Hot Topic Mining based on the Heat of Micro-blog

Wang Siyao

627188726@qq.com

Abstract—**With the popularity of social networks, users interact with each other and comment on current events through online social network more and more frequently,how to extract the hot topic has become the focus of natural language processing research.In this paper, we propose a hot topic extraction method based on micro-blog popularity. Combining the heat of micro-blog and word2vec model to assign weight for each word, and we use bidirectional LSTM to conduct document semantic coding and single-pass method for topic mining.The experiment results show that the proposed method performs better and has stronger robustness than the traditional topic detection method.**

Keywords—OSN，topic mining, word2vec, Neural Network

I. Introduction

Online Social Networks(OSN) have gradually become popular in daily life and become one of the most important media for users to interact with each other, comment on current events and keep track of topics,like Facebook ,Twitter as well as Weibo. By 2016 there are more than 2.2 billion registered users in Facebook. In the second quarter of 2017, there are more than 2 billion active users in Facebook. How to effectively extract the topic of the user's published content has become the focus of the current natural language processing research.

Social media text data contain many new words, abbreviations and emoticons and the length of text is relatively short [1,2] .This type of text data contains more noise,so the text matrix yield by topic model is extremely sparse and difficult to analysis and calculate.Commonly used probability topic model LDA and its variants [3] can display the semantic information of the texts through taking topic as the expression of the middle layer features.However, this model treats each word as an entity without considering the contextual relationships between words and the temporal dependencies.The machine-learning based approach is mainly used to process ordinary texts, using words as a basic attribute, represent texts as sets of words, and apply machine learning methods for matrix decomposition (LSI,NMF) [23] ,topic clustering(like k-means, incremental clustering, hierarchical clustering, single-pass clustering, etc.) [24-26] In addition,the development of deep learning provides a new research direction and technical method for the topic mining . Recently, word embedding and recurrent neural network (RNN) have become a research focus of deep learning in natural language processing.The word embedding method is to express the vocabulary as a dense real number vector on a low-dimensional space,in this way, the expression of lexical semantic features and the construction of language model can be realized.

However, the current study of hot topics mining in the OSN media mainly focus on the level of text, ignoring the interactivity of online social network:every text published by users hide a lot of social information.For example,the more the text was commented and forwarded ,the more probability that text contain the hot topic.Therefore, based on the above ideas, we propose a heat-based topic mining method.In this paper,the main works are as followed:

Given that the word2vec model can not distinguish the importance of words in the text,we give each word a value based on the tweets popularity and reconstruct word embedding model.

Using LSTM to encode documents and characterize the dependencies between words, and we apply single-pass clustering method to mine the topic model.

.

II. Related work

In this section, we review the prior study on the issue of hot topic mining. Social media text expression is not standardized, and there is a large number of short text messages, to solve such problems, Danushka and Liu [11,12] et.al proposed the use of search engines and hownet to expand the original text with the aim of weakening the low-frequency feature words on the clustering results.However, the methods are implemented by introducing a large number of external features,as a result the excessive time consumption is not suitable for large-scale expected research.

In terms of the expression of words, word embedding has become a very popular tool in recent years and is widely used in every aspects of natural language processing.Collobert & Weston et al. [27] proposed a multi-layer neural network model structure based on recurrent neural network for processing variable length word sequence input. In 2012, Eric H. Huang improved the method of C&W and proposed a word embedding training model based on cyclic neural network considering the local and global contexts [28] . In 2013, Google's Mikolov team proposed the word2vec word embedding model [6,7] , expressing the vocabularies as a dense real number vector on a low-dimensional space, in order to achieve a lexical semantic feature expression. In 2014, Jeffrey and Socher fused the idea of local context information and

DOI reference number: 10.18293/SEKE 2018-049

global word co-occurrence matrix decomposition to further explore a global linear regression model Glove [29] . Other methods to consider the global context include adding text vectors such as PV-DM and PV-DBOW [30] models and introducing external knowledge base, but due to the particularity of words in natural language processing in terms of semantic expression and multi-word synonymy and so on, how to obtain high-quality representation of word features has always been an important topic in natural language processing and text mining. Sundermeyer et. al [8] explained how to build language models using LSTM neural networks. Kim [9] used CNN to complete the task of sentence classification with the pre-process of word vector.In 2015, Tang et.al [10] used the neural network model consisting of CNN and LSTM to conduct text topic mining.After experimental verification, neural networks composed of CNN and LSTM have achieved good results in topic mining.

