View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by AIS Electronic Library (AISeL)

Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2017 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

Summer 7-19-2017

WSO-LDA: An Online "Sentiment + Topic" Weibo Topic Mining Algorithm

Jing Ma
Nanjing University of Aeronautics and Astronautics, majing5525@126.com

Zhaoxu Yao
Nanjing University of Aeronautics and Astronautics, 269708110@qq.com

Mingzhu Sun
Nanjing University of Aeronautics and Astronautics, mingzhu61@163.com

Follow this and additional works at: https://aisel.aisnet.org/pacis2017

Recommended Citation

Ma, Jing; Yao, Zhaoxu; and Sun, Mingzhu, "WSO-LDA: An Online "Sentiment + Topic" Weibo Topic Mining Algorithm" (2017). *PACIS 2017 Proceedings*. 223.
https://aisel.aisnet.org/pacis2017/223

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2017 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

WSO-LDA: An Online "Sentiment + Topic" Weibo Topic Mining Algorithm[*]

Type: Completed Research Paper

Jing Ma
Economics and Management College of Nanjing University of Aeronautics and Astronautics
Jiangsu Nanjing China
majing5525@126.com

Zhaoxu Yao
Economics and Management College of Nanjing University of Aeronautics and Astronautics
Jiangsu Nanjing China
269708110@qq.com

Mingzhu Sun
Economics and Management College of Nanjing University of Aeronautics and Astronautics
Jiangsu Nanjing China
mingzhu61@163.com

Abstract

In order to accurately excavate the micro-blog (Weibo) topic information and emotional information, we put forward Weibo Sentiment Online-LDA model on the basis of LDA. The model prejudges the emotional tendencies of the words in the text as a priori information of emotions and expands LDA model according to the emotional layer to get the topic information and the different emotional information of the topic. It also considers the influence of text information on the current time, dynamically adjusts the genetic coefficient of the topic, and ensures that the hot topic features are inherited to the next moment. The experiments show that WSO-LDA model mining matches the topic information and emotion information, and the model confusion degree is superior to other topic models.

Keywords: Topic Model, sentiment analysis, weibo, text mining

[*] This paper is the National Natural Science Fund Project " Research on Internet public opinion adaptive Topic Tracking Method basing on evolutionary Ontology" (project number: 71373123);Key Projects of Philosophy and Social Science Research in Jiangsu Colleges and Universities" Research on the Evolution Model and Application of Multi -Opinion of public in Jiangsu Education basing on hyper net"(Item number: 2015ZDIXM007);University Major Project Fostering Fund" complex social network behavior large data analysis method basing on" model - data dual drive " " (Item number: NP201630X).

Introduction

Micro-blog (Weibo) community has become an important platform for the dissemination of public opinion, and micro-blog topic has become an important channel for users to obtain the topic of event information and express relevant views. Text sentiment analysis, also known as opinion mining, is the process of analysis, processing, inducing and reasoning of the subjective text with emotional color. In recent years, sentiment analysis based on topic granularity has become a hotspot in the field of research. Researchers try to dig out the content of the discussion and corresponding emotional tendencies from the topic. In 2003, Beli [1] proposed Latent Dirichlet Allocation (LDA) model. LDA model as the bag of words model maps the high-dimensional textual information to the low-dimensional latent semantic space and uses the probability distribution of words to represent the topic information. LDA model is not only widely used in topic mining but also further extended to the sentiment analysis, and achieved good results. On the basis of LDA model, Yan Sun et al. [2] designed an unsupervised topic emotion mixture model. Different from the conventional LDA model, the model extracts the topic of the document and analyzes the emotion tendency of the sentence granularity. Finally, the experiment proves that the result is the best in unsupervised emotion classification. Jo [3] proposed ASUM which adds the emotional factors into the topic model, and he thinks topic information and emotional information influence each other. Po-Wei Liang et al. [4] used the Twitter API to collect the relevant text of mobile phones, cameras and movies on Twitter and marked them as positive, negative and neutral emotions. Based on a simple naive Bayesian model, they used mutual information and chi-square distribution to eliminate irrelevant characteristics, and ultimately predicted Twitter's emotional tendencies. Mei [5] proposed topic-sentiment mixture model (TSM), which added the emotional information of the document into the document generation process. TSM can simultaneously represent the topic and emotion implied in the document, where the emotional layer is the sub-layer of the topic, in order to extract the topic and emotional words. Based on LDA, Lin[6] proposed Joint Sentiment-Topic Model in 2010, which adds the emotion layer to LDA and extends it into a four-layer Bayesian network. Li [7] proposed sentiment-LDA, which is the same model as Reverse-JST, and proposed Dependency-Sentiment-LDA on its basis to describe the change of adjacent emotion in the document, and by identifying a priori of conjunctions to improve the accuracy of the word emotion sampling.

