Author-Topic-Sentiment Mixture(ATSM) model for Author's Sentiment Analysis

KeHua Yang
College Of Information Science and Engineering
Hunan University
ChangSha, China
khyang@hnu.edu.cn

Xiang Yang
College Of Information Science and Engineering
Hunan University
ChangSha, China
jlf997@163.com

***Abstract*—In this paper,we propose a probabilistic modeling framework,called Author-Topic-Sentiment Mixture(ATSM) model,which based on Latent Dirichlet Allocation (LDA) to include authorship information and sentiments information.The proposed model can reveal the sentiment-topic and author's sentiment.Each author associated with a distribution of the sentiment-topics,and each sentiment-topic is associated with a distribution of the words.Unlike other approaches to sentiment classification which often require labeled corpora or sentiment seed words,the proposed ATSM model is unsupervised .We show sentiment-topics recovered and the author's distribution of sentiment-topic by the ATSM model.We compare the performance with two other generative models for documents :LDA and ATM,and illustrative a possible application of the ATSM.**

Keywords-LDA; author-topic; ATSM; Sentiment analysis; probabilistic topic models;Gibbs sampling; LDA

I. INTRODUCTION

With the advent of WEB 2.0 era, more and more people express their views and opinions through the network, such as weblogs,news comments and product reviews from shopping websites. As the amount of text message grows rapidly, the demand for efficient text analysis and sentiment analysis is urgent. In fact, we are concerned not only with the sentiment of the text itself, but with the sentiment expressed by the author. By analyzing the feelings of author on each topic, we can know the topic that the author is interested in and their attitude to it. The text we analyze varies from books,blog,weblogs and web comments.

Probabilistic topic model , a text modeling approach to find topic structure in large document sets. Latent Dirichlet Allocation Model (LDA)[1] is a quite simple model in Probabilistic topic models. LDA has a solid foundation in mathematics and its extensibility is good. Many variations of LDA are applied to the field of text analysis. Nowadays, in the field of text sentiment analysis, there are more studies on Probabilistic topic model, a type of text presentation model.

The paper combines an extended model of LDA, author topic model, and finds out not only the attitudes the authors on each topic ,but the relations between the sentiment of each word and its topic. In addition, we define a formula to calculate the distance between authors, thus we can further make a clustering analysis to authors according to the formula.

II. RELATED WORK

The traditional text sentiment analysis method is based on rules, this method pay attention to the semantic and contextual relevance.But because of the complexity of this method, in recent years there are few studies on it.There is more research methods are based on machine learning methods.Machine learning methods include:the semi-supervised, the unsupervised and the supervised.Rule-based method and the method based on supervised or a semi-supervised machine learning classifier training needs in a certain number emotional tags of the training sample, but artificial marking process relatively time-consuming, and unsupervised machine learning training samples without annotation.In the Web 2.0 era, new Internet words are increasingly emerges and people chang their ways to express feelings frequently. Under this circumstances,unsupervised machine learning method has a better application prospect.

In the text analysis method based on machine learning, text representation model is one of the important research contents.Probabilistic Topic Models is a kind of common text representation model.LDA is one of the most classic probabilistic topic models.The document is the probability distribution of topics in LDA,and the topic is the probability distribution of word. The Author-Topic Model (ATM)[2] is an extension of the LDA Model.As in LDA,the document is the probability distribution of topics as well. Considering the relationship between the author and topic,ATM express both authors and documents using topics.

In recent years, there are many research that extend probability models with sentiment, getting a mixture model of sentiment and topic.In TSM[3] model ,words can be divided into common words and words related to a topical theme.They further categorize words related to a topical theme into three sub-categories:words about the topic related to neutral opinions;words about the topic related to the positive opinions;and words about the topic related to the negative opinions .But the TSM does not model sentiment directly. The JST[4] don't need to calculate the sentiment coverage in a document in order to identify its polarity.In JST,every word is generated from one topic and one sentiment.The ASUM[5] assumed that every word in same sentence has the same sentiment and topic.The assumption ignores the connection between words.UTSU[6] assumed that all words in a sentence are generated from one sentiment and each word is generated from one topic. The assumption conforms to the sentiment expression of language and will not limit the topic relation of words[5].

The model proposed in this paper is the author-sentiment-topic mixture model.The Author-Topic-Sentiment Mixture(ATSM) can be seen as the ATM model joined the sentiment factors. We assume that all words in a sentence have the same sentiment and each word is generated from one topic. We can discover the sentiment-topic word and the author's sentiments of all topics.

