Mining Social Emotions from Affective Text

Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu

Abstract—This paper is concerned with the problem of mining social emotions from text. Recently, with the fast development of web 2.0, more and more documents are assigned by social users with emotion labels such as happiness, sadness, and surprise. Such emotions can provide a new aspect for document categorization, and therefore help online users to select related documents based on their emotional preferences. Useful as it is, the ratio with manual emotion labels is still very tiny comparing to the huge amount of web/enterprise documents. In this paper, we aim to discover the connections between social emotions and affective terms and based on which predict the social emotion from text content automatically. More specifically, we propose a joint emotion-topic model by augmenting Latent Dirichlet Allocation with an additional layer for emotion modeling. It first generates a set of latent topics from emotions, followed by generating affective terms from each topic. Experimental results on an online news collection show that the proposed model can effectively identify meaningful latent topics for each emotion. Evaluation on emotion prediction further verifies the effectiveness of the proposed model.

Index Terms—Affective text mining, emotion-topic model, performance evaluation.

1 Introduction

Recent years have witnessed a rapid growth of online users and their increasing willingness to engage in social interactions. This inspires numerous social websites, e.g., Sina.com.cn[1] and People.com.cn,[2] to provide a new service that allows users to share their emotions after browsing news articles. Figs. 1a and 1b give two examples of eight social emotions, which are collected from 378 Sina users and 32 People users, respectively.

The user-generated social emotions provide a new aspect for document categorization, and they cannot only help online users select related documents based on emotional preferences, but also benefit a number of other applications such as contextual music recommendation [1]. Several studies have been carried out to automatically predict the most probable emotions for documents [2], [3], [4], [5]. Strapparava and Mihalcea [2] claimed that all words can potentially convey affective meaning. Every word, even those apparently neutral, can evoke pleasant or painful experiences because of their semantic relation with emotional concepts or categories. However, the way how text documents affect online users' social emotions is yet to be unveiled. In this paper, we refer to the problem of discovering and mining connections between social emotions and online documents as *social affective text mining*, including predicting emotions from online documents, associating emotions with latent topics, and so on.

To this end, a straightforward method is to manually build a dictionary of affective terms for each emotion, e.g., SentiWordNet [6] and WordNetAffect [7]. But building a dictionary is not only labor consuming, but also unable to quantize the connection strengths between affective terms and social emotions. As an alternative, the Emotion-Term model, a.k.a. Naïve Bayes, provides a principled way to estimate term-emotion associations using their cooccurrence counts. It was shown to be one of the best approaches in SemEval 2007 [2]. However, since such an approach treats each term individually, most relevant terms mined for each emotion are usually mixed with background noisy terms that do not convey much affective meaning. Moreover, it is more sensible to associate emotions with a specific emotional event/topic instead of only a single term.

In this paper, we propose a joint emotion-topic model for social affective text mining, which introduces an additional layer of emotion modeling into Latent Dirichlet Allocation (LDA) [8]. In more details, the model follows a three-step generation process for affective terms, which first generates an emotion from a document-specific emotional distribution, then generates a latent topic from a Multinominal distribution conditioned on emotions, and finally generates document terms from another Multinominal distribution based on latent topics. Because its exact inference is intractable, we develop an approximate inference method based on Gibbs sampling [9]. As a complete generative model, the proposed emotion-topic model allows us to infer a number of conditional probabilities for unseen documents, e.g., the probabilities of latent topics given an emotion, and that of terms given a topic.

We evaluate the proposed model on an online news collection containing 2,858 articles collected from the Sina society channel. It has 659,174 votes distributed in eight kinds of social emotions as shown in Fig. 1. Experimental results show that the proposed model can effectively

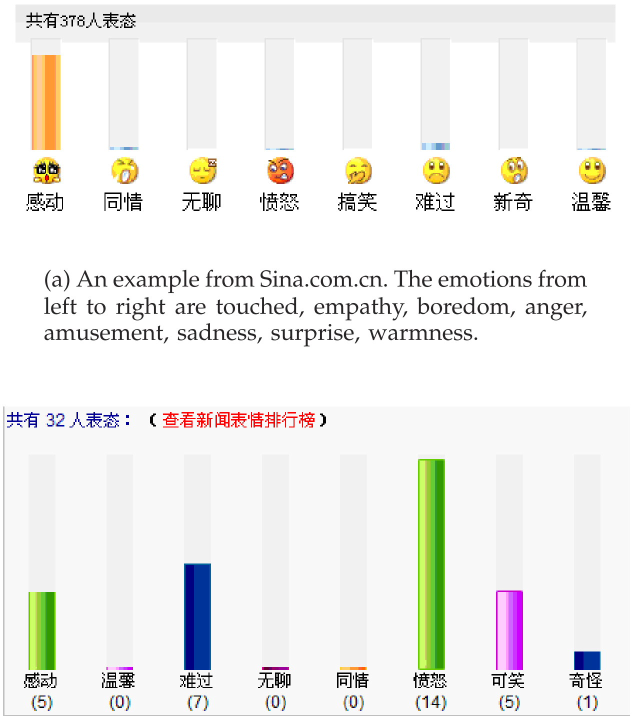
[1] One of the largest news portal in China. http://news.sina.com.cn/society/.