Researchers found that, view of emotional static access is not enough, and the timeliness is also an important factor. Some researchers have introduced the time factor into the sentiment analysis: Based on the social network, Daniel [8] constructed a public opinion model of the change of individual trust value. The experiment proved that the evolution of public opinion view was influenced by the individual with high self-confidence in the network. Considering the trust relationship between individuals to others, Qing Li et al. [9] built an Internet public opinion evolution model based on BA network. Weidong Huang[10] extracted the text content according to the emotional words at different times, took into account the adverbs and other emotional words on the modification and each emotional word in different moments of the contribution of the topic, and ultimately reflected emotional change of the sub-topic in different moments. Xianghua Fu et al [11] think that the user not only concerned about the topic of text content but also want to find the topic evolution model. On the basis of HDP model, adding time factors, they think that the topic content evolution is divided into intrinsic mode and mutual model. They achieved a dynamic online HDP model. Ping Lin et al. [12] used the PLSA model to extract the topic of network public opinion and its feature words and combined with TF-IDF to correct feature words. Based on the topic feature words that have been extracted, they construct the corresponding emotional vocabulary and apply How Net similarity algorithm to calculate the positive and negative emotional tendencies corresponding to each emotion word. They considered all the emotional words corresponding to the feature words. The emotional values of the feature words are calculated, and the emotional orientation and the change of the public opinion participants are accurately positioned. However, we found no research results relating to a combination of topic content and emotional timeliness.

Based on the above analysis: we propose parallel topic mining information and opinion information, and consider the real-time information and opinion information of micro-blog topic. Based on LDA model, we add the emotional factors and time factors, and put forward Weibo Sentiment Online LDA model. It can not only describe evolution and emotion but also describe the different emotional tendencies under the same topic, and realize the sentiment analysis of the topic of micro-blog.

Idea of WSO-LDA Topic Model

Introducing Emotional Factors to Improve LDA Topic Model

In this paper, when we introduce emotional factors to improve LDA topic model, it is suggested that we should first determine the emotional tendency of the words before determining the distribution of emotion.

Emotional word is the word that expresses the subjective emotion information in the text. Emotional words, as the most basic granularity in sentiment analysis, are usually selected as features to analyze emotional information of the topic. Emotional words are usually concerned with the emotional tendencies and emotional intensity, that is, the emotional meaning of the word is a commendatory or derogatory, and the degree of emotional tendencies. On the basis of the text pretreatment, we use so_pmi to determine the emotional tendencies of words. First of all, according to the emotional word ontology, K pairs of reference words are selected from the sentiment dictionary, and each pair of reference words contains a positive word and a negative word. Then, the correlation between the candidate word and the seed word is calculated. The semantic tendency of the word w is denoted by o. The default threshold is denoted by 0, the tendency is greater than 0 as the commendatory term, and the lesser than 0 is the derogatory term.

$$\text{so\_pmi}(w) = \sum_{pword \epsilon Pwords} PMI(w, pword) - \sum_{nword \in Nwords} PMI(w, nword) \qquad (1)$$

In this paper, the lexical semantic sentiment tendency O is determined by so_pmi. The so_pmi values of the word and the seed word are added together then a tendency of the emotion was judged according to the calculation result. If the polarity of emotion words is positive, χ will be χ_1; if the polarity of emotion words is negative, χ will be χ_2; if not, χ will be χ_3.

$$O_w = \begin{cases} \chi_1, so\_pmi(w) > 0 \\ \chi_2, so\_pmi(w) < 0 \\ \chi_3, so\_pmi(w) = 0 \end{cases} \qquad (2)$$

Introducing Time Factors to Improve LDA Topic Model

The traditional Online-LDA model [13] assumes that the "topic-word" distribution in a time slice was affected by the "topic-word" distribution from the previous time slices. Specific performance in the model has shown in Equation 3.

$$\beta_k^t = B_k^{t-1} \otimes \omega_k^{t-1} \tag{3}$$

That is, the priori parameter β^t of the "topic-word" distribution in the current time t can obtained by multiplying the evolution matrix B^t of the first δ time slices by the weight vector ω. Where the weight vector ω is a constant and its value can be pre-determined experimentally. Because the closer the document from the current time t has the greater impact on the current topic content, we assume that the document probability distribution at current time t is only affected by the previous moment, that is, the prior distribution of t only affected by the time t-1, independent of the earlier time slice. Based on the idea of Online-LDA model, the matrix B that formed by the topic-sentiment distribution ϕ which is the matrix of | V | * t. For each topic k, the topic emotional distribution ϕ is filled into the matrix as a column value to reflect the topic content and emotional change. ω_k^{t-1} is the genetic factor of the topic k at the time t, representing the priori information of the topic.

