Postprint: Research on 'Theme+Opinion' Term Extraction Algorithm for Weibo Topics

Authors: Yao Zhaoxu, Ma Jing

Date: 2017-11-08T00:00:00+00:00

Abstract

Purpose Automatically extract Weibo topic information and comprehensively reveal Weibo topic content and viewpoints from the two dimensions of topics and viewpoints. **Method** Apply topic models to Weibo topics, combine with an improved TF-IDF algorithm to construct topic feature word vectors; automatically extract topic lexical chains based on the correlation between feature words in the feature word vectors; introduce a sentiment dictionary to extract topic viewpoints, and unsupervised construct "topic+viewpoint" entries. **Results** Using a web crawler tool, 24,598 Weibo posts related to 4 specific popular Weibo topic events during June 2014–June 2015 were collected, "topic+viewpoint" entries were extracted, achieving an average accuracy of 80.3% and a recall of 76.7%. **Limitations** The dataset is still relatively small, and the effectiveness of topic models in feature extraction for Weibo short texts still needs improvement. **Conclusion** The algorithm proposed in this paper can accurately and effectively describe topic event content and its corresponding viewpoints.

Full Text

Preamble

Extracting "Topic + Opinion" Entries from Microblog Topics
Yao Zhaoxu, Ma Jing
(School of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211106)

Abstract

[**Objective**] This study aims to automatically extract microblog topic information, integrating both thematic and opinion dimensions to reveal the content and viewpoints of microblog topics. [**Methods**] We applied a topic model to

microblog topics and constructed a thematic feature word vector using an improved TF-IDF algorithm. Based on the correlation between feature words in the vector, we automatically extracted thematic lexical chains. By introducing a sentiment dictionary, we extracted topic opinions and constructed "topic + opinion" entries in an unsupervised manner. [**Results**] Using a crawler tool, we collected 24,598 microblog posts related to four specific trending topics between June 2014 and June 2015. The extracted "topic + opinion"entries achieved an average precision of 80.3% and a recall rate of 76.7%. [**Limitations**] The dataset remains relatively small, and the effectiveness of topic models for extracting features from short microblog texts still needs improvement. [**Conclusions**] The proposed algorithm can accurately and effectively describe the content of topic events and their corresponding viewpoints.

Keywords: Text mining; Keyword extraction; Topic model; Microblog topic
Classification Codes: TP391; G350

1. Introduction

With the popularization and development of the Internet, platforms such as blogs, microblogs, and social networks have become important sources of information for netizens. By June 2015, China's internet user population had reached 668 million, with 249 million microblog users, of which 64.6% had participated in discussions on trending topics. This demonstrates that the microblog community has become a crucial platform for public opinion dissemination, where microblog topics serve as important channels for users to obtain information and express viewpoints on events. However, due to the mixed quality of information in microblog topics and the inherent characteristics of short, loosely structured texts with severe data sparsity, there is an urgent need for an appropriate organizational model or framework to help users rapidly extract and express microblog topic information and multi-dimensionally present public opinion content when facing information overload.

The extraction and representation of topic information can be traced back to the topic detection phase of Topic Detection and Tracking (TDT), whose main task is to detect and organize topics to address information overload. Recent research on topic information extraction methods has primarily approached from two perspectives: data mining methods and NLP text mining methods. Data mining methods mainly extract information from structured and semi-structured data. Becker et al. utilized Twitter' s historical data over a period of time, obtained event clusters through clustering algorithms, extracted cluster features, and employed a support vector machine model to identify new texts online. Popescu et al. extracted product features from review information for specific product categories by calculating the pointwise mutual information (PMI) between nouns in reviews and product characteristic words, using Bayesian classification. NLP text mining methods discover new knowledge from unstructured open texts and convert it into understandable and useful information. Ritter et al. proposed an open-domain event extraction method tailored to Twitter' s characteristics,

using a latent variable model to discover important event categories.

