Analysis of Hot Topics on Microblog Based on Findings and Social Governance

Chan Xu*¹, Rodrigo Moreno²
¹Wuhan University of Technology, Wuhan, 430070, China
²Teledyne Scientific Company, Leon, France
*Corresponding author, e-mail: 253591338@qq.com

Abstract

In the finding of hot topics on microblog, the short text, less word, non-standard word use and other features of microblog have made the traditional identification method of hot topic powerless. To solve this problem, discovery method of microblog hot topic based on speed increase has been put forward. First, divide pretreated microblog according to windows with equal quality, count word frequency of each word in each window and express as two-tuple sequence of time; then calculate increase slope of each word in every two adjacent windows to find words with faster increasing speed; later calculate increasing speed of users and article number of microblogs related to the word to make sure whether the word is hot subject term; finally produce hot topic from the cluster of hot subject term. The feasibility of this method has been verified by experiment. Experimental results show that this method has improved identification efficiency and lowered omission ratio and fall-out ratio, which can effectively and promptly discover the hot topic of microblog.

Keywords: Topic identification; Two-tuple sequence of time; Hot topic of microblog; Analysis on public sentiment

1. Introduction

As a new research direction of information retrieval, topic identification aims to monitor reports to realize automatic identification [1] for new topics in information stream of news media. Comparing with information retrieval, information extraction, text mining and several cross-over studies, topic identification puts more emphasis on discovery ability of new information, and concerns information related to a particular topic rather than relatively broad subject categories. Frequently-used topic models include vector space model (VSM) and Language Model (LM) currently. Vector space model is a topic model based on bag-of-words description and language model is probabilistic model [2] [3]. Typically, researches of topic identification focus on clustering algorithm of relevant reports. CMU [4], Weihua Luo [5], Xiang-Ying Dai [6], Bin Ma and others all use clustering algorithm for topic identification. The defect of clustering algorithm is that complexity of time and space is relatively high, and the quality of clustering center directly affects the accurate identification of follow-up reports. Apart from clustering method, probabilistic method can also be used for topic identification. Loulwah Alsumait has used empirical Bayes Theorem in topic identification [8,9], but the relevant researches are very few.

2. Discovery Model of Hot Topic on Microblog

In basic belief network model [10-12], the term set \( U = \{ k_1, k_2, ..., k_i \} \) is a discourse, and has defined sample space for belief network model. \( u \subset U \) is a subset of \( U \), and \( g_i(u) = l \iff k_i \in u \). Every index term is considered as a basic conception, thus \( U \) is seen as a concept space and concept \( u \) is the subset of \( U \). Document and user queries are expressed with the concept of concept space \( U \).

Probability distribution defined on sample space \( U \) is shown below. \( c \) is a concept on \( U \), which means a document or a user query.

\[
p(c) = \sum_{u} p(c \mid u) \times p(u)
\]
Equation (1) defines $p(c)$ as coverage degree of $c$ in space $U$; equation (2) represents that all concepts in concept space have equal probability to occur.

Figure 1 shows the topological structure of basic belief network; query $q$ is modeled as network node $q$ associated with random variable of binary system; node $k_j$ which constitutes term of query concept points to query node $q$. Document $d_j$ is also modeled as network node $d_j$ associated with random variable of binary system; node $k_t$ which constitutes term of query concept points to document node $d_j$.

The sorting of document $d_j$ related to given query $q$ is understood as a concept matching relationship, which reflects the coverage degree of concept $d_j$ provided by concept $d_t$. Therefore, use $P(d_j|q)$ to calculate the sorting of document $d_j$ concerning query $q$ in belief network. According to conditional probability, formula (2) and Bayes Theorem, this can be obtained:

$$P(d_j | q) = \eta \sum_{u} P(d_j | u) \times P(q | u)$$

$\eta$ is a normalization factor wherein. Different provisions on $P(d_j|u)$ probability and will correspond to sorting policies of different information retrieval models.

![Diagram](image)

Figure 1. Basic belief network

3. Identification Model of Hot Topic on Microblog

3.1. Modeling Basic

Topic identification model on the basis of Bayesian Belief Network proposed by the work is based on the following basic work:

(1) Apply terminology bank of report $S = \{s_{k_1}, s_{k_2}, \ldots, s_{k_n}\}$ built through samples reports. Certain word segmentation technology can be use in its establishment process to obtain some terms to represent each sample report $s_i$. The weight of term can be calculated by formula (4). The whole terminology bank is composed of all terms obtained; if a reduplicative term appears, then use the average value of weight to represent the weight of the term, and label it to indicate its source.

$$w(s_{k_j}) = \frac{freq(s_{k_j}) + 0.5N_{beg} + 0.5N_{end} + N_{title}}{\sum freq(s_{k_j})}$$

Wherein, $w(s_{k_j})$ is the weight of term $s_{k_j}$; $freq(s_{k_j})$ is the frequency of term $s_{k_j}$ appearing in report $s_i$; $N_{beg}$, $N_{end}$, and $N_{title}$ respectively represent the frequency of the term appearing at the beginning, end and title of report $s_i$; $\sum freq(s_{k_j})$ represent sum of frequencies of all items in report $s_i$. 