[2] One of the official news portal of Chinese government. http://opinion.people.com.cn/.

- *S. Bao, L. Zhang, and Z. Su are with IBM Research-China, Diamond Building, ZGC Software Park, #19, ShangDi, Beijing 100193, China. E-mail: {baoshhua, lizhang, suzhong}@cn.ibm.com.*
- *S. Xu, D. Han, and Y. Yu are with Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China. E-mail: {slxu, handy, yyu}@apex.sjtu.edu.cn.*
- *R. Yan is with Facebook Corp, 1601 S. California Ave, Palo Alto, CA 94304. E-mail: yanrong@gmail.com.*

Manuscript received 23 Nov. 2010; revised 28 Apr. 2011; accepted 3 May 2011; published online 23 Aug. 2011.
Recommended for acceptance by S. Parthasarathy.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2010-11-0625.
Digital Object Identifier no. 10.1109/TKDE.2011.188.

1041-4347/12/$31.00 © 2012 IEEE Published by the IEEE Computer Society



(a) An example from Sina.com.cn. The emotions from left to right are touched, empathy, boredom, anger, amusement, sadness, surprise, warmness.

(b) An example from People.com.cn. The emotions from left to right are touched, warmness, sadness, boredom, empathy, anger, amusement, surprise.

Fig. 1. Examples of social emotion statistics.

discover meaningful latent topics from news documents. The model can also distinguish the topics with strong emotions from background topics. For social emotion prediction, the proposed model outperforms the emotion-term model, term-based SVM model, and topic-based SVM model significantly by 34.3, 8.10, and 15.31 percent, respectively, in terms of the top-1 accuracy, which further verifies the effectiveness of jointly modeling emotions, topics, and affective terms.

The rest of the paper is organized as follows: Section 2 surveys the related work. Section 3 formalizes the problem of social affective text mining and describes the joint emotion-topic models with the emotion-term and LDA baseline models. Section 4 presents the experimental results. Section 5 gives several discussions. Finally, we make some concluding remarks in Section 6.

2 Related Work

This section reviews some of the related work on affective text mining and topic modeling, followed by discussing their connections and differences with the proposed model.

2.1 Affective Text Mining

There is a large body of previous work on mining affective content from text documents, e.g., product reputation mining [10], customer opinion extraction/summarization [11], [12], and sentiment classification [13]. However, none of these studies explores the connection between social emotions and affective terms.

The most related direction to our work is emotion prediction/classification. For example, SemEval introduced a task named “affective text” in 2007 [2], aiming to annotate short headline texts with a predefined list of emotions and/or polarity orientation (positive/negative). Alm et al. [4] explored text-based emotion prediction using supervised learning approaches. Strapparava and Mihalcea [5] evaluated several knowledge-based and corpus-based methods for the automatic identification of six emotions. Tokuhisa et al. [14] proposed a two step model for emotion classification using emotion-provoking event instances extracted from the web. Liu et al. [15] introduced a multiple corpus-based linguistic analysis and classified the emotion of e-mails at the sentence level.

Particularly, emotion predication in the context of web blog has been extensively studied. Yang et al. [3] investigated the emotion classification of blogs using machine learning techniques. Gill et al. [16] explored the emotion rating activities of 65 judges from short blog texts. An online system MoodViews has also been developed for tracking and searching emotion annotated blog posts [17], [18], [19], [20].

Despite the success of previous work on emotion prediction, existing approaches usually model documents under the “bag-of-word” assumption, so that the relationship across words is not taken into account. This also prevents us from further understanding the connections between emotions and contents in the topic level, because it is arguable that emotions should be linked to specific document topics instead of a single keyword.

Some applications of emotion mining have also been studied. For example, Liu et al. [21] introduced an approach for graphically visualizing the affective structure of a text document. Quite differently, with the help of proposed model, we are able to visualize the emotion assignments at the term level.