However, traditional text extraction methods are not suitable for microblog topics due to their brief content, loose structure, and severe data sparsity compared to conventional texts. With the introduction of topic models, an increasing number of scholars have incorporated LDA models into research on topic extraction and representation to address these issues. LDA is a three-layer Bayesian probabilistic model that represents implicit topic information in topics through probability distributions of characteristic words. To further enhance model applicability and topic extraction effectiveness, researchers have introduced external factors such as sentiment, topic popularity, author information, and user relationships between microblogs into the traditional LDA model for short text research on microblogs, achieving good results. Current research on topic information extraction and representation based on topic models primarily focuses on topic tag extraction, topic threading, and topic evolution. Kou Wanqiu et al. proposed a topic tag extraction method based on seed words, re-ranked topic feature word weights, extracted seed words, adopted a Bootstrapping approach to generate key phrase collections, and finally generalized and selected topic tags to represent topic content. Ramage et al. utilized the Labeled LDA model to map blog content to four dimensions based on content characteristics in Twitter to extract tags reflecting topic information. Darling et al. proposed the PoSLDA model, extending LDA and HMMLDA by categorizing vocabulary in documents into three types (adjectives, verbs, and nouns) to represent objects, actions, and descriptive information involved in topics. Yan Zehua adjusted word weights, considered background words and N-gram phrases, and extracted news thread labels based on the LDA model. These studies extracted topic information from the perspective of topic content without considering the introduction of opinion dimensions to improve microblog topic information extraction and representation, thereby more comprehensively displaying topic information.

To further enhance the effectiveness of topic information extraction and representation, this paper designs a "topic + opinion" representation model for microblog topics and proposes an unsupervised "topic + opinion" entry extraction algorithm. Experimental results demonstrate that the proposed algorithm achieves good results across different microblog topics, reflecting thematic information and viewpoints in microblog topics from multiple dimensions.

2. "Topic + Opinion" Entry Model for Microblog Topics

The microblog semantics in topic posts can be divided into two categories: objective description information about topic events and subjective opinion information. Considering the inherent characteristics of microblog topics, this paper proposes a "topic + opinion" topic representation model for microblog topics. A "topic + opinion" entry consists of a thematic lexical chain and topic opinions. The thematic lexical chain characterizes the content information of each thematic event in microblog topics in the form of lexical chains, while topic opinions reflect users' viewpoint tendencies toward thematic events.

Definition 1: Thematic Lexical Chain $LexicalChain\{k\}$ consists of a set of representative words or phrases automatically constructed based on the correlation $cor(w_i, w_j)$ between words in the feature word set, used to characterize the content information of thematic events in microblog topics.

Definition 2: Topic Opinion $Viewpoint\{j\}$ is an opinion word representing the viewpoint information of theme z_i, reflecting netizens' opinions and attitudes toward thematic events.

Definition 3: "Topic + Opinion" Entry $Entry(n)$ reveals topic information from two dimensions—topic content and topic opinion—composed of a thematic lexical chain $LexicalChain\{k\}$ and topic opinion $Viewpoint\{j\}$. The model structure is as follows:

$$Entry(n) = \{LexicalChain\{k\}, Viewpoint\{j\}\}, \quad n = 1, 2, \ldots, K \tag{1}$$

where $LexicalChain\{k\}$ represents the thematic lexical chain of the n-th theme z_n, $Viewpoint\{j\}$ represents the opinion information corresponding to the thematic information, and K denotes the number of themes.

3. Unsupervised "Topic + Opinion" Entry Extraction Algorithm

Current common approaches for topic information extraction include supervised, semi-supervised, or unsupervised text mining methods. Supervised methods have only theoretical value because it is difficult to construct appropriate training sets to build classifiers in practical applications. Ontology, as an effective formal semantic model and knowledge representation form, has also been applied in topic extraction in recent years. However, constructing topic-related ontologies often adopts a semi-supervised approach requiring substantial domain information, resulting in low accuracy and mostly prototype systems that have not been deployed in practice. In contrast, unsupervised detection algorithms require fewer prior demands and possess stronger generalization capabilities, better aligning with the practical context of topic extraction.