(2) Terminology bank of event built; each event is composed by several reports. It is more complicated if simply use feature items of all reports belonging to the event to represent, so certain methods can be used in this step to simplify set \( E \) of event term. We can reselect the first \( m \) feature items with more weight to constitute set \( E' \) from selected feature items of each report to describe event which it belongs, namely:

\[
e' = \bigcup E'
\]

\[
E = \bigcup e_i = \{ e_k, e_k, \ldots, e_k \}
\]

The weight of term \( s_jk_i \) in set \( e_i \) is mainly based on the calculation of frequency appearing in report which it belongs, but all reports are developed around a core report. Thus in event description, the weights of terms appearing in core reports shall be increased. Increasing method proposed by formula (6) can be adopted:

\[
w(e_i k_i) = \alpha \times w(s_jk_i)
\]

Wherein, \( w(e_i k_i) \) is the weight of term \( e_i k_i \) in event \( e_i \); \( w(s_jk_i) \) is the weight value of term \( s_jk_i \) in core report \( s_j \); \( \alpha \) is a factor for increasing, and its value is more than 1.

(3) Topic bank built \( T = \{ t_1, t_2, \ldots, t_t \} \). Its establishment method is similar to the establishment from terminology bank of report to terminology bank of event. Description terms of each topic \( t_i \) in topic bank root in the corresponding terminology bank of event, and these terms can be obtained by the partial extraction principle above. As for the weight calculation of term in topic \( t_i \), increasing strategies given by formula (6) can also be adopted; that is, if term \( k_i \) describing topic \( t_i \) comes from core events, then its weight is:

\[
w_i = \beta \times w(e_i k_i) \quad (\beta > 1)
\]

3.2. Topological Structure and Probability Calculation of Model

Based on built terminology bank of report, terminology bank of event and topic bank, topological structure of belief network model that is suitable for topic identification has been established in the work as shown in Figure 2; and probability calculation for the identification of new report \( s_i \) has been provided. Decide whether \( s_i \) is existing topic or new seeded topic through the comparing of threshold values.

![Figure 2. Belief network for topic identification](image_url)
pointing directions of arcs indicate index relationship, for example, term \( s_k \), \( s_j \), \( s_k \), \( s_n \) in terminology bank of report all have arcs pointing to new report, thus. It can be found from above figure that the whole network is belief network, but the part enclosed by gridlines is inference network; thus relevant knowledge has also been involved during the inference procedure of probability. Assuming that prior probabilities of all topics \( t_i \) are equal and based on Bayesian probability, conditional independence assumptions and related knowledge of inference network, the following formula can be obtained:

\[
\text{Sim}(t, s_n) = P(s_n | t_i) = \alpha P(t, s_n) = \alpha \sum_{e} P(t_i | s_n, e) P(s_n, e) = \eta \sum_{e} P(t_i | e) P(s_n | e) P(e | s)
\]

Wherein, \( s_n \) is the subset of set of report term; \( \eta \) is a normalization factor; the premise of above formula establishment is to assume that the occurrence of all subsets occur at an equal probability. Conditional probability \( P(t_i | e) \), \( P(s_n | e) \) and \( P(e | s) \) can be calculated by formulas (9), (10), (11):

\[
P(s_n | e) = \begin{cases} 1 & \text{if } \forall i (g(s_n) = g(u)) \\ 0 & \text{otherwise} \end{cases}
\]

\[
P(t_i | e) = \frac{\sum_{i} w_i \times w_{e_i}}{\sqrt{\sum_{i} w_i^2} \times \sqrt{\sum_{e_i} w_{e_i}^2}}
\]

\[
P(e | s) = \frac{\sum_{i} w_i \times w_{e_i}}{\sqrt{\sum_{i} w_i^2} \times \sqrt{\sum_{e_i} w_{e_i}^2}}
\]

Weight \( w_i \), \( w_e \) and \( w_{e_i} \) can be obtained by the calculation of formula (4), (6), (7) respectively. Besides, in order to simplify calculation, only consider subsets composed by at least \( \delta \) terms from subset \( s \) for subset \( e \) of event term, which means that both of them shall have \( \delta \) same elements at least.

Obtain \( \max_{i,j} \text{Sim}(t_i, s_n) \) from formula (8); if its value is greater than given threshold \( \theta \), classify new report \( s_n \) as topic with greatest similarity, and otherwise output it as a new seeded topic.

