed. It includes the use of word similarity measure and text clustering analysis which is combined with design strategy in specific period, the use of the main idea of the sudden vocabulary hot topic detection algorithm has to be used and improved in order to generate the final clustering. The experimental results show that the algorithm can obtain an accurate blog topic detection results.

Keywords: hot topic; bursty words; detection algorithm; design strategy

INTRODUCTION

Blog is online diary as a typical application of Web 2.0 technology which integrates personalized management, information sharing, views communication and other functions together organically. Blogs and online forums, although all belong to social networking platform, but the blog gives the user rights of owning private space and autonomous management and is more likely to show a user personalized features[1]. Blog topic presents the characteristics of diversity. Different subject reflects the different hobbies or blog using purpose. This paper analyzes the characteristics of different types of blog topics. How to classify topics in the blog space is a primary problem. Referring to the blog division standard of professor fangxing dong who divides the blog into the type of timeliness content, professional knowledge and personal communication mainly. Characteristics of different types of blogs which corresponds to the topic analysis are as follows:

1）Timely topics
On many blog sites, news blog which has typical and temporal features and the application of blog media property in the field of news is the platform in spreading the news information and news comment.

2）Professional topics
As the professional blog is concerned, its marked characteristics are that topics last time longer and burst characteristics of the subject is not obvious, the user's behavior take the knowledge sharing and technical communication as the main purpose, and the user's professional background is more specific[2].

3) Personal feelings topics
According to the blog behavior study record by China Internet network information center offering, recording personal emotional life is intentions that most Internet users register blog, so personal ex-change blog topics focus on personal life, work and study.

Due to the diversity of users' interests, blog has become an important platform to discuss the topic in the different domains. Topics on the blogosphere often correspond to a specific field, so mining the topic on the blogosphere to understand current needs of society and the social people from all works of life has great realistic significance [3]. Distinguish standard to classify the blog is actually relatively vague, Take news blog as example, for professional journalists, the blog has become a part of the personal work to some extent, professional standards on its blog writing constraint is very strict. They can use blog platform to communicate between peers or sponsored some advocacy campaigns, so journalists of the blogosphere are considered as professional blogosphere. For ordinary users, merging into those experts in the blogosphere can not only meet the demand of entertainment, but also enrich their knowledge, and ordinary users can also resort their own thoughts to cause the attention of professors in the relevant departments and gain more supports [4].

Due to the diversity of users' interests, blog has become an important platform to discuss the topic in the different domains. Topics on the blogosphere often correspond to a specific field, so mining the topic on the blogosphere to understand current needs of society and the social people from all works of life has great realistic significance [5]. Distinguish standard to classify the blog is actually relatively vague, Take news blog as example, for professional journalists, the blog has become a part of the personal work to some extent, professional standards on its blog writing constraint is very strict. They can use blog platform to communicate between peers or sponsored some advocacy campaigns, so journalists of the blogosphere are considered as professional blogosphere. For ordinary users, merging into those experts in the blogosphere can not only meet the demand of entertainment, but also enrich their knowledge, and ordinary users can also resort their own thoughts to cause the attention of professors in the relevant departments and gain more supports [6].

Because of the intention of the ordinary users to join the blogosphere is for leisure and entertainment, that is, the user interest occupies the dominant position, so the type of news blogosphere can be considered as professional blogosphere. This paper expects to distinguish blog type according with the behavior of the majority of users in the blogosphere, the certain types of topics on the granularity of semantic expression of positioning will be difficult, such as the entertainment star himself and relevant topic of such event can be used as a unit[7]. This paper embarks from the topic of temporal features, the characteristics of temporal topic is that the more popular it is, the features of topic content keywords are more obvious, so for a specific topic in the field, this paper proposes a blog hot topic detection method based on bursty vocabulary.

THE KEYWORD EXTRACTION TECHNOLOGYBLOG TOPIC MODEL METHOD

A major goal in the field of topic detection and tracking is to study how to accurately identify from the sample new topic and track some topics[8]. Whether text clustering method is adopted to discover new topic or text classification technology is used to track the existing topic, the text expressing has a direct influence on the effect of text mining technology. Vector space model (VSM) is a commonly used method to represent text, the basic idea is taking the text as a collection of words, each word in the text are endowed as different weights, the more vocabulary expressing the theme of the text, the greater its weight. Heavy words has played an important roles as the main characteristics of the text when calculating the text similarity, therefore the process of extracting keywords is the assessment of blog different vocabulary weights and the lexical selection process according to the weights[8].