The genetic intensity of the topic is closely related to the importance of the topic. When a topic is more important than other topics, it has a greater impact on topic content in the next time, that is, when the distance between the topic and other topics is the smallest, you can think the topic z is at the core position at the current time t, and it can affect the next moment. The KL distance formula is usually used in LDA model. However, the KL distance formula is an asymmetric distance formula, so we use JS distance formula in this paper. The distance between topic and topic is the same.

$$D_{JS}(P,Q) = \frac{1}{2}\left[\sum_i D_{KL}\left(P(i), \frac{P(i),Q(i)}{2}\right) + D_{KL}\left(Q(i), \frac{P(i)+Q(i)}{2}\right)\right] \tag{4}$$

Thus, the problem of solving the importance is transformed into the problem of distance between topics. The importance of the topic is the shortest distance to other topics. When the sum of distances that a topic to the remaining topics is the minimum, the topic is the most important. When the topic distance is smaller and the weight ω^t of the topic is larger, the Equation 4 is derived.

$$d_{max}^t = min \sum D_{JS}(P(i), P_{z\neq i}) \tag{5}$$

Since the prior parameter is less than 1, it is necessary to normalize ω^t and to convert ω^t to the appropriate range. ω^t is the topic of importance λ^t divided by the sum of the importance of all topics at the current time t.

$$\lambda^t = \frac{1}{d^t} = \frac{1}{\sum D_{JS}(P(i), P_{z\neq i})} \tag{6}$$

According to the above formula, each topic at the different time in a sliding window appears different importance degrees. At a certain time the important degree is greater, and "genetic" ability of the word is stronger in this time; on the contrary, if the word of the importance degree is lower, the "genetic" ability is weaker in this time.

$$\omega^t = \frac{\lambda^t}{\sum_1^k \lambda^t} \tag{7}$$

Construction of WSO-LDA Topic Model

Description of WSO-LDA Topic Model

In this paper, we propose Weibo Sentiment Online LDA model, which is a dynamic topic emotion mixture model. WSO-LDA divides the text flow into fixed time slices, takes into account the influence of the text information on the current moment, provides the guidance for mining the topic at the current time, and excavates the micro-blog topic text information in different time slices; for the same time slice, it analyzes the emotional tendencies of each topic in the text, and simultaneously excavates the different emotional tendencies for each topic and presents it in the form of distribution. That is, WSO-LDA model finally excavates the evolution and development of the micro-blog topics at a different time. On the one hand, it integrates the emotional information, LDA from the traditional three-layer "document-topic-vocabulary" structure, expands to "document-topic-sentiment-vocabulary "four-layer structure. On the other hand, in the current time slice, this model adjust the intensity between topics so that strong topic in the time slice still can guarantee the content continue in the time before and after, and popular topic disappeared in the next time in order to observe different topics and the evolution and development of emotions. WSO-LDA model reduces the dimension of micro-blog topics from text dimension to topic dimension by model calculation, calculates lexical probability distribution and modifies prior parameters of by probability distribution of vocabulary which prepares for calculation model next moment.

Generation Process of WSO-LDA Topic Model

WSO-LDA model assumes that text order is sorted by time, divides the corpus into t time slices, records the text features using eigenvectors and finally detects the change of topic information by topic distribution. The distribution of topics in each time slice is affected by the content of the previous time slice. In the time slice t, there are M documents, denoted as $\mathrm{D} = \{d_1, d_2, \dots, d_M\}$. At the time t, the text corpus d consists of word vector $w_d^t = \{w_1^t, w_2^t, \dots, w_n^t\}$. WSO-LDA model assumes that all documents are generated by the same topic-sentiment distribution at the same time. Every micro-blog in the document contains l emotional tendencies. There are K topics, denoted as $\mathrm{T} = \{z_1, z_2, \dots, z_k\}$; L emotional labels are denoted as $\mathrm{S} = \{s_1, s_2, \dots, s_l\}$. The main symbols in this paper are described in the following table.

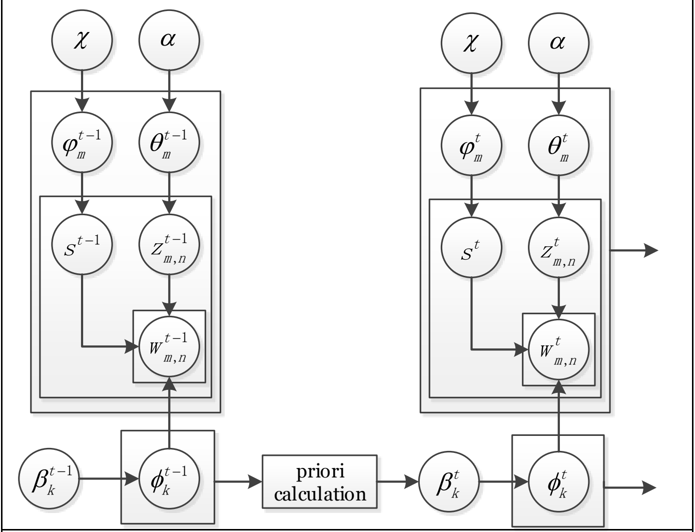

Figure 1. Bayesian Graph of WSO-LDA Model

In the evolution of the topic, the topic and emotion are continuity, not casually producing and disappearing, and will evolve to the next moment, so the content and emotion are affected by the previous moment. Thus, the topic and emotional probability distributions are affected by the prior probability distribution. In this paper, we use the evolution vector $E_{z,l}^t$ of the topic z and the emotion label l at time t to record the first l "document-sentiment-topic" distribution θ^{t-1}. Each column of the matrix represents the "document-topic" distribution ϕ^{t-1} at the time slice t-1. According to the weight distribution of different time slices, we calculate the priori parameters $\alpha^t, \beta^t, \chi^t$ of the model at the time t, get the "document-sentiment-topic" distribution at the current time, and then model the document of the current time slice. The Bayesian diagram of the model is shown in Figure 1.