4. Experiment

4.1. Test for Set

TDT4 is used for evaluation of experiments in the work. TDT4 includes 98,245 news reports, with news reports from 20 news sources including APW, NYT, ABC, VOA, Xinhua and others from October 2000 to January 2001, which involves three languages as English, Chinese and Arabic. American Language Data Consortium (LDC) has labeled 40 news topics wherein and provided relevant and irrelevant reports of the 40 topics. The Chinese reports in TDT4 have been used in experiments to conduct performance test for the model proposed in the work.

4.2. Evaluation Method

The first influence factor of model identification performance is the expression of new reports. If there are more terms in vector quantity of new reports, then there is more query information; if the information is all useful, then the identification performance can be improved; if there are too many terms in vector quantity of new reports, there may be some redundant
information and it will reduce the identification performance. The influence of term number in vector quantity of new reports on precision ratio will be verified with different recall ratios at experimental part to make sure the optimal number of term in vector quantity of new reports. The calculation formula of precision ratio and recall ratio is as follows:

\[ \text{Precision} = \frac{a}{a+b} \quad \text{Recall} = \frac{a}{a+c} \]  \hspace{1cm} (12)

Wherein, \( a \) is the number of relevant documents identified; \( c \) is the number of relevant documents unidentified; \( b \) is the number of irrelevant documents identified.

Missing report rate and false positive rate involved in experiments are two coordinates in DET curve. The former is the vertical coordinate and the latter is abscissa. The calculation method is as follows:

\[ \text{MP} = \frac{c}{a+c} \quad \text{FAP} = \frac{b}{b+d} \]  \hspace{1cm} (13)

The meanings of \( a \), \( b \) and \( c \) are the same as above; \( d \) is the number of irrelevant documents unidentified. Although two different methods of evaluation have been used in the experimental part of the work, certain relationship can be found between the two from observing formula (12), (13).

4.3. Experiment Content and Analysis

The method of repeated training is adopted in the first experiment to obtain optimal number for expression. Experimental procedures of the second experiment are as follows: first, generate report vector, event vector and topic vector on the basis of test set, and use formula (4) to calculate term weight in vector; then carry out test in new model and vector space model (VSM) by using Chinese reports in test set, and calculate missing report rates and false positive rates of the two models under different threshold values. As for the third experiment, first adjust term weights for event bank and topic bank according to formula (5), (6), and then compare missing report rates and false positive rates before and after adjustment under different threshold values.

![Figure 3. Influence of different number of term expression used by new report on precision ratio](image)

(1) Influence of new report vector on identification performance of model: Whether using vector space model (VSM) or new model proposed by the work, expression of new report vector is a direct factor influencing identification performance when the identification upon new report is
carried out. 5, 10, 15, 20, 25, 30, 50, 100 feature items are respectively used to represent for new report vectors when recall ratios are 20%, 40%, 60%, 80%, 100% respectively in experimental part to verify the influence of precision ratio, thus to obtain optimal number for expression of new report vector. The result of the experiment is shown in Figure 3.

Figure 3 shows that choice of feature item number of new report will influence precision ratio. It can be found from the observation on dash lines in figure that the precision ratio is the highest when the number of feature items is around 25 under different recall ratios. Based on this experiment, the expression of report all adopt first 25 terms with bigger weight to represent in experiment of comparing the performance of vector space model (VSM) and new model, before and after weight adjustment of new model.

(2) Performance comparison between vector space model (VSM) and new model: What the experiment verifies is the performance of new model weight before adjustment and vector space model (VSM). DET curve is adopted as evaluation method. In this experiment, obtain missing report rates and false positive rates of the two models under 12 threshold values and draw DET curve. Finally, two curves can be obtained; and the closer the curve to the origin, the better performance it has; the result of experiment is shown in Figure 4. The black full line represents the performance of new model. It can be found from the observation on Figure 4 that new model is closer to origin than basic vector model as for most points on curves, thus the performance of new model is better than vector space model as a whole.

(3) Performance comparison between weights of new model before and after adjustment: Based on the above work and according to basic weight calculation method, increasing factors are used to adjust the term weight of different layers. Figure 5 is the performance curve before and after weight adjustment of new model drew by using DET curve.
Figure 5. Performance comparison between weights of new model before and after adjustment.

It can be found from the observation on Figure 5 that pre-adjustment performance is slightly higher than post-adjustment one at the beginning part with higher missing report rate. However, in the actual system design, a condition with very high missing report rate will not be chosen in the setting of threshold value since the threshold is relatively high under this condition, and the threshold will filter out a lot of useful information. Therefore, the weight performance before adjustment is obviously higher than the weight performance after adjustment from application and overall effect point of view.

5. Conclusion

Bayesian belief network model for topic identification has been proposed in the work according to the relevant knowledge of Bayesian belief network and topic identification. In the modeling process mapping from report layer to event layer, there may be some deviations due to term reduction. But when terms are filtered by the work, the terms selected have relatively large weight in report, and can basically represent the report; compared with time cost of large-scale news report processing, the deviation can be accepted. In the future work, we will further improve the model and better balance time cost and deviation.

References