With the deep application of deep neural network in the field of natural language processing, Encoder-Decoder framework [16] has been the mainstream method of text sequence modeling in recent years, It also has far-reaching effects on reading comprehension [17] , text abstract [18] , machine translation [19] , automatic question answering [20] . Wang et al. Proposed an end-to-end deep learning framework for fusion question matching to model reading comprehension questions [21] .The framework includes match-LSTM, a match expression model for questions and sentences, and a web-oriented Point Net for answer constraints that effectively enable reading comprehension on large datasets. The popularity of the Encoder-Decoder framework has led to many enhancements to codecs, and attention models [22] are the most powerful and most powerful enhancements available today.The attention model decodes the Decoder output to give a different weight to each input to the Encoder so that the weight of the input that is more important to the current output becomes larger.

III. PROPOSED METHOD

A. Definition of Micro-Blog Heat

Our thinking about the way of calculating blog heat comes from the description of self-information in information theory:a small probability event contains a large amount of information, at the same time a large probability event contains a small amount of information.So the valve of the information in event A is defined as formula(1) :

$$I(A) = -\log P(A) \quad (1)$$

Learning from the idea of calculating the valve of information, assuming that the number of comments on a micro-blog m is r, the forwarding number is s, the definition of the micro-blog heat is shown in formula(2)

$$Heat\ (m) = -\log \frac{1}{r+s+1} \quad (2)$$

The heat of micro-blog is equal to the sum of the heat of the term in the micro-blog. Therefore, when there are N words in the micro-blog, the heat calculation formula of a single word is described as formula(3):

$$Heat\ (w) = \frac{Heat\ (m)}{N} \quad (3)$$

B. Word2Vec model

Word embedding is a good way to express lexical features, and the vocabularies are expressed as dense real number vectors in low-dimensional space.Word2Vec is the most widely used word embedding technology,including two kinds of models:CBOW and Skip-Gram.The CBOW model uses the remaining words in the context to predict the probability of generating the target word,and Skip-Gram uses the target vocabulary to predict the probability of other terms in the context.Compared with the CBOW model, Skip_gram has higher semantic accuracy at the expense of higher computational complexity.This paper is based on the Skip-Gram model to improve the training of word vectors, so the prediction of contextual probability is defined as (4):

$$p(context\ (w) \mid w) = \prod_{\overline{w} \in context\ (w)} p(\overline{w} \mid w) \quad (4)$$

Word2vec applies a layered softmax function to improve computational efficiency,combined with the layered softmax function, equation (4) can be expanded to formula(5):

$$p(context(w) \mid w) = \quad (5)$$

$$\prod_{\overline{w} \in context(w)} \prod_{j=2}^{l_{\overline{w}}} [\sigma(v_w^T x_{\overline{w},j-1})]^{y_{\overline{w},j}} [1-\sigma(v_w^T x_{\overline{w},j-1})]^{1-y_{\overline{w},j}}$$

Where w denotes the target vocabulary, $\overline{w}$ indicates the context of the target vocabulary, l_w denotes the the path length of the context in the output layer hierarchy tree, v_w denotes the input word vector of target vocabulary. $x_{\overline{w},j}$ represents the output word vector at the corresponding level under a certain contextual vocabulary. $y_{\overline{w},j}$ is logistic output variable.

$$p(y_{\overline{w},j} \mid v_m, x_{\overline{w},j-1}) = \sigma(v_w^T x_{\overline{w},j-1}) \quad y_{\overline{w},j} = 0$$
$$p(y_{\overline{w},j} \mid v_m, x_{\overline{w},j-1}) = 1 - \sigma(v_w^T x_{\overline{w},j-1}) \quad y_{\overline{w},j} = 1$$

Where $\sigma(\cdot)$ represents sigmod function.