This paper proposes an unsupervised microblog topic information extraction algorithm, which mainly consists of three steps: (1) Adjust feature word weights based on the differential representativeness of thematic feature words across different topics within a topic, constructing a thematic feature word vector; (2) On the basis of the feature word vector, unsupervised generation of thematic lexical chains according to correlations between feature words to represent thematic content information; (3) Introduce a sentiment dictionary to construct an opinion word set, and automatically extract topic opinions by combining the thematic lexical chains from step (2) with the opinion strength of opinion words.

3.1 Construction of Thematic Feature Word Vector Based on Improved TF-IDF Algorithm

In the LDA topic model, dimensionality reduction transforms topic information from a massive text space to a topic space, representing a theme (Topic) in a topic through a probability distribution of a set of vocabulary words—that is, describing a thematic event in a topic through a set of feature words. Assume the topic text collection is $D = \{d_1, d_2, \dots, d_M\}$, where M is the number of documents. After topic modeling, we obtain the topic-word probability distribution θ and document-topic probability distribution φ, where $p(w_j|z_k)$ represents the contribution of vocabulary w_j to theme z_k, i.e., the probability that w_j belongs to theme z_k.

The LDA model assumes equal weight for each vocabulary word, but in reality, each word' s representativeness varies across themes. Traditional calculation of word representativeness typically uses the TF-IDF algorithm, but TF-IDF cannot effectively identify high-frequency keywords nor filter uniformly distributed keywords. Drawing on the ideas from literature [12], this paper introduces coverage and characteristic features into the traditional TF-IDF algorithm to better distinguish between thematic feature words and background words.

Coverage indicates the degree to which a word covers the document collection. Words with high coverage are more representative in the corpus. Coverage is calculated as the number of documents N_i containing the word divided by the total number of documents N:

$$coverage = \frac{N_i}{N} \tag{2}$$

Characteristic reflects the degree to which a word' s text represents a particular theme. $p(z_i|d_i)$ is the topic-document probability distribution, representing the probability that text d_i containing word w_i belongs to theme z_i. The calculation formula is as follows:

$$characteristic = \frac{\sum_{d_i \in D} p(z_i|d_n)}{N_i} \tag{3}$$

By calculating word weights through the improved TF-IDF algorithm and sorting them in descending order, we select the top n feature words to form the thematic feature word vector. The adjusted feature word vector for theme z_n is represented as:

$$\{(w_1, weight_1), (w_2, weight_2), \dots, (w_n, weight_n)\}$$

A comparison between the basic LDA model and weight calculation results is shown in [Figure 1: see original paper]. The upper part shows the LDA topic

modeling results, while the lower part shows the thematic features after weight calculation.

3.2 Generation of Thematic Lexical Chain Based on Feature Word Vector

Words often describe topic information around specific themes. Such collections of semantically interrelated words centered on a theme are called lexical chains. This paper uses the correlation magnitude between feature words to reflect the strength of semantic associations between different words. Common correlation calculation formulas between words are as follows:

Combined with the improved TF-IDF algorithm, the expression is:

$$cor(w_i, w_j) = \frac{c(w_i, w_j)}{c(w_i) \cdot c(w_j)} \cdot \frac{|W|}{N} \cdot coverage \cdot characteristic \tag{4}$$

where $c(w_i, w_j)$ represents the number of times w_i and w_j appear in the same window, $c(w_i)$ and $c(w_j)$ are their respective word frequencies, $|W|$ represents the total number of document words, and N is the total number of documents. Since negative correlations may be obtained in calculations, and considering that $c(w_i)$ and $c(w_j)$ are typically small compared to N, the latter part of formula (4) can be ignored.

Traditional correlation calculations neglect the influence of the weights of w_i and w_j themselves. When high-weight feature words have high correlation, their composed phrases more easily reflect thematic information. Therefore, this paper improves the correlation calculation formula as shown in formula (5), introducing feature word weights $weight_i$ and $weight_j$ into the original correlation calculation:

$$cor(w_i, w_j) = \frac{c(w_i, w_j)}{c(w_i) \cdot c(w_j)} \cdot weight_i \cdot weight_j \tag{5}$$

In this paper, we count vocabulary co-occurrence within the same microblog post. If the correlation is positive, it indicates that the two words are related; the larger the positive value, the higher the correlation. When two high-weight feature words have high co-occurrence probability, the correlation between words increases, and their composed phrases can more accurately reflect topic semantic information.