At present, there are several main vocabulary weights calculation methods in the document (referred to as TFIDF), information gain, mutual information, and the complex network method[9]. TFIDF method considers that the frequencies of a word in a document related to the importance of it in the document, the more frequently used the words, the closer the relationship between vocabulary and document it will be, but due to the frequency some commonly used function words in the document collection is very high, so the document in terms of number of document (the document frequency) was used to measure the particularity of vocabulary, word document frequency is lower, the greater the importance of vocabulary is in the document.

Information gain method mainly consider the text in contains a word after the change of information entropy, so as to determine the importance of vocabulary for text classification judgment. Mutual information considers the relationship between words and text category, if the word appears only in a certain category samples, the value of the mutual information between the categories is the largest[10]. Card square inspection method considers both the words appearing in the text categories or not. The study indicates that a kind of complex network is formed between the vocabularies of the text, therefore some keywords extraction method based on complex network have been put forward. Such as Zhu build word co-occurrence network with a text word co-occurrence relation, and then extracted text keywords by assessing the effect of vocabulary to vocabulary network average path length to. Zhao peng use document language network characteristics of the complex networks, such as accessing the center degree of the vocabulary and clustering coefficient to extract the Chinese document keywords[11]. On the whole, the method based on complex networks is still in its infancy, and the method is more complex. Due to the information gain, mutual information and the square inspection methods need a lot of the known categories of training samples, and news is unpredictable and new words may appear in events. Hence this article chooses the simple and practical

TFIDF method to put forward the key words in the text.

THE KEYWORD EXTRACTION TECHNOLOGYTEXT SIMILARITY MEASURE METHOD

The methods of text similarity measure are the commonly used Euclidean distance, cosine similarity and Jaccard coefficient methods. Euclidean distance and cosine similarity are mainly used in the vector space model which feature vector is used to represent text. Each dimension of vector corresponds to a key word, the dimension of value corresponds to the weightof the key words in text. In n dimensional vector space, for example, the text D_i represents (term$_1$, weight$_1$, I; term$_2$, weight$_2$,term$_n$, weight$_{n,i}$).where weight, j, i are for vocabularies,term j expresses weights in text D_i.Euclidean Distance measures documents similarities based on document feature vector Distance. In n dimensional vector space, Euclidean Distance calculation formula of text D_1 and D_2 is as follows

$$Euc(D_1, D_2) = \| D_1 - D_2 \| = \sqrt{\sum_{i=1}^{n} (weight_{i,1} - weight_{i,2})^2} \tag{1-1}$$

The smaller the Euclidean distance is,the greater the similarity of text D_1 and D_2 is

Cosine similarity in which vector Angle cosine is used measure the similarity of two texts. In n dimensional vector space, text D_1 and D_2 cosine similarity computation formula is as follows：

$$Cos(D_1, D_2) = \frac{D_1 \bullet D_2}{\| D_1 \| * \| D_2 \|} = \frac{\sum_{i=1}^{n} weight_{i,1} * weight_{i,2}}{\sqrt{\sum_{i=1}^{n} weight_{i,1}^{2}} * \sqrt{\sum_{i=1}^{n} weight_{i,2}^{2}}} \tag{1-2}$$

The bigger cosine similarity is, the greater texts D_1 and D_2 are, the greater the similarity. When using the set of words to represent text, Jaccard coefficient is the common way to measure the similarity between the texts. The text similarity calculation principle based on Jaccard coefficient is for calculating the contact ratio between two collections, the greater the contact ratio is, and the greater the similarity between the texts is. D_X represents text word set X, set D_Y expressed words text Y, Jaccard coefficient calculation formula is as follows:

$$JC(D_X, D_Y) = \frac{| D_X \cap D_Y |}{| D_X \cup D_Y |} \tag{1-3}$$

where $D_X \cap D_Y$ is intersection between set D_X and D_Y, $D_X \cup D_Y$ is denoted as collections of D_X and D_Y, formula of molecular expresses intersection capacity and the denominator expresses set capacity.