SYMBOL	DESCRIPTION
M	total number of documents
K	number of topics
V	total number of unique words
N	total number of words
L	number of emotions
t	time t
θ_m^t	the "document-sentiment-topic" multinomial distribution of dimension $M^t \times K \times L$ at time t
ϕ_m^t	the "topic-sentiment-word" multinomial distribution of dimension $V^t \times K \times L$ at time t
φ_m^t	the "topic-sentiment" multinomial distribution of dimension $\mathrm{L} \times \mathrm{K}$ in the t-th time slice
$w_{m,n}^t$	the n-th word in document d_m at time t
$z_{m,n}^t$	the distribution of topic m and emotional label n at time t
α^t	the K dimensional priors of the "document-sentiment" distribution at time t
β_k^t	the priori parameters of the "topic-sentiment-word" polynomial distribution of topic k at time t
χ_k^t	the priori parameters of the "topic-sentiment " polynomial distribution of topic k at time t

Table 1. Notation used in the paper

The generation process of WSO-LDA model in the first t time slice can be expressed as follows:

(1) For each emotion label $\mathrm{l} = 1, \dots, \mathrm{L}$

A) For each topic $\mathrm{z} = 1, \dots, \mathrm{K}$

B) Compute $\beta_k^t = B_k^{t-1} \cdot \omega_k^{t-1}$

C) Draw distribution φ from a Dirichlet prior β, which is $\varphi_{l,z}^t \sim Dir(\beta_{l,z}^t)$

(2) For each document $\mathrm{d} = 1, \dots, D^t$

A) Draw the samples from the Dirichlet distribution of the parameter χ, which is $\varphi_d^t \sim Dir(\chi_d^t)$

B) For each emotion label in document d, draw $\theta_{d,l}^t$ from a Dirichlet prior α, which is $\theta_{d,l}^t \sim Dir(\alpha^t)$

C) For each word n in document d

i. Draw an emotion label $l_n \sim Mult(\chi_d^t)$

ii. Draw a topic $z_n \sim Mult(\theta_{d,l_n}^t)$

iii. Draw a word $w_n \sim Mult(\varphi_{z_n,l_n}^t)$

Parameter Derivation of WSO-LDA Topic Model

In this paper, we use Gibbs sampling to estimate the distribution of hidden variables θ, φ, ϕ in the model. Compared with the EM algorithm, the Gibbs sampling algorithm can converge quickly to ensure the stability of the results. Different from the traditional topic model, this paper introduces the emotional factors, so the topic-sentiment-word distribution needs to be estimated. At the time t-1, the weighting factor ω_k^{t-1} is adjusted dynamically according to the distance between topics, and further adjusted β^t to ensure that the hot topic can be better maintained to the time t. The joint probability distribution of the emotional distribution of topic z, emotion s, and word w at time t is as follows:

$$\mathrm{P}(W^t, S^t, Z^t) = P(W^t|S^t, Z^t, \beta^t) P(Z^t|S^t, \alpha) P(Z^t|\chi) \tag{8}$$

According to the Gibbs formula, WSO-LDA model of the Gibbs sampling formula can be written:

$$\mathrm{P}(Z_i^t = j, S_i^t = k \middle| W^t, Z_{-i}^t, S_{-i}^t, \alpha, \beta^t, \chi) \propto \frac{N_{j,k,-i}^t+\beta^t}{N_{j,k}^t+V\beta^t} \cdot \frac{N_{k,j,d,-i}^t+\alpha}{N_{k,d}^t+T\alpha} \cdot \frac{N_{k,d,-i}^t+\chi}{N_d^t+S\chi} \tag{9}$$

The approximate probabilities of φ , θ and ϕ are obtained:

$$\varphi_{i,j,k}^t = \frac{N_{i,j,k}^t+\beta^t}{N_{j,k}^t+V\beta^t} \quad \theta_{j,k,d}^t = \frac{N_{j,k,d}^t+\alpha}{N_{k,d}^t+T\alpha} \quad \phi_{k,d}^t = \frac{N_{k,d}^t+\chi}{N_d^t+S\chi} \tag{10}$$

Experiment Design and Result Analysis Based on WSO-LDA

Experimental Data Sources and Pretreatment

In this paper, we do experiments using some micro-blog text which include 11,184 micro-blogs of "Baoqiang Wang divorce", 9,230 micro-blogs of" Chengdu driver was beaten ", 6932 micro-blogs of "Nepal 7.5 earthquake",4367 micro-blogs of "A passenger ship sank in the Yangtze River " and 4069 micro-blogs of" Na Li gave birth to a daughter ". Considering that too short text contributes very little to the micro-blog topic mining, we select more than 10 words of micro-blogs to test, and finally get 27453 experimental corpus.