Literature [19] suggests that the news domain typically uses the six elements of news to describe an event: What, Who, Where, When, Why, and How. Literature [20] proposes that opinion holders in comments are generally named entities and suggests using named entity recognition technology to obtain opinion holders. Drawing on these ideas, this paper considers that texts describing topic events typically contain noun words. Therefore, we select noun words w_i^n from

the feature word set as seed words to automatically generate thematic lexical chains.

The thematic lexical chain $LexicalChain\{k\}$ is generated based on the correlation between seed word w_i^n and other feature words in the feature word vector, as well as feature word weights. When the correlation $cor(w_i, w_j)$ between seed word w_i^n and feature word w_j exceeds a threshold, the seed word and feature word are combined into phrase P_t and added to the lexical chain candidate set P_i, with the phrase weight updated to the sum of the word weights. After iterative calculation, the phrase P_t with the maximum weight is selected as the thematic lexical chain:

$$LexicalChain\{k\} = \arg\max_{P_t \in P_i}(weight_t), \quad k = 0, 1, \ldots, K \tag{6}$$

The thematic lexical chain generation algorithm is as follows:

Input: Feature word set V_i, feature phrase set P_i
Output: Thematic lexical chain $LexicalChain\{k\}$

1. Set $P_i = V_i$;
2. For each w_j in V_i, calculate $weight_{i,j}$ and add all w_i to the list;
3. For each w_i in the list and each P_t in P_i, if $cor(w_i, w_j) \geq threshold$, calculate (w_i, w_j) as a phrase into P_i;
4. For each P_t in P_i, set the maximum weight P_t as $LexicalChain\{k_i\}$.

3.3 Topic Opinion Extraction Based on Sentiment Dictionary

Opinion extraction refers to the use of computer technology to automatically analyze sentences or documents containing opinion information on the Internet and extract the viewpoints or attitudes expressed by users. Viewpoint tendencies in topic texts are mainly conveyed through opinion words, which are mostly sentiment words. Opinion words manifest along two dimensions: viewpoint tendency (positive, negative, and neutral) and viewpoint strength.

Drawing on the Dalian University of Technology Sentiment Word Ontology [21], this paper constructs a sentiment dictionary. The sentiment word ontology is described through triples:

$$Lexicon(B, R, E)$$

where B represents basic word information, R represents synonym relationships between words, and E represents word sentiment information, describing from three dimensions: sentiment category, polarity, and strength. Sentiment strength is determined by pointwise mutual information (PMI) between candidate sentiment words and benchmark sentiment words in large-scale corpora, with strength divided into five levels: 1, 3, 5, 7, and 9. In microblog texts, an increasing number of users employ microblog emoticons instead of

text to express personal viewpoints. In our experimental corpus, 46.7% of texts contain microblog emoticons. Therefore, based on the Dalian University of Technology sentiment word ontology, we expanded the sentiment dictionary by adding commonly used microblog emoticons. Microblog emoticons are represented in the form `[emoticon content]`, such as `[applause]` or `[love you]`, and stored in the sentiment dictionary with their text content representing the emoticon semantics. After processing, the sentiment dictionary contains 28,466 words, including 16,074 positive words and 12,392 negative words, with sentiment strength also divided into five levels (1, 3, 5, 7, 9).