When training the model of word embedding, the robustness of the model is often enhanced by introducing some words that are not found in the corpus as negative samples,these negative samples can be pre-generated by negative sampling [13] .The objective function of Skip-Gram model trained with negative sample is shown in formula(6):

$$L = \sum_{w \in C} \sum_{\bar{w} \in context} \sum_{(w)u \cup \{w\} \bigcup N_{\bar{w}}}$$

$$\{y_{wu} \log[\sigma(v_{\bar{w}}^{T} x_u)] + (1 - y_{wu}) \log[1 - \sigma(v_{\bar{w}}^{T} x_u)]\} \quad (6)$$

Where $N_{\bar{w}}$ is a negative sampling set under the vocabulary $\bar{w}$, y_{wu} is logistic output variable, $v_{\bar{w}}$ is the sum of word embedding of context, x_u denotes the parameter of the model.

C. Word2Vec based on micro-blog heat

Given training corpus dictionary $vocab = \{t_i \mid i \in 1...N\}$ and document $d_i = < w_1, w_2, ...w_j >$where N is the word vector dimension.We use Word2vec model to train the corpus to get t he word vector,accumulate the word vectors in the text d_i to g et the vector representation of the text d_i shown in formula(7)

$$R(d_i) = \sum_t word2vec(t) \quad where \quad t \in d_i \quad (7)$$

Next, we introduce the heat model to calculate the word weight in Word2vec model according to the heat of words,and the the weighted word vectors are accumulated to obtain a new vector representation of document d_i shown in formula (8)

$$weight\_R(d_i) = \sum_t word2vec(t) \times Heat_t(w) \quad (8)$$

D. Document Semantic Coding Based on LSTM

Automatic coding machine is a artificial neural network,using self-supervised learning to encode input samples ,in order to achieve the purpose of reducing the data sample dimension.It is mainly divided into two parts:Encoder and Decoder.The encoder mainly compresses the original data into the output code:

$$\phi \ : \ X \rightarrow Z$$

The decoder will restore the output code closely to the original data:

$$\varphi \ : Z \rightarrow X'$$

In order for the automatic coder to retain the primary information in the original sample, the optimization goal is set as formula(9):

$$\phi, \ \varphi = \arg \min_{\phi, \ \varphi} \| X - \varphi(\phi(X)) \|^2 \quad (9)$$

LSTM coding framework is divided into five layers,taking the word embedding expression of all the keywords in the document as input and the overall semantic embedding of the document as output. The details of these five levels are as follows:

Word embedding presentation layer:This layer is the input layer. Because some documents tend to have more lexicons and some unimportant words that have no direct effect on the expression of the document's theme,in this paper we only select the word embedding with larger heat valve as input,and the word embedding for all words is obtained from word2vec model above.

Bidirectional LSTM hidden layer:Contains two LSTM hidden layers, forward and backward layer,at the same time, each word embedding is connected to both the forward and backward LSTM hidden layer unit,these two hidden layers are connected to the same output.The input word at the moment t is embedded as E_t ,the output to the forward LSTM hidden layer cell is h_{t-1}^{f} ,output to the backward LSTM hidden layer cell is h_{t-1}^{b} for the last moment.The output of the forward and backward layers at the current moment are shown in formula(10,11):

$$h_t^f = H(E_t, h_{t-1}^f, c_{t-1}, b_{t-1}) \qquad (10)$$

$$h_t^b = H(E_t, h_{t+1}^b, c_{t-1}, b_{t-1}) \qquad (11)$$

Where $H(\cdot)$ denotes the function of LSTM hidden layer, c_{t-1} indicates the status value of the Cell unit at the last moment, b_{t-1} denotes the offset at the last moment.