Text preprocessing for micro-blog, we use regular expressions to remove irrelevant text, such as, @ a micro-blog user, "second video" sign or "web link" sign and a large number of text that does not reflect semantic information. Then, we use word segmentation NLPIR of Chinese Academy of Sciences to introduce the user dictionary, segment the micro-blog text and tag part of the tag, and remove irrelevant words according to the disabled vocabularies to ensure that the experimental corpus is not disturbed by unrelated words.

Parameter Setting of WSO-LDA Topic Model

In this paper, we use the emotional seed word to judge the emotional information in different contexts, and to guide the choice of emotional prior information. We assume that the emotion is classified into two categories, that is, the positive emotion and negative emotion, and 15 pairs positive and negative emotion words are chosen as emotional seed words. As shown in Table 2 and Table 3. On the basis of the 30 emotional seed words, we can distinguish the emotion prior parameters by the correlation between the experimental corpus words and the emotional seed words.

love	praise	eulogize	compliment	adore

satisfaction	grateful	pleased	congratulations	care
good	joyous	admire	happy	gay

Table 2. Positive Emotional Seed Words

sad	sorry	grieved	sorrowful	contempt
suspicion	worry	anxious	diatribe	angry
anxiety	annoy	indignant	regret	despair

Table 3. Positive Emotional Seed Words

In order to validate WSO-LDA model, the experiment aims at the topic of "Baoqiang Wang divorce". In this experiment, we set the time slice t to one day and the parameter $\alpha = 50/K$ and the initial values $\beta = 0.01, \chi_{pos} = 5, \chi_{neg} = 0.1$. The number of iterations of the Gibbs sampling is 1000 times, where K is the number of topics. In order to analyze the influence of the number of topics on WSO-LDA topic modeling, perplexity was used to measure the experimental results. And the number of topics K is gradually increased to calculate the hash of WSO-LDA topic under the different topic number. With the increase of number of topics, the perplexity decreases continuously. The experimental results are shown in figure 2. In this paper, we set K=50.

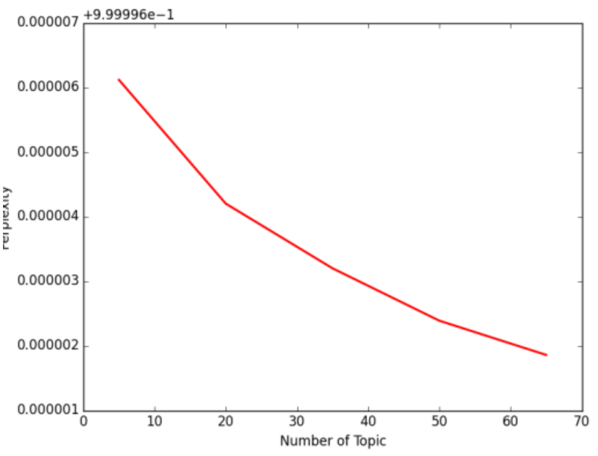

Figure 2. Perplexity Curve

WSO-LDA Mining Results Analysis

Based on Online-LDA, we consider that the topic and emotion of the current moment will continue to the next moment, and the content and emotion of the next moment will be affected by the corresponding topic of the current time, that is, the topic matter of the topic label in the previous time slice will affect the distribution of the topic content of the same topic label at the next moment. The results of WSO-LDA modeling at different time slices are shown in figure 3 and figure 4. Figures from two aspects of positive and negative emotions at different time slice, show the evolution in the topic, reflect the content and emotion as time progresses to produce change, which depicts the generation, development and decline of the event.

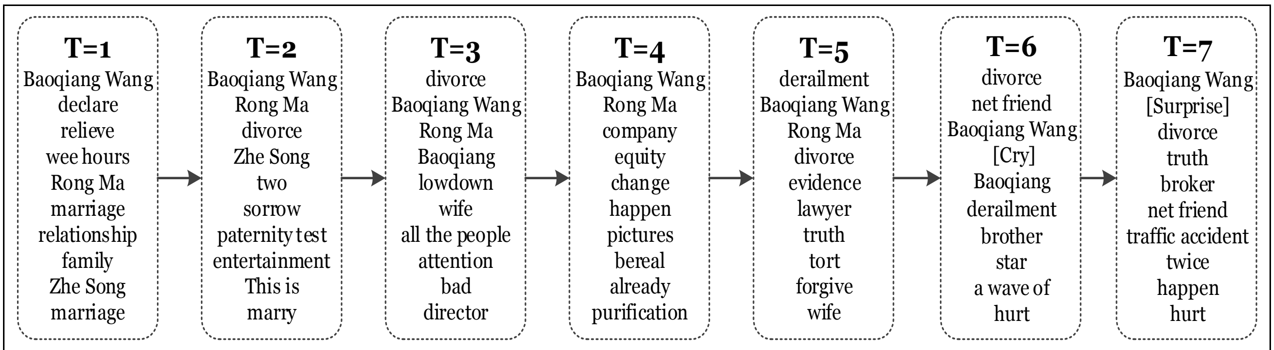

Figure 3. The Evolution Process of Negative Emotion Topic Probability Distribution