2.2 Topic Modeling

Numerous approaches for modeling latent document topics are available, especially in the context of information retrieval, e.g., probabilistic Latent Semantic Indexing (pLSI) [22] and LDA [8].

Because of its flexibilities and completeness, LDA has been extended to more advanced application domains with additional sampling steps. Rosen-Zvi et al. [23] merged author factors with document generation to jointly estimate document contents as well as author interests. From the perspective of model generation, their author variable shares some similarity with the emotion variable in our model. The key difference lies in different sampling distributions. Their author variable is chosen uniformly from a set of authors while emotion variable is sampled from multinomial distributions by the emotions contributed by web users. Titov and McDonald [24] described a new statistical model called the Multiaspect Sentiment model (MAS), which consisted of two independent components. Differently, the model proposed in this paper unifies the process of generating topics and associating emotions with texts. Wang and McCallum [25] combined time information with latent topics discovering in order to capture topic variance over time. Mei et al. [26] introduced Topic Sentiment Mixture model for sentiment analysis on the topic level. Lin et al. [27]

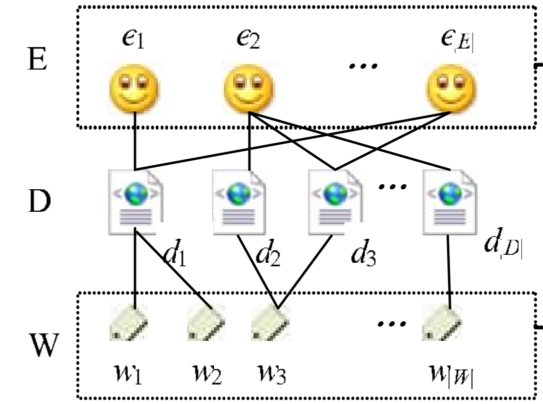


Fig. 2. Illustration of social affective text mining.

studied the problem of ideological discourse. Specifically, they proposed a joint topic and perspective model to uncover the words that represented an ideological texts topic as well as words that revealed ideological discourse structures. In contrast, this work extended LDA with a different set of information, i.e., social emotions contributed by online users, in the latent topics modeling process, of which the details are discussed as follows:

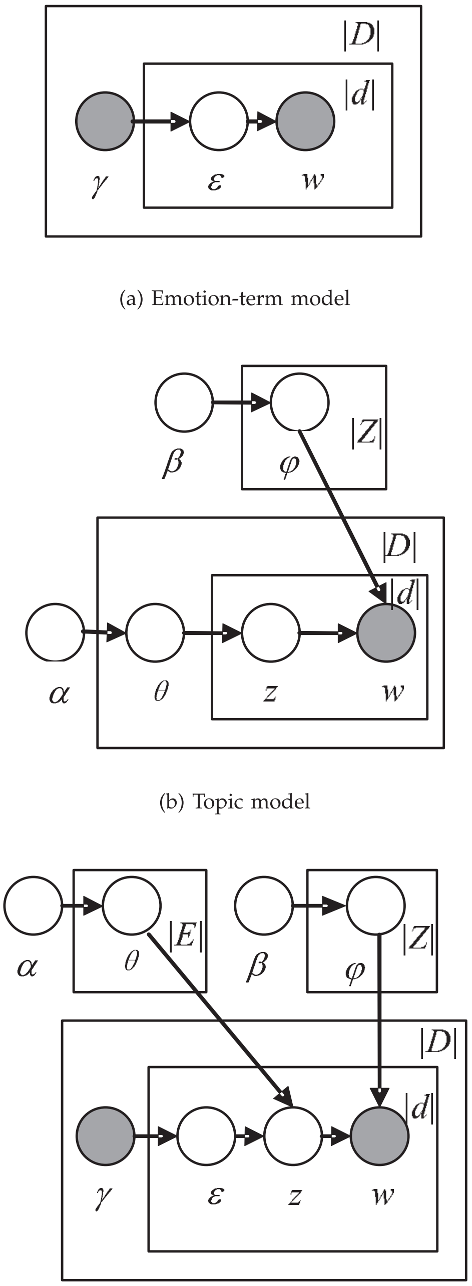
3 Social Affective Text Mining

Let us introduce our notations and terminologies for social affective text mining. An online text collection D is associated with a vocabulary W, and a set of predefined emotions E. In particular, each document $d \in D$ consists of a number of words $\{w_i\}, w_i \in W$, and a set of emotion labels $\{e_k\}, e_k \in E$. For each emotion e, $\gamma_d = \{\gamma_{d,e}\}$ represents its frequency count collected by the social websites. Similarly, the count of each word w is denoted as $\delta_d = \{\delta_{d,w}\}$. Our objective is to accurately model the connections between words and emotions, and improve the performance of its related tasks such as emotion prediction. An illustration of the social affective mining problems is shown in Fig. 2.