In a text vector space model, high dimension calculation and data sparseness seriously affected the Euclidean distance and the effect of cosine similarity method, on the contrary Jaccard coefficient method is more flexible, which based on the Jaccard coefficient method to compare similarity between the posts.

Hierarchical clustering is one of the more popular texts clustering method,text clustering can be realized by calculating the distance of feature vector between text. According to the execution strategy, hierarchical clustering can be divided into the condensing and hierarchical clustering two clustering methods.

In the current field of topic detection, most of the hierarchical clustering methods adopt the hierarchical clustering strategy. As shown in figure 3-1, condensed hierarchical clustering adopts step by step from a bottom to up merging principle, firstly, take each text as a separate class cluster, and then merge the highest two similarities of class cluster, until class clusters reach to a certain size of the clustering, or to a termination condition.

In minimum distance method, if the shortest distance between two kinds of clusters in the sample is less than a certain threshold, then merge two clusters. Its formula is:

$$MinDis = \min_{Z_i \in C_i, Z_j \in C_j} \| Z_i - Z_j \| \tag{1-4}$$

In maximum distance method, if all sample distances between two clusters is less than a certain threshold,then merge two clusters. Its computation formula is

$$MaxDis = \max_{Z_i \in C_i, Z_j \in C_j} \| Z_i - Z_j \| \tag{1-5}$$

if average distance between all samples in one clustering and another clustering is less than a certain threshold, then merge two clusters. Its computation formula is as follows:

$$Dis = \frac{1}{n_i * n_j} \sum_{Z_i \in C_i} \sum_{Z_j \in C_j} \| Z_i - Z_j \| \tag{1-6}$$

Hot topic is usually a very small proportion, if used to identify the topics to each blog, the system efficiency will be very low. The more popular the topic is, the higher expression of topic words frequency is, the greater the correlation between the related words will be, hence vocabulary network based on the correlation between the high frequency keywords in the specified period can reflects the period of the hot spots to a certain degree. According to the above ideas, this paper constructs vocabulary network based on the blog keywords co-occurrence relation in a specific period, and then realize the vocabulary in the edge network clustering to determine the topic content, according to the blog and comments from users interaction to assess subject bursty degree and select hot topic. The specific process is shown in table 1. Hot topic detection strategy based on sudden vocabulary work center of gravity is to extract and evaluate the topics in different periods; the extraction of topic depends on vocabulary network construction in different period, the subject hot evaluation emphasis topic eruption in this period of time [12]:

So the algorithm's time complexity is O(n) , the process as shown in the algorithm 1, the following four subsystems tasks need to be solved:
1).How to extract the post keywords;
2). How to measure the degree of lexical co-occurrence between;
3).How to implement the vocabulary the edge clustering in the net;
4). How to assess the topic of unexpected degree.

BLOG TOPIC DESIGN SCHEME OF THE MODELTHE REPRESENTATION OF A TOPIC MODEL

Events and activities by the TDT topic is subject component elements by the definition of TDT. Events in the field of topic detection and tracking which are defined by some reasons and conditions occur in special time and place, involving some objects and associating with some inevitable result. In real life, natural disasters and traffic accidents happen in emergency situations, and phenomenon which is processed timely by relevant departments or prevented from subsequent events occurring is generally classified as a category. Activities in the field of topic detection and tracking which is defined as occurs in a particular time and place, have a common purpose and focus on a collection of related events. In real life, the events due to political or social needs of the large sports games, initiated by the state or social phenomena, such as political campaign and war, often trigger a series of subsequent events, and the related events often are recorded as a whole or to be talked about.

Look from the level of the semantic expression, topic that is more abstract concept than events and activities can understand into induction and summary of the related event information. In terms of the habit of bloggers, when bloggers capture the news event information, the purpose of bloggers publishing blog posts is opening the latest progress of events or comment on the effects to the society. Due to bloggers reporting or commentary Angle is different, the business of the post corresponding to the different aspects of events, and other user's attention and participation also stimulate the subsequent events reported events, so that in the whole process of the incident reporting and discussing, topic is reinterpreted unceasingly, therefore looked from the time distribution, temporal news topic model has the characteristics of multicenter or multi-sided. That is subject presents different characteristics in different periods.