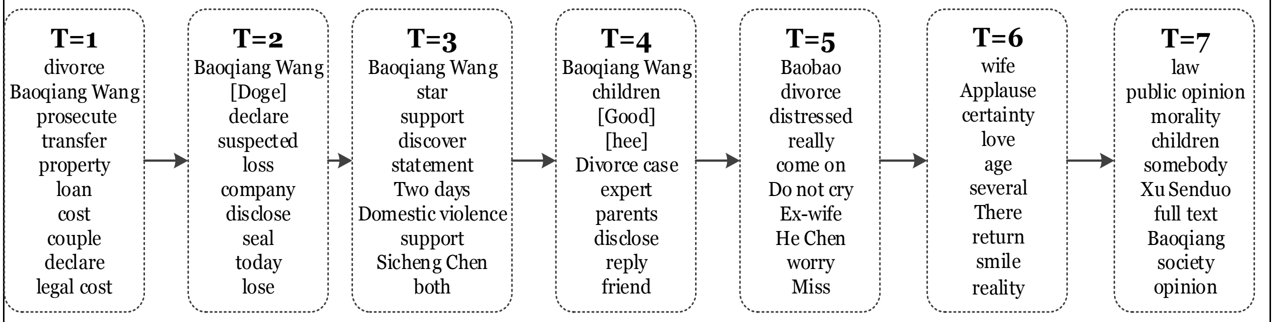

Figure 4. The Evolution Process of Positive Emotion Topic Probability Distribution

Table 4 shows the "topic-sentiment-word" distribution in the topic of "Baoqiang Wang divorce" at the same time slice. As the negative content is far more than the positive emotions, the topic of negative emotional labels is far more than the positive, and the word is also more closely related with a higher explanatory.

Topic 1		Topic 2		Topic 3	
Tag：Neg		Tag：Neg		Tag：Neg	
w	p	w	p	w	P
Baoqiang Wang	0.032248	Baoqiang	0.039010	Baobao	0.036568
real	0.021213	relieve	0.014576	distressed	0.014154
Baoqiang	0.018868	declare	0.011982	son	0.012667
two	0.016524	broker	0.011604	mother	0.011896
pitiful	0.014981	marital	0.011321	chat	0.011180
sorrowful	0.014752	Zhe Song	0.011227	support	0.010960
women	0.013322	marriage	0.010991	farmer	0.010850
daughter	0.013151	betray	0.010661	It's so......	0.010519
program	0.010864	wee hours	0.009859	popular	0.010464
all the way	0.009091	destroy	0.009859	picture	0.010134

Topic 1		Topic 2		Topic 3	
Tag：Pos		Tag：Pos		Tag：Pos	
w	p	w	p	w	P
divorce	0.051325	love	0.021709	movie	0.032403
Baoqiang Wang	0.027357	Baoqiang	0.021709	think of	0.028215

hahaha	0.022923	happiness	0.021294	lifetime	0.021649
entertainment	0.022527	influence	0.019978	the film	0.019342
sina	0.021085	estimate	0.017970	distressed	0.017994
reporter	0.019211	life	0.016239	like a play	0.016077
hahahaha	0.016364	somebody	0.014265	smile	0.014161
morning	0.014814	wechat	0.012569	Cannot help	0.012386
Wei Zhuo	0.014453	weekday	0.012188	midnight	0.010328
Rong Ma	0.014237	message	0.012153	remember	0.009512

Table 4. Experimental Results Show

WSO-LDA and JST, ASUM Contrast Experiment

In this paper, we introduce the emotion and time factors to improve the model. Then, we respectively compare WSO-LDA model with Online-LDA considering the time factor and JST and ASUM model considering the emotion factor to verify the validity of the model.

(1) Comparison with Online-LDA model

In order to verify the genetic and emotional characteristics of WSO-LDA model, the results of the original LDA model and Online-LDA model in different time slices are selected to compare with WSO-LDA model. The topic corpus of different topic events corresponding to the time slice experiments respectively, the mean value is calculated and the model result is synthetically evaluated.

In figure 5, the abscissa is the time slice, and the ordinate is the degree of confusion at each time. Series 1 is the original LDA without improvement, series 2 is Online-LDA model, and series 3 is WSO-LDA model. In the experiment, we can clearly see that the perplexity value of series 2 and series 3 are significantly smaller than the perplexity value of series 1. The perplexity value of series 3 is always less than the value of series 2, except that it is close in a certain time slice. The perplexity values of series 2 and series 3 are decreasing. These show that we can use WSO-LDA model to achieve good modeling effect in the topic of micro-blog, whether relating to genetic or emotion. In the topic of micro-blog, adding the time and topic factors, it can greatly enhance the ability of text modeling. Therefore, this experiment proves that WSO-LDA model and can significantly improve the effect of original LDA model and Online-LDA model.