III. AUTHOR-TOPIC-SENTIMENT MIXTURE MODEL

The author-topic model(ATM) is a generative model for documents that extends Latent Dirichlet Allocation.The ATM is based upon the idea that a document can be expressed as a mixture of topics, where each topic is a probability distribution of words.The author associated with the probability distribution of the topics.In the ATM model,a group of authors decide to write the document. For each word in the document , choosing an author at random. Then choosing a topic from a distribution of topics specific to the chosen author. Finally,generating the word according to the chosen topic.

The ATSM model is a generative model for documents that extends ATM.The ATSM can reveal the latent topic and the author's sentiment to it.The framework of ATM has three hierarchical layers, where authors are associated with documents,topics are associated with authors ,and words are associated with topics.In order to model sentiments, we propose the model by adding an additional sentiment layer.We imposes a constraint on ATSM that all words in a sentence are generated from one sentiment and each word is generated from one topic. The topic is a probability distribution of words.We denote the topic associated with sentiment by sentiment-topic.Each sentiment-topic is a probability distribution of words.The document can be expressed as a mixture of the sentiment-topic. The author is associated with a probability distribution of sentiment-topic.

A corpus with a collection of M documents denoted by $C = \{d_1, d_2, ..., d_M\}$;A document di is a vector of S sentences denoted by $D = (s_1, s_2, ..., s_s)$; each sentences in the document is a vector of N words denoted by $S=(w_1,w_2,...,w_N)$;each words chosen from a vocabulary of size V denoted by {1, 2, ..., V }; a vector of authors ad is a sequence of authors of size A. Also, let L be the number of sentiment , and T be the number of topics.

A. *The generative process*

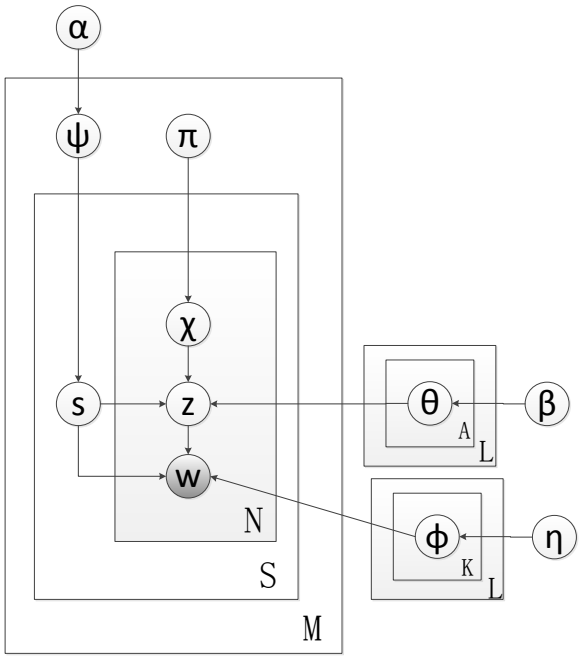

Figure 1. The hierarchical Bayesian model of ATSM.

In this model,each author is associated with a distribution of sentiment-topic ,denoted by ϕ , chosen from a symmetric Dirichlet(η) prior.Each word is associated with a distribution of sentiment-topic,denoted by θ, chosen from a symmetric Dirichlet(β) prior.For every document d,one chose a sentiment distribution $\psi_d \sim Dir(\alpha)$.Assume that a group of authors decide to write some documents.Firstly,for each sentence in the document,they choose a sentiment that they will express in this sentence. Then ,for each word in the sentence,choosing an author to write this word. According to the chosen sentiment, the author decide what topic he want to describe.Finally the author write down the word according to both the topic and the sentiment.

Name: Generative model for ATSM

1.For all sentiment s $\in$[1,...,L] ,topic z $\in$[1,...K] and word w$\in$[1,...,V] ,do sample mixture components $\theta_{s,z,w} \sim Diri(\beta)$,draw a author distribution $\phi \sim Dir(\eta)$

2. For all sentiment s $\in$[1,...,L] ,topic z $\in$[1,...K] and author $\chi \in$[1,...,A] ,do sample mixture components $\phi_{s,z,\chi} \sim Dir(\eta)$

3.For each document d $\in$ [1,...,M],

Draw the document's sentiment distribution $\psi_d \sim Dir(\alpha)$;

For every sentences s $\in$ [1,...,S], choosing a sentiment l~Multinomial (ψ);

For every word w$\in$ [1,...,N],choosing a author $\chi \sim$ Multinomial (π), choosing a topic z~Multinomial (θ_s, χ), chose a word~Multinomial (ϕ_s, z);

Figure 2. The pseudo code for the generative process of ATSM.