Model is divided into topics layer, events layer, and subtopic layer. Topic layer shows the topics information’s reported in different periods, event layer reflects the related event information of incurring topic or composing subject, subtopic layer reflects talking about the topic for events in different aspects. The arrows of the event in events layer indicates the evolution relationship, such as temporal relationship or causal relationship. The arrow on the center of the Subtopic layer indicates the topic change. Topic model in the event the cause and effect of layer and the subject shows the topic. In terms of storage structure, topic model reflects the structure characteristics of tree, namely the topic layer depositing root node directing to layer nodes of the events, events on the node to child topic layer on layer of nodes, sub topic layer on the node to a leaf node (post), so the topic building model creation process is to build from leaf nodes to root node index structure.

TOPIC BUILDING PROCESS

From post focus to build and update model is also the process of topic detection and tracking, by topic model hierarchy is visible, event is the foundation of the topic model build, so blog topic model to build two main work is: based on post event information on clustering; Based on the topic information for clustering event set (post). Around the two parts work topic detection and tracking algorithm is proposed as shown in Table 1.

Table 1：Topic detection and tracking based on information model

```
Input：Blog sample set S，Number of time units n
Output：S The topic list Test
For i in S
    Extract the post I corresponding subject model；
End for
For j=1 to n
    detect events within the time unit j Based on the sub topics；
    If j==1 then
        model was constructed and deposited in the Test Based on the event topic；
    Else
        Find new and tracking the old events；
        Create a new model and existing in the Test or update Test subject model；
    End if
```

CONCLUSION

Due to the efficiency of the news topic and report or comment on the same event post will be concentrated in a specific time period, words associated with an event so will present a sudden, and the more hot topics, vocabulary characteristics of sudden, the more obvious. These words reflect the post event information of the important characteristics, according to the characteristics of the different information layer, topic model building strategies are to extract the keywords of each post to judge blog corresponding subtopic, and then post on event level (a subject) clustering and calculate the weights of each topic, the final judgment on topic hierarchy relationship between news events is to combine related events. The model has good accuracy in detection microblogging topic and easy to implement.

REFERENCES

[1]C.C. Yang, X.D. Shi, C.P. Wei. IEEE Transaction on Systems, Man, and Cybernetics-Part A: *Systems and Humans*, **2009**,39(4): 850-863.
[2]K.Y. Chen, L. Luesukprasert, S.C.T. Chou. *IEEE Transactions on Knowledge and DataEngineering*, **2007**, 19(8): 1016-1025.
[3]C.H. Wang, M. Zhang, S.P. Ma, et al. Automatic online news issue construction in webEnvironment[C]. Proceedings of the Seventeenth International Conference on World WideWeb, Beijing, New York: *Association for Computing Machinery*, **2008**: 457-466.
[4]K.Y. Chen, L. Luesukprasert, S.C.T. Chou. *IEEE Transactions on Knowledge and DataEngineering*,**2007**, 19(8): 1016-1025.
[5]S. Patterson, B. Bamieh. Interaction-driven opinion dynamics in online social networks[C].Proceedings of the First Workshop on Social Media Analytics, New York: Association forComputing Machinery, **2010**: 98-105.
[6] M.D. Mumford. *Personnel Psychology*, **1983**, 36(4): 867-881.
[7] NingZhong,Jian Hua Ma,Run He Huang,Ji Ming Liu,Yi Yu Yao,YaoXueZhang,JianHui Chen. *The Journal of Supercomputing* . **2013** (3)
[8] Li Zhao,XiaohongGuan,Ruixi Yuan. *Knowledge and Information Systems* . **2012** (2)
[9] Yung-Ming Li,Cheng-Yang Lai,Ching-Wen Chen. *Information Sciences* .**2011** (23)
[10] JamyLi,MarkChignell. *International Journal of Human - Computer Studies* . **2010** (9)
[11] ZhongfengZhang,Qiudan Li. *Expert Systems With Applications* .**2010** (6)
[12] C. Jensen,,C. Sarkar,C. Jensen, et al.Tracking Website Data-collection and Privacy Practices with the Iwatch Web Crawler. Symposium On Usable Privacy and Security . **2007**