Bidirectional LSTM output layer： Each output cell is connected to both the forward and backward LSTM hidden layer cells at same moment.

$$g_1 = \sigma(W_{hg}^{f} h_t^{f} + W_{hg}^{b} h_t^{b} + b_g) \quad (12)$$

Where W_{hg}^{f} and W_{hg}^{h} are respectively the connection weights between the forward, backward hidden layer and the bidirectional LSTM output layer, b_g denotes the offset.

Average pooling layer:We use average pooling to process the original eigenvalues and construct new features,as well as realizing the dimension reduction, enhancement and noise filtering of the original valid features.Through averaging all cell values over a certain range,local information can be taken into account,calculation is shown in formula(12):

$$pool(g)=\sum_{t=1}^{T}\frac{g_t}{T} \quad (12)$$

Where T is the length of the input word embedding sequence.

Semantic encoding output layer:The result of the average pooling layer can be calculated by activating the function to get the final semantic coding vector of the entire document

The dimension of the semantic code vector is consistent with the dimension of the input word embedding in order to facilitate similar calculation.

E. Hot topic clustering

After applying LSTM to semantic encoding, each document exists in the form of a vector in the hidden subject space,and its dimension is much lower than that of the feature space in vector space model.Therefore, we use Single-Pass clustering algorithm to conduct document clustering. After that we get the number of clusters K (that is, the number of topics), and the results of the division of each document on the K topics in the document set.

IV. Result evaluation

Data set: We used crawlers to crawl over 10,000 micro-blogs in 15 hot topics in 2016 in Sina Weibo, and selected 200 micro-blogs from each topic,recording the number of each comments and forwarding on each micro-blog.As a result we take a total of 3000 micro-blogs as the experimental data set.

The data set need to be pre-processed before model training,including stop words, high and low frequency words, illegal characters.The word with the frequency of less than 5 in the corpus is considered as the low frequency word, and the word with the frequency more than 20% of the total number of words is considered as high frequency word.

Evaluation Measurement

For the results of the hot topic mining, we use the purity and the normalized information (UMI) as the evaluation measurement, and these two evaluation methods are applicable to the text data with label.

Purity:Purity is used to measure the proportion of correctly clustered documents in the total document. The greater the purity is, the better effect the topic clustering yields.The purity value is calculated as formula(13):

$$purity=\frac{1}{D}\sum_{i=1}^{k}\max_{j}|p_i\cap c_j| \qquad (13)$$

Where D is the total number of micro-blogs, k is the number of topic clusters, p_i represents the set of words contained in the i-th cluster, c_j denotes the j-th micro-blog in the corpus.

normalized mutual information (UMI):The NMI value is calculated as formula(14):

$$NMI=\sum_{t=1}^{K}\sum_{1<ijK}\frac{\log\frac{P(w_i,w_j)}{P(w_i,w_j)}}{-\log P(w_i,w_j)} \qquad (14)$$

Where K represents the number of topics, N is the first N words under the topic. $P(w_i,w_j)$ denotes the co-occurrence probability of word w_i,w_j . $P(w_.)$ is the probability of the word w under the topic k.

Point-wise Mutual Information,PMI is the most commonly used semantic coherence evaluation measurement [14,15] .The results of PMI evaluation are often highly consistent with manual evaluation. The higher PMI scores, the stronger semantic coherence topic have.The PMI value is calculated as formula(15):

$$PMI(\varphi_k)=\frac{2}{V(V-1)}\sum_{1\le i<j\le V}\log\frac{p(w_i,w_j)}{P(w_i)p(w_j)}(15)$$

Where $p(w_i)$ is the probability of appearance of vocabulary w_i in the test document set. $p(w_i,w_j)$ represents the joint probability of vocabulary w_i and w_j in the test document set,V is the dimension of the vocabulary list. Therefore, in this paper, for each subject word distribution, we only select the first 10 words with the highest probability value to calculate PMI.