B. Parameter estimation

In ATSM,we have three sets of unknown parameter to infer,including: the sentiment-topic-author distribution θ, the sentiment-topic-word distribution ϕ, and the sentiment-document distribution ψ.

The latent variables are the assignments of words to topics z ,sentiment l and author x.

We apply Gibbs sampling[9,11] to estimate the parameters , we construct a Markov chain that converges to the posterior distribution on the assignment of ith word to topics ,author and sentiment labels. Then use the results to infer θ, ϕ and ψ.The sampling distribution for ith word given the topic ,sentiment label and author is:

$$p(z_i = k, s_i = \mathrm{l}, \chi_i = t / z_{-i}, s_{-i}, \chi_{-i}, w_i = \mathrm{n}, \pi) = \frac{p(z, s, \chi, w / \pi)}{p(z_{-i}, s_{-i}, \chi_{-i}, w / \pi)} \quad (1)$$

where $z_i = k$ represents the assignments of the ith word in a document to topic k ,s_i= l represents the assignments of the ith word to sentiment l,χ_i=t represents the assignments of the ith word to author χ and w_i = n represents the observation that the ith word in documents is the nth word in the vocabulary, and X−i (X could be the s,z or χ)represents the assignments of X for all the words except for the ith word.

The joint probability of the assignments of words to a topic-sentiment-author label is:

$$\begin{aligned} & p(W, Z, S, X / \alpha, \beta, \eta, \pi) \\ & = p(W / Z, S, \eta) p(Z / S, X, \beta) p(S / \alpha) p(X / \pi) \\ & = \prod_{k=1}^{K} \prod_{l=1}^{L} p(\phi_{k,l} / \alpha) \prod_{a=1}^{A} \prod_{l=1}^{L} p(\theta_{k,l} / \eta) \\ & \prod_{d=1}^{D} p(\psi_d / \beta) \prod_{l=1}^{L} p(s_{d,l} / \psi_d) \\ & \prod_{n=1}^{N} p(x_{d,l,n} / \pi_d) p(z_{d,l,n} / \theta_{x, s_{d,l}}) p(w_{d,l,n} / \phi_{z_{d,l,n}, s_{d,l}}) \end{aligned} \quad (2)$$

Integrating out ϕ, we obtain:

$$p(\vec{w} / \vec{z}, \vec{s}, \vec{\eta}) = \frac{B(\vec{n}_k + \vec{\eta})}{\mathrm{B}(\vec{\eta})}; \vec{n_k} = (n_k^{(1,1)}, .., n_k^{(S,Z)}) \quad (3),$$

Integrating out θ, we obtain:

$$p(z / s, \chi) = \frac{B(\vec{n_\chi} + \vec{\beta})}{B(\vec{\beta})}; \vec{n_\chi} = (n_\chi^{(1,1)}, ..., n_\chi^{(S,Z)}) \quad (4)$$

Integrating out ψ, we obtain:

$$p(s / \alpha) = \frac{\mathrm{B}(\vec{n_s} + \vec{\alpha})}{\mathrm{B}(\vec{\alpha})}; \vec{n_s} = (n_1^{(\mathrm{d})}, ..., n_{\mathrm{S}}^{(\mathrm{d})}) \quad (5)$$

where $n_s^{(d)}$ refers to the number of sentences that have been assigned to sentiment i in document d,the notation $\mathrm{n}_{\mathrm{k}}^{(i,j)}$ refers to the number of times that word k has been assigned to topic j and sentiment i, $n_\chi^{(\mathrm{i,j})}$ refers to the number of times that author χ has been assigned to topic j and sentiment i,V refers to the size of the vocabulary,S refers to the number of sentiment labels,.Z refers to the number of topics..B is the beta function.