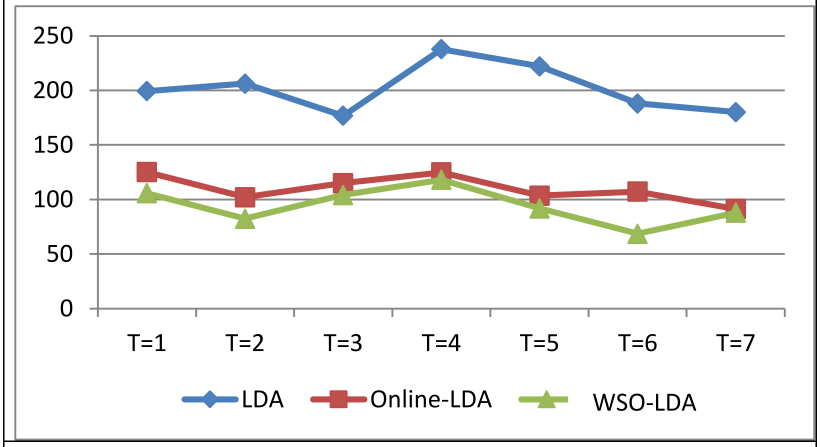

Figure 5. Perplexity Comparison

Compared with the original LDA model and Online-LDA model, it is proved that WSO-LDA model is more effective and modeling effect is more prominent. On this basis, we discuss the most suitable application field of WSO-LDA model in different types of micro-blog topic. For different types of micro-blog topics, the contents published by the users are different. In this paper, four topics, such as "Baoqiang Wang divorce", " Nepal earthquake ", " A passenger ship sank in the Yangtze River " and "Na Li gave birth to a daughter", were compared from entertainment, news, social and sports. In this experiment, the time slice is set to 1 day, and the change trend of perplexity for the four micro-blog topics in the different time was observed.

Figure 6 shows the change of perplexity value of each topic with time. It can be seen that the perplexity value of each topic decreases slowly with time, which indicates that the genetic characteristics of the topic have an impact on the next moment and reduce the confusion of the topic. At the same time, we can see that perplexity value of Nepal earthquake is obviously lower than that of other topics. Because the Nepal earthquake belongs to the news topic, the semantics is more normative, and the user's discussion is news-related information, event evolution is not obvious, the discussion is more focused, more closely linked to the topics. For the event of "Baoqiang Wang divorce", because of the sudden outbreak of the event, the user participation is wide, so the topic content is more discrete. With the development of the event, in each time slice evolve new topic events, the perplexity value will be significantly higher. In the event of “Na Li gave birth to a daughter”, because of the low content of the topic and the low participation of the users, the topic and the content of the discussion are different, so the Perplexity value is larger.

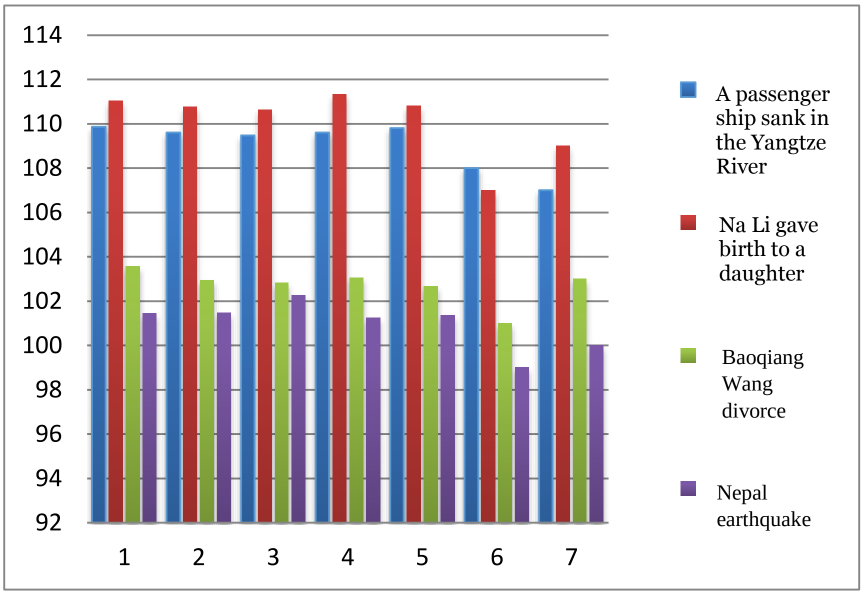

Figure 6. Perplexity Value Comparison of Different Topics

(2) Comparison with the topic model considering emotional factors

On the basis of traditional LDA model, ASUM and JST are put forward by the assumption of different emotional information. We use precision P, recall R and F1 (Precision / Recall / F1-Measure) to compare ASUM’s accuracy of sentiment classification and JST’s. $P = C_{correct}/C_{extract}$, $R = C_{correct}/C_{standtard}$, $F1 = 2PR/(P + R)$. Among them, $C_{correct}$ is number of correctly extracted results, $C_{extract}$ is number of all extracted results, and $C_{standtard}$ is number of all manually marked.