Assume the opinion word set is $SW = \{(sw_1, sw_{weight1}), (sw_2, sw_{weight2}), \dots, (sw_m, sw_{weightm})\}$, where sw represents an opinion word and sw_{weight} represents the corresponding opinion strength. However, the expression of topic opinions is related not only to opinion strength but also to the closeness between opinion words and thematic content. This paper uses the thematic lexical chain $LexicalChain\{k\}$ to represent the semantic information of thematic events. Therefore, the opinion extraction process is transformed into a correlation calculation process between opinion words and thematic lexical chains. We define the opinion value Q_i of a topic opinion as:

$$Q_i = sw_{weight} \cdot cor(sw, w_i), \quad sw \in SW, w_i \in LexicalChain\{k\} \tag{7}$$

For all annotated opinion words sw, if one of the following conditions is met, it is automatically extracted as a topic opinion $View\{j\}$:

1. If sentiment words exist in the thematic lexical chain, select the feature word with the maximum weight as the topic opinion:

   $$Viewpoint\{j\} = \arg\max\{weight_i\}, \quad sw_i \in LexicalChain\{k\}, i \in \{1, 2, \dots, n\}$$

2. If $sw \in SW$ and $sw_i \in LexicalChain\{k\}, i \in \{1, 2, \dots, n\}$, select the opinion word with the maximum opinion value in the opinion word set as the topic opinion:

   $$Viewpoint\{j\} = \arg\max\{Q_i\}$$

That is, in condition (1), if sentiment words exist in the thematic lexical chain, the feature word with the maximum weight is selected as the topic opinion; in condition (2), the opinion word with the maximum opinion value in the opinion word set is selected as the topic opinion.

4. Experiments

4.1 Experimental Setup

This study used a crawler tool to collect 24,598 microblog posts on trending topics from June 2014 to June 2015, including 9,230 posts on the "Chengdu

female driver beaten" topic, 6,932 posts on the "Nepal earthquake" topic, 4,367 posts on the "Yangtze River cruise ship sinking" topic, and 4,069 posts on the "Li Na gives birth" topic. In the preprocessing stage, we used the NLPIR2015 Chinese word segmentation system from the Institute of Computing Technology, Chinese Academy of Sciences, to segment and part-of-speech tag the microblog text. Stop words were removed according to the Harbin Institute of Technology stop word list, while microblog short links and low-frequency words were also eliminated, retaining nouns, verbs, and adjectives as candidate words.

The experimental parameters were set as follows: the number of topics $K = 50$, and the Gibbs sampling iteration count was 1,000. To analyze the impact of topic number settings on LDA topic modeling, we used the Perplexity metric to evaluate experimental results. Perplexity is a common metric for measuring topic model performance, indicating the uncertainty in predicting data; smaller values represent better performance. The calculation formula is:

$$Perplexity(W) = \exp\left(-\frac{\sum_{m=1}^{M} \ln p(w_m)}{\sum_{m=1}^{M} N_m}\right) \tag{8}$$

where W is the test set, w_m is the observable word in the test set, and N_m is the number of words. By gradually increasing the topic number K and calculating the perplexity of LDA topics under different values according to formula (8), the perplexity value continuously decreased as the topic number increased, as shown in [Figure 2: see original paper]. This study ultimately selected $K = 50$, and the experimental results are presented in .

4.2 Results Analysis

Microblog topic content often contains multiple different thematic contents, i.e., sub-topics. The experimental results show that the LDA model performs well in thematic mining, with high independence between themes and high summarization ability of thematic feature words, fully reflecting text content across different themes and effectively eliminating the impact of spam microblogs on topic events. For example, in the "Nepal earthquake" event, the model reflected thematic information such as "Chinese tourists returning home," "Tibet region disaster," and "rescue team rescue," effectively distinguishing different thematic information surrounding the microblog topic.

In feature word set construction, the improved TF-IDF algorithm enhanced the weights of topic feature words and reduced the weights of irrelevant background words, highlighting thematic features. For instance, in the "Li Na gives birth" topic, words like "China" and "champion" had low relevance to the topic. Our method reduced the impact of irrelevant words while increasing the weights of words like "Li Na" and "daughter," better reflecting topic content.

Opinion word extraction reflects users' viewpoints and attitudes toward topic content, showing attitudes toward different themes during event development.

For example, in the "Chengdu female driver beaten" event, the initial opinion on the female driver being beaten was "curse," expressing condemnation of the beating. As the event developed and it was revealed that the female driver used charity as an excuse for illegal lane changes, the opinion in this thematic event became `[cry with laughter]`, expressing irony and disbelief.