Using the equation (3)(4)(5),equation (2) become:

$$\begin{aligned} & p(z_i = k, s_i = \mathrm{l}, \chi_i = t / z_{-i}, s_{-i}, \chi_{-i}, w, \pi) \\ & = \frac{p(z, s, \chi, w / \pi)}{p(z_{-i}, s_{-i}, \chi_{-i}, w / \pi)} \\ & = \frac{p(w / z, s, x, \pi)}{p(w_{\mathrm{i}}) p(w_{-i} / z_{-i}, s_{-i}, \pi)} \frac{p(z, s, \chi / \pi)}{p(z_{-i}, s_{-i}, \chi_{-i} / \pi)} \\ & \propto \frac{\mathrm{B}(\overrightarrow{n_k} + \vec{\eta})}{\mathrm{B}(\overrightarrow{n_{k,-i}} + \vec{\eta})} \frac{B(\overrightarrow{n_a} + \vec{\beta})}{\mathrm{B}(\overrightarrow{n_{a,-i}} + \vec{\beta})} \frac{\mathrm{B}(\overrightarrow{n_s} + \vec{\alpha})}{\mathrm{B}(\overrightarrow{n_{s,-i}} + \vec{\alpha})} \end{aligned} \quad (6)$$

$$\propto \frac{\Gamma(n_{\mathrm{k}}^{(z,s)} + \eta)}{\Gamma(\sum_{k=1}^{\mathrm{V}} (n_k^{(z,s)} + \eta))} \frac{\Gamma(\sum_{k=1}^{\mathrm{V}} (n_{k,-i}^{(z,s)} + \eta))}{\Gamma(n_{\mathrm{k,-i}}^{(z,s)} + \eta)}$$

$$\frac{\Gamma(n_\chi^{(z,s)} + \beta)}{\Gamma(\sum_{\chi=1}^{\mathrm{A}} (n_\chi^{(z,s)} + \beta))} \frac{\Gamma(\sum_{\chi=1}^{\mathrm{A}} (n_{\chi,-i}^{(z,s)} + \beta))}{\Gamma(n_{\chi,\mathrm{-i}}^{(z,s)} + \beta)}$$

$$\frac{\Gamma(n_{\mathrm{s}}^{(d)} + \alpha)}{\Gamma(\sum_{\mathrm{s}=1}^{\mathrm{L}} (n_{\mathrm{s}}^{(d)} + \alpha))} \frac{\Gamma(\sum_{\mathrm{s}=1}^{\mathrm{L}} (n_{\mathrm{s},-i}^{(d)} + \alpha))}{\Gamma(n_{\mathrm{s,-i}}^{(\mathrm{d})} + \alpha)}$$

where Γ is the gamma function.

By using the properties of gamma function ,equation (2) becomes:

$$\begin{aligned} & p(z_i = k, s_i = \mathrm{l}, \chi_i = t / z_{-i}, s_{-i}, \chi_{-i}, w, \pi) \\ & \propto \frac{n_{\chi,-i}^{(z,s)} + \beta}{\sum_{\chi}^{A} n_{\chi,-i}^{(z,s)} + \beta} \frac{n_{\mathrm{k,-i}}^{(z,s)} + \eta}{\sum_{k}^{V} n_{k,-i}^{(z,s)} + \eta} \frac{n_{\mathrm{s,-i}}^{(d)} + \alpha}{\sum_{s}^{L} n_{s,-i}^{(d)} + \alpha} \end{aligned} \quad (7)$$

Where $n_{\chi,-i}^{(z,s)}$ refers to the number of times all words have been assigned to topic z sentiment s and author χ except for the ith word; $n_{\mathrm{k,-i}}^{(z,s)}$ refers to the number of times that the word ,not including word k ,has been assigned to topic z and sentiment s; $n_{\mathrm{s,-i}}^{(d)}$ refers to the number of sentences that have been assigned to sentiment i in document d,and the sentences not include the word i.

For any sample obtained from the Markov chain,being an assignment of every word to a topic-sentiment -author label , we can estimate θ, ϕ and ψ using:

$$\theta_{z,s,\chi} = \frac{n_{\chi}^{(z,s)} + \beta}{\sum_{\chi=1}^{A} (n_{\chi}^{(z,s)} + \beta)} \quad (8)$$

$$\phi_{z,s,k} = \frac{n_{k}^{(z,s)} + \eta}{\sum_{k=1}^{V} (n_{k}^{(z,s)} + \eta)} \quad (9)$$

$$\psi_{s,d} = \frac{n_{s}^{(d)} + \alpha}{\sum_{s=1}^{L} (n_{s}^{(d)} + \alpha)} \quad (10)$$

where $n_{\chi}^{(i,j)}$ refers to the number of times that author a has been assigned to topic j and sentiment i, $n_{k}^{(i,j)}$ refers to the number of times that word k has been assigned to topic j and sentiment i, $n_{s}^{(d)}$ refers to the number of sentences that have been assigned to sentiment s in document d.