Result analysis and comparison

We also implement the classic statistical-based TF*IDF, NTM and LDA algorithm to exact key words from micro-blog.We conducted several sets of comparative experiments.In the experiment, we set the number of topics as 6, 20, 40, 80 respectively for the experiment of purity, and we set the number of topics as 20, 40, 60, 80, 100 respectively for experiment of NMI.The dimension of word vector size is 300, the experimental results are shown in Fig.1 and Fig.2

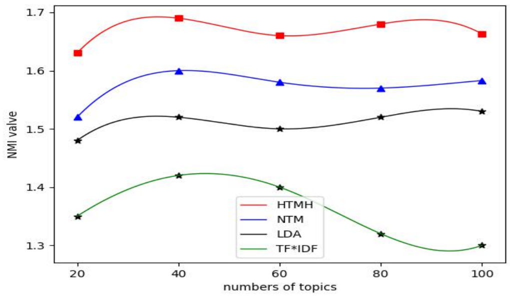

Fig.1 NMI valve

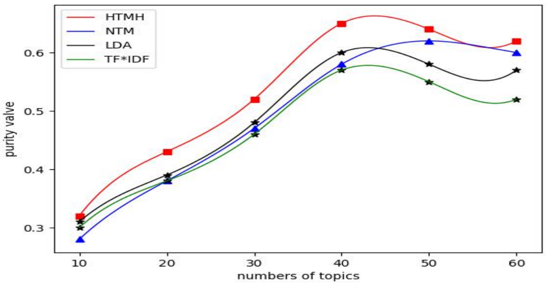

Fig.2 Purity valve

From the above two figures we can see,under the Sina Weibo corpus, the proposed HTMH(Hot Topic Ming Based on the Heat of Micro-blog) model has some improvement over other models in terms of purity and normalized mutual information,the purity reaches the maximum when the number of topics is 40,and the normalized mutual information (NMI) gets the highest accuracy when the number of topics is 20,This is also consistent with other models, HTMH has increased by 3% compared with TF * IDF, and increased by 5% compared with LDA.This is because word embedding technique is introduced as a semantic supplement.This makes the semantic relationship between words and subject strengthened.And in the follow-up training process, word embedding and topic model training promote each other, so we can identify the topic more accurately.

We can see from Fig.3, the HTMH model proposed in this paper has a generally higher PMI valve.It shows that the extracted hot topic has a strong semantic coherence and as the number of subjects increases, the PMI value basically remains unchanged.The traditional LDA model has the lowest PMI because it does not consider the semantic reinforcement of documents and words,NTM is a neural network reconstruction of the LDA model, but it does not take into account the reinforcement of vocabulary and semantics.So it performs relatively poorly.

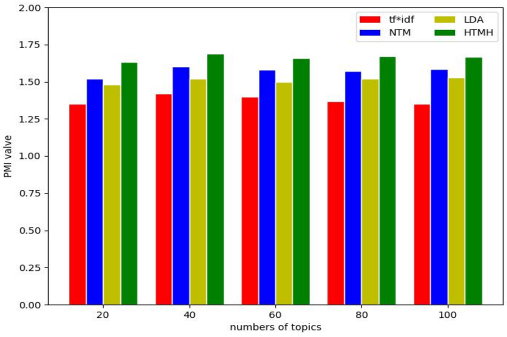

Fig.3 PMI valve

[16] Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks[J]. 2014, 4:3104-3112.

V. CONCLUSION

In this paper, we proposed a hot topic mining method based on micro-blog heat,based on the word2vec, we combined the idea of information theory and gave weight to each word. After using bidirectional lstm for semantic encoding, we applied single-pass clustering algorithm to conduct hot topic mining.