The "topic + opinion" entries in this study can effectively reflect topic information, basically covering thematic event content in topics and characterizing topics from both thematic content and opinion dimensions. For example, in the "Yangtze River cruise ship sinking" event, the automatically extracted entries "Yangtze cruise ship + safe" and "Yangtze cruise ship + capsized" had the same thematic lexical chain but belonged to different discussion contents within the microblog topic—the former praying for the cruise ship's safety, the latter describing the event.

4.3 Comparative Experiments

We compared our method with Sina Weibo topic tags and the method proposed in literature [12]. Sina Weibo topic tags are generally manually edited as summaries of microblog topic events. Literature [12] extracts seed words for each theme, iteratively generates key phrase collections, and finally generalizes and selects topic tags to describe topic information. The results are shown in .

The comparison results demonstrate that our method can accurately extract and express topic content and opinions. For example, in the "Nepal earthquake" topic, due to the sudden outbreak and relatively concentrated discussion, user opinions were basically consistent. Sina Weibo simply described it as "Nepal 8.1 magnitude earthquake," lacking multi-dimensional expression of thematic events within the topic. Literature [12]'s topic tags only described topic event content, such as "Nepal earthquake" and "Tibet earthquake," while our method extracted "Chinese tourists returning home + [heart]" and "Tibet disaster + donation," reflecting user opinions on corresponding themes while expressing thematic information.

Compared with literature [12]'s topic tag extraction method, our method reflects not only thematic content information but also corresponding opinion tendencies, facilitating users' understanding of the full picture of topics. For example, in the Nepal earthquake event, although both methods extracted "Nepal earthquake," our method reflected that most microblogs mentioning the earthquake were praying for the disaster area while expressing the event. Additionally, our method extracted the event of trapped Chinese tourists returning home, which literature [12]'s method failed to mine.

In some semantic expressions, our method is inferior to Sina Weibo topic tags. For example, Sina Weibo's topic tag "Chengdu female driver lane change beaten" is more complete and semantically fluent than our extracted entry "female driver lane change + beaten." Additionally, as users increasingly use emoticon icons instead of text to express opinions, microblogs containing emoticons often lack

obvious syntactic structures, affecting the interpretability of entries containing microblog emoticons.

We used precision P, recall R, and $F1$ to compare the extraction effects of literature [12] and our method. The calculation formulas are:

$$P = \frac{|correct \cap extract|}{|extract|}, \quad R = \frac{|correct \cap extract|}{|standard|} \tag{9}$$

where *correct* is the number of correctly extracted results, *extract* is the number of automatically extracted results, and *standard* is the total number of manually annotated entries. The results are shown in .

The results in show that our method achieves higher accuracy than literature [12]. In the "Chengdu female driver beaten" event, topic discussions mostly involved spontaneous participation from netizens, lasted a long time, and were accompanied by event evolution and development, with user emotions changing at different stages. Literature [12]'s topic tags, lacking opinion dimension expression and only describing topic content, therefore had lower precision and recall. In Li Na' s childbirth microblogs, a large number of irrelevant background words such as "Grand Slam" and "China tennis" interfered with topic information to some extent. Our method effectively reduced the impact of background words on event extraction, thus achieving higher precision. In the "Nepal earthquake" and "Yangtze River cruise ship sinking" events, most microblogs in the topics were news report-style microblogs with generic formats and high semantic similarity, resulting in similar precision for both methods.

Both our method and literature [12] use LDA model for topic modeling, but the modeling results contain some themes with unclear semantic expression and mixed spam microblog information, such as a large number of simple verbs like "go," "eat," "walk," and "love" that do not possess the ability to express topic semantics, affecting the reflection of main topic events.

5. Conclusion and Outlook

This paper proposes a "topic + opinion" entry model for microblog topics and its unsupervised extraction algorithm. The algorithm employs LDA modeling, constructs a feature word set after word weight calculation, automatically extracts thematic lexical chains based on correlations between feature words to represent thematic content information, introduces a sentiment dictionary to obtain topic opinions, and constructs "topic + opinion" entries from content and opinion dimensions to characterize microblog topic information. Experimental data validates the practicality of "topic + opinion" entries in topic information extraction and representation and the effectiveness of their unsupervised extraction algorithm. Future work includes further accurate extraction of topic opinions and improvement of topic models to enhance thematic extraction effectiveness.