IV. EXPERIMENTAL

A. Data Set

Our data set is a collection of hotel reviews from tripadvisor.com/ that contain 10,000 reviews(105,734 sentences,2,875,213 words in total) from 2,741 authors. We chose 90% reviews as train date ,and the rest of reviews as the test data. Every review was about at least three aspects: service, location and rooms.

B. The Discover Of Sentiment-Topic

The result in Figure 2 is from a sample ran 1000 iterations.we set the number of sentiments L= 2,the number of topics K = 10.we set the hyper-parameters α = 1 , η = 1 and β = 0.01 through experience.

TABLE I. EXAMPLE OF SENTI-TOPIC DISCOVERED BY ATSM

Topic 1				**Topic 2**			
positive		negative		positive		negative	
WORD	PROB	WORD	PROB	WORD	PROB	WORD	PROB
staff	0.0541	reception	0.0321	room	0.0547	room	0.0481
friendly	0.0402	problems	0.0311	comfortable	0.0471	small	0.0354
helpful	0.0361	staff	0.0202	bed	0.0342	bed	0.0324
service	0.0298	no	0.0110	lobby	0.0282	no	0.0221
desk	0.0213	hotel	0.0104	window	0.0264	noisy	0.0214
concierge	0.0186	service	0.0098	water	0.0255	rooms	0.0195
excellent	0.0151	desk	0.0065	hotel	0.0244	terrible	0.0155
extremely	0.0101	poor	0.0047	nice	0.0194	hard	0.0074
hotel	0.0086	nobody	0.0014	rooms	0.0148	hotel	0.0041
great	0.0062	attitude	0.0005	clean	0.0117	cockroaches	0.0027

Topic 3				**Topic 4**			
positive		negative		positive		negative	
WORD	PROB	WORD	PROB	WORD	PROB	WORD	PROB
hotel	0.0647	location	0.0514	experience	0.0641	experience	0.0514
walk	0.0554	far	0.0384	excellent	0.0411	average	0.0210
location	0.0551	miles	0.0321	amazing	0.0384	bad	0.0187
station	0.0541	away	0.0278	free	0.0311	nothing	0.0154
walking	0.0497	center	0.0211	pleased	0.0291	no	0.0132
away	0.0421	city	0.0184	recommend	0.0255	only	0.0127
minutes	0.0411	hotel	0.0174	nice	0.0214	conditions	0.0117
close	0.0321	station	0.0024	hotel	0.0211	unacceptable	0.0024
bus	0.0211	long	0.0014	surprise	0.0104	never	0.0018
block	0.0191	minute	0.0007	food	0.0087	recommend	0.0014

TABLE II. EXAMPLE OF AUTHORS ASSOCIATED WITH A PROBABILITY DISTRIBUTION OF SENTIMENT-TOPICS

	Starina75		renodecor8r		kliles		cvguru	
Topic	positive	negative	positive	negative	positive	negative	positive	negative
1	0.0241	0.0142	0.0167	0.0084	0.0671	0.0014	0.0014	0.0021
2	0.0554	0.0234	0.0091	0.0014	0.0314	0.0214	0.0034	0.0214
3	0.0144	0.0195	0.0624	0.0241	0.0251	0.0101	0.0247	0.0347
4	0.0314	0.0114	0.0157	0.0041	0.0021	0.0310	0.0514	0.0124

We show the discovered sentiment-topic that are specific to some aspects such as location,rooms,experience and service in TABLE I.Each sentiment-topic is shown with the top 10 words that highest probability conditioned on that sentiment-topic.We can learn some interesting knowledge from those sentiment-topic such as people expressed their

positive sentiments about the service of the hotel with the words ‘staff’ ,’friendly’ and ‘helpful’.But using the word ‘reception’ ‘problem’ ’or ‘staff ’ to express their negative sentiment about the service.we can see that the word ‘away’ could used to express the both negative and positive sentiment about the location. Because the hotel may ‘ just a few miles away’ or ‘is far away’.

Seeing from the relationship(in TABLE II) between the author and probability distribution of sentiment-topics,we could know that Starina75 is most concerned about the rooms(topic 2) and he was satisfied with the rooms he has used .

C. *Perplexity*

An important evaluation in the probability model is Perplexity.The lower this value is, the better the performance of the model will be. The perplexity is:

$$Perplexity = \exp(-\frac{\sum_{N=1}^{N_d} \ln p(W_d, Z_d, S_d, X_d)}{N_d}) \quad (11)$$

where N_d is the number of test documents,P(W,Z,S,X) can get from equation (2).