To achieve this, we first present two baseline models: 1) emotion-term model that uses Naïve Bayes to model social emotion and affective terms via their cooccurrences and 2) a LDA topic model which utilizes the term cooccurrence information within a document and discovers the inherent topics within affective text. Then, we describe the proposed emotion-topic model that can jointly estimate the latent document topics and emotion distributions in a unified probabilistic graphical model.

3.1 Emotion-Term Model

A straightforward method to model the word-emotion associations is called the emotion-term model, which follows the Naïve Bayes method by assuming words are independently generated from social emotion labels. Fig. 3a illustrates the generation process of the emotion-term model. It generates each word w_i of document d in two sampling steps, i.e., sample an emotion e_i according to the emotion frequency count γ_d, and sample a word w_i given the emotion under the conditional probability $P(w|e)$.



(a) Emotion-term model

(b) Topic model

(c) Emotion-topic model

Fig. 3. Graphical representation of models for affective text mining.

The model parameters can be learned by maximum-likelihood estimation. In particular, the conditional probability of a word w given an emotion e can be estimated as follows:

$$P(w|e) = \frac{|(w,e)|}{\sum_{w' \in W} |(w',e)|}, \tag{1}$$

where $|(w,e)|$ is the cooccurrence count between word $w \in W$ and emotion $e \in E$ for all the documents. It can be formally derived based on the word and emotion frequency counts

$$|(w,e)| = S + \sum_{d \in D} \delta_{d,w} \cdot \gamma_{d,e}, \tag{2}$$

where S is a small smoothing constant that is set to 1. To use the emotion-term models for predicting emotion on a new document d, we can apply the Bayes theorem under the term independence assumption

$$\begin{aligned} P(e|d) &= \frac{P(d|e)P(e)}{P(d)} \propto P(d|e)P(e) \\ &= P(e)\prod_{w \in d} P(w|e)^{\delta_{d,w}}, \end{aligned} \tag{3}$$

where $P(e)$ is the a priori probability of emotion e. It can again be calculated by maximum likelihood estimation (MLE) from the emotion distribution of the entire collection.

3.2 Topic Model

Many topic models have been proposed and well studied in previous work, of which, LDA [8] is one of the most successful models. LDA addresses the overfitting problem faced by other models like pLSI by introducing a Dirichlet prior over topics and words. Although LDA can only discover the topics from document and cannot bridge the connection between social emotion and affective text, for the ease of understanding in the following description, we make a simple review of LDA here.

As illustrated in Fig. 3b, LDA assumes the following generative process for each word w_i from topic z_i in document $d_i \in D$

$$\begin{array}{lcl} \theta & \sim & \text{Dirichlet}(\alpha) \\ z_i|\theta_{d_i} & \sim & \text{Multi-Nominal}(\theta_{d_i}) \\ \varphi & \sim & \text{Dirichlet}(\beta) \\ w_i|z_i, \varphi_{z_i} & \sim & \text{Multi-Nominal}(\varphi_{z_i}), \end{array}$$

where α and β are hyperparameters, specifying the Dirichlet priors on θ and φ.

In the first study of LDA, Blei et al. [8] proposed a convexity-based variational inference method for inference and parameter estimation under LDA. In this paper, we employ an alternative parameter estimation method, Gibbs sampling [28], which is more computationally efficient. Equation (4) shows the full conditional distribution used in the Gibbs sampling

$$P(z_i = j|z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + |W|\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + |Z|\alpha}, \tag{4}$$

where, n_{-i} means the count that does not include the current assignment of z_i, $n_j^{(w)}$ is the number of times word w has been assigned to topic j, and $n_j^{(d)}$ is the number of times a word from document d has been assigned to topic j.

3.3 Emotion-Topic Model

Emotion-term model simply treats terms individually and cannot discover the contextual information within the document. However, intuitively, it is more sensible to associate emotions with a specific emotional event/topic instead of only a single term. While topic model utilizes the contextual information within the documents, it fails to utilize the emotional distribution to guide the topic generation. In this paper, we propose a new approach called emotion-topic model. As illustrated in Fig. 3c, the emotion-topic model accounts for social emotions by introducing an additional emotion generation layer to Latent Dirichlet Allocation. For each document d, this model follows a generative process for each of its words w_i as follows:

$$\begin{array}{lcl} \theta & \sim & \text{