References

[1] China Internet Network Information Center. The 36th Statistical Report on the Network Development of China Internet[R/OL]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/20150

[2] iResearch. The 2014 Research on China Weibo User Behavioral Report[R/OL]. http://www.iresearch.com.cn/report/2183.html.

[3] Hong Y, Zhang Y, Liu T, et al. Topic Detection and Tracking Review[J]. Journal of Chinese Information Processing, 2007, 21(6): 71-87.

[4] Becker H, Naaman M, Gravano L. Beyond Trending Topics: Real-World Event Identification on Twitter[C]. Proceedings of the 5th International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain. AAAI Press, 2011.

[5] Popescu A M, Etzioni O. Extracting Product Features and Opinions from Reviews[A]//Natural Language Processing and Text Mining[M]. Springer London, 2007.

[6] Ritter A, Mausam, Etzioni O, et al. Open Domain Event Extraction from Twitter[C]. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2012.

[7] Blei D M, Ng A Y, Jordan M I, et al. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[8] Lin C H, He Y L. Joint Sentiment/Topic Model for Sentiment Analysis[C]. Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM, 2009: 375-384.

[9] Tang X, Xiang K. Topic Mining Based on LDA Model and Popularity of Weibo[J]. Library and Information Service, 2014, 58(5): 58-63.

[10] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The Author-Topic Model for Authors and Documents[C]. Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. 2012.

[11] Zhang C, Sun J, Ding Y. Topic Mining for Microblog Based on MB-LDA Model[J]. Journal of Computer Research and Development, 2011, 48(10): 1795-1802.

[12] Kou W, Li F. Topic Label Extraction Based on Seed Word[J]. Journal of Chinese Information Processing, 2013, 27(5): 114-121.

[13] Qian Z, Li F. Keyword and Name Entity Identification Based News Topic Thread Extraction[J]. Computer Applications and Software, 2011, 28(12): 168-171.

[14] Hoffman M D, Blei D M, Bach F R. Online Learning for Latent Dirichlet Allocation[C]. Proceedings of the 24th Annual Conference on Neural Information Processing Systems. 2010.

[15] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora[C]. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore. 2009.

[16] Darling W, Song F. Probabilistic Topic and Syntax Modeling with Part-of-Speech LDA[OL]. arXiv: 1303.2826.

[17] Yan Z. News Threading Based on LDA Model[D]. Shanghai: Shanghai Jiaotong University, 2012.

[18] Wang Y. Research on Algorithm of Adaptive Chinese Topic Tracking Based on Ontology Evolution[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2013.

[19] Guo Y, Lv X, Li Z. Burstyn Topics Detection Approach on Chinese Microblog Based on Burst Words Clustering[J]. Journal of Computer Applications, 2014, 34(2): 486-490.

[20] Kim S M, Hovy E. Determining the Sentiment of Opinions[C]. Proceedings of the 20th International Conference on Computational Linguistics. 2004.

[21] Chen J. The Construction and Application of Chinese Emotion Word Ontology[D]. Dalian: Dalian University of Technology, 2008.

Author Contributions

Yao Zhaoxu: Proposed research ideas and design, conducted experiments, wrote and revised the paper.
Ma Jing: Collected data, expanded research ideas, reviewed and revised the paper.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data can be found in the online version of the journal at http://www.infotech.ac.cn:
[1] Yao Z, Ma J. lda.zip. LDA modeling JAVA program.
[2] Yao Z, Ma J. corpus.txt. Segmented dataset.
[3] Yao Z, Ma J. ldaresult.towords. LDA result data.
[4] Yao Z, Ma J. tfidfresult.xls. Feature word vector result data.
[5] Yao Z, Ma J. sentimentdictionary.sql. Sentiment dictionary.
[6] Yao Z, Ma J. finalresult.xls. Result data.

Received: 2016-01-28
Revised: 2016-05-23