We then compared MTMH(Hot Topic Ming Based on the Heat of Micro-blog) model with TF * IDF, LDA and NTM algorithms,and introduced measurement of the topic mining quality,such as purity, NMI and PMI.Through the weibo corpus we proved that the HTMH model performs better on topic mining.

Future research focuses on the impact of sequence on the distribution of topics,in addition,collaborative training of feature expression of topic detection, emotion analysis and word embedding is also the trend of large-scale social media data analysis.

REFERENCES

[1] Losada D E. The challenge of understanding the flow of sentiments in social media documents[C]// International Workshop on Search and Mining User-Generated Contents. ACM, 2011:1-2.

[2] Liu H. Mining social media: issues and challenges[C]// ACM Sigmm International Workshop on Social Media. ACM, 2011:1-2.

[3] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:993-1022.

[4] Bengio Y, Vincent P, Janvin C. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2006, 3(6):1137-1155.

[5] Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks[J]. 2014, 4:3104-3112.

[6] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[7] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.

[8] Sundermeyer M, Schlüter R, Ney H. LSTM Neural Networks for Language Modeling[C]// Interspeech. 2012:601-608.

[9] Kim Y. Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014

[10] Tang D, Qin B, Liu T. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification[C]// Conference on Empirical Methods in Natural Language Processing. 2015:1422-1432

[11] Bollegala D, Ishizuka M, Matsuo Y. Measuring semantic similarity between words using web search engines[J]. Computer Science, 2015:757-766.

[12] Liu Z, Yu W, Chen W, et al. Short Text Feature Selection for Micro-Blog Mining[C]// International Conference on Computational Intelligence and Software Engineering. IEEE, 2010:1-4.

[13] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.

[14] Newman D, Lau J H, Grieser K, et al. Automatic evaluation of topic coherence[C]// Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA. DBLP, 2010:100-108.

[15] Jey Han Lau,David Newman,Timothy Baldwin.Machine reading tea leaves:Automatically evaluating topic coherence and topic model quality.In EACL'14,pp.530-539,2014

[17] Seo M, Kembhavi A, Farhadi A, et al. Bidirectional Attention Flow for Machine Comprehension[J]. 2016.

[18] Lopyrev K. Generating News Headlines with Recurrent Neural Networks[J]. Computer Science, 2015.

[19] Cho K, Merrienboer B V, Bahdanau D, et al. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches[J]. Computer Science, 2014.

[20] Pengcheng Yin, Zhengdong Lu, Hang Li, et al. Neural Enquirer: Learning to Query Tables with Natural Language[J]. Computer Science, 2016.

[21] Wang S, Jiang J. Machine Comprehension Using Match-LSTM and Answer Pointer[J]. 2016.

[22] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer Science, 2014.

[23] Wang Q, Xu J, Li H, et al. Regularized latent semantic indexing[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2011:685-694.

[24] Yamron J P, Knecht S, Mulbregt P V. Dragon's Tracking and Detection Systems for the TDT2000 Evaluation[J]. Proceedings of the Broadcast News Transcription & Understanding Workshop, 2000:75--79.

[25] Charikar M, Chekuri C, Motwani R. Incremental clustering and dynamic information retrieval[C]// Twenty-Ninth ACM Symposium on the Theory of Computing, El Paso, Texas, Usa, May. DBLP, 1997:626-635.

[26] Corpet F. Multiple sequence alignment with hierarchical clustering.[J]. 1988.

[27] Collobert R, Weston J, Karlen M, et al. Natural Language Processing (Almost) from Scratch[J]. Journal of Machine Learning Research, 2011, 12(1):2493-2537.

[28] Huang E H, Socher R, Manning C D, et al. Improving word representations via global context and multiple word prototypes[C]// Meeting of the Association for Computational Linguistics: Long Papers. Association for Computational Linguistics, 2012:873-882.

[29] Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation[C]// Conference on Empirical Methods in Natural Language Processing. 2014:1532-1543.

[30] Le Q V, Mikolov T. Distributed Representations of Sentences and Documents[J]. 2014, 4:II-1188.