Event Name	Precision			Recall		
	WSO-LDA	JST	ASUM	WSO-LDA	JST	ASUM
Na Li gave birth to a daughter	68.7%	64.6%	60.2%	63.3%	59.6%	56.2%
Chengdu driver was beaten	65.3%	58.7%	49.8%	61.7%	56.2%	50.6%
Nepal earthquake	75.1%	74.1%	60.4%	68.3%	66.4%	58.3%
A passenger ship sank in the Yangtze River	70.3%	68.9%	63.2%	66.7%	67.8%	61.9%
Baoqiang Wang divorce	72.3%	67.5%	63.2%	70.2%	65.4%	60.7%

Table 5. Test Precision Evaluation Results

It can be seen from table 5 that the precision of sentiment classification of WSO-LDA is higher than that of JST and ASUM, which proves that the model can correctly judge the emotion information in the text. Compared to JST and ASUM, our model can accurately distinguish the emotional tendencies and positive and negative emotional information in different micro-blog topics.

For the event of "Chengdu driver was beaten", with the evolution and development of it, the emotion of users is change at different stages. The emotional factors and the impact of the previous time are considered in this model, so precision and recall of WSO-LDA is higher than JST's and ASUM's. There is a large number of "Grand Slam" and "Chinese Tennis" and other unrelated background words in the micro-blog of " Na Li gave birth to a daughter ", which interferes with topic information to a certain extent. This model effectively reduces the impact of background words on event extraction, so the precision is higher than the other two models'. In the event of "Nepal earthquake" and "A passenger ship sank in the Yangtze River ", because these topics are news reports and have universal format and high content of semantic similarity, the precision of each model is similar. In the topic of "Baoqiang Wang divorce", it contains a lot of negative emotional words, so the emotion of the topic tends to negative emotion, resulting in higher classification precision.

Conclusion

In this paper, we consider the characteristics of micro-blog topics, add emotional factors to the traditional topic model, take into account the temporal features, analyze the similarity between the topics under each time slice and dynamically determine genetic factors. At the same time, the differences of emotion labels between the words in different contexts are considered, and the close relationship between vocabulary and emotional seed words is determined. Experiments show that WSO-LDA model excavates topic information and emotional information to match each other. Its precision of emotion classification is 10% higher than that of JST, and 7% higher than that of AUSM. Its confusion is lower than that of Online-LDA, so it can better extract topic information.

Of course, WSO-LDA only takes into account influence of previous time slice on the distribution of topics at the next moment, but in reality the text at all times will have an impact on current text before development of topic evolution. The evolution of a topic is closely related with the previous text data. In next study, we will consider a wide range of time windows and impact of the previous text on the current. We only use the topic similarity calculation method to ignore the topic of the spam comments. However, in order to further enhance the effect of comment space mining, the research needs to pre-filter the spam comments before modeling the topic.

References

[1]Blei D M, Ng A Y, and Jordan M I. 2003. "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3:2003.

[2]Yan Sun, Xueguang Zhou, and Wei Fu. 2013. "Unsupervised Topic and Sentiment Unification Model for Sentiment Analysis,"*Acta Scientiarum Naturalium Universitatis Pekinensis* (49:1), pp. 102-108.

[3]Jo Y and Oh A. 2011. "Aspect and sentiment unification mode for online review analysis," *in Proceedings of the 4th ACM International conference on Web search and data mining*. New York: ACM, pp. 815–824.

[4]Liang P W and Dai B R. 2013. "Opinion Mining on Social Media Data," *in 2013 IEEE 14th International Conference on Mobile Data Management*, IEEE, pp.91-96.

[5] Qiaozhu Mei, et al. 2007. Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs[C]. Alberta, Canada.

[6] Chenghua Lin, Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. CKIM, pp. 375-384.

[7] Fangtao Li, et al. 2010. "Sentiment analysis with global topics and local dependency," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence,* pp.1371-1376.

[8]SIEWIOREK Daniel P., Yang Xiaozong, CHILLAREGE Ram, and KALBARCZYK Zbigniew T.. 2007. "Industry Trends and Research in Dependable Computing," *Chinese Journal of Computers*, pp.1645-1661.

[9]Qing Li and Huangmin Zhu. 2012."Study on the Evolution Model of the Online Public Opinion Viewpoint Based on BA Network," *Journal of Intelligence*, pp.6-9+35.

[10]Weidong Huang, Ping Lin, Yi Dong, and Hongwei Li. 2015. "Analysis on the Feature Words Based Evolution of Netizens Sentiments in Network Public Topics," *Journal of Intelligence,* pp.117-122.

[11] Fu X, Li J, Yang K, et al. 2015. Dynamic Online HDP model for discovering evolutionary topics from Chinese social texts. *Neurocomputing*, pp.412-424.

[12]Ping Lin, and Weidong Huang. 2013. "Event Topic Evolution of Network Public Opinions: An Analysis Based on LDA Model," *Journal of Intelligence,* pp.26-30.

[13]Jianyun He, Xingshu Chen, min Du, and Hao Jiang. 2015. "Topic evolution analysis based on improved online LDA model," *Journal of Central South University (Science and Technology)*, pp.547-553.