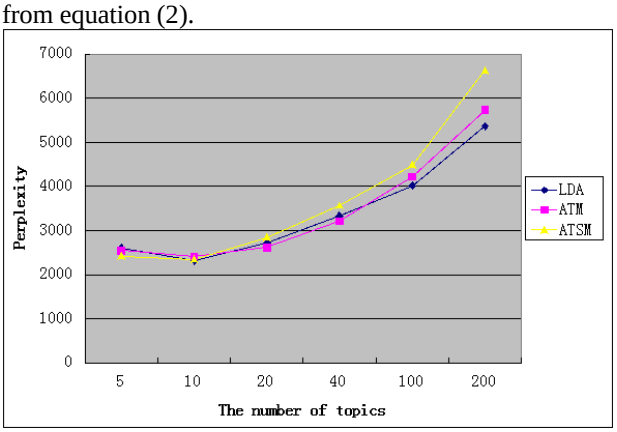

Figure 3. Perplexity of the test documents for different numbers of topics, for the LDA, ATM, and ATSM models.We set the number of sentiment L = 2.

We compare the LAD,ATM and ATSM by setting different number of topics. ATSM approaches the LAD and ATM,as illustrated by the similar perplexity in Fig. 3 .The performance will be poor when the number of topics is too large.

D. *Illustrative applications of the model*

The ATSM model could have many possible applications，such as we can automatically get all similar author to a given author .So we must know how to compute the similarity between authors. We define the the distance between authors i and j as follows:

$$KL(i,j) = \sqrt{\sum_{s=1}^{L}(\sum_{k=1}^{K}(\phi_{i,z,k}\log\frac{\phi_{i,z,k}}{\phi_{j,z,k}} + \phi_{j,z,k}\log\frac{\phi_{j,z,k}}{\phi_{i,z,k}}))^2} \quad (12)$$

Equation(12) is a form of the symmetric KL divergence.By using the distance measure above, we could classify the authors with the data mining .

TABLE III. EXAMPLE OF THE DISTANCE BETWEEN AUTHORS

	Starina75	renodecor8r	kliles	cvguru
Starina75	0.0	0.0433	0.1586	0.0099
renodecor8r	0.0132	0.0	0.0609	0.0433
kliles	0.0816	0.0132	0.0	0.1586
cvguru	0.0099	0.0609	0.0816	0.0

V. CONCLUSION

ATSM model proposed in this paper is a relatively simple probability model exploring the relationships between authors,sentiment,documents,topics and words.The experiment shows that ATSM can discover the sentiment-topic word relatively accurately and the Perplexity is acceptable.The primary benefit of the author-sentiment-topic mixture model is that it allows us to explicitly include authors’s sentiment in topic models. Possible future directions for this work include adding time sequence to the ATSM to model the change of the author’s sentiment over time.

ACKNOWLEDGMENT

This research is supported by program for the growth of young teachers of Hunan University.

REFERENCES

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

[2] 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. AUAI Press 2004: 487--494.

[3] Mei Q Z, Ling X, Wondra M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs //Proceeding of WWW’ 07. New York: ACM, 2007:171-180

[4] .Lin C H, He Y L. Joint sentiment/topic model for sentiment analysis // Proceeding of the 18th ACM conference on Information and knowledge management. New York: ACM, 2009: 375-384

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

[6] SUN Yan, ZHOU Xueguang. Unsupervised Topic and Sentiment Unification Model for Sentiment Analysis Acta Scientiarum Naturalist Universalistic Pekinensis ,2012:1-017

[7] Pang B, Lee L, Vaithyanathan S. Thumbs up?:sentiment classification using machine learning techniques // Proceedings of the Conference on Empirical Methods in Natural Language Processing.Stroudsburg: Association for Computational Linguistics .Philadelphia, PA, 2002: 79-86

[8] Resnik P, Hardisty E. Gibbs sampling for the uninitiated[R]. MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES, 2010.

[9] Shalinie S M, Sundarakantham K, Pushparathi S. A author topic model based unsupervised algorithm for learning topics from large text collections[C]//Recent Trends in Information Technology (ICRTIT), 2011 International Conference on. IEEE, 2011: 360-363.

[10] Heinrich G. Parameter estimation for text analysis[R]. Technical report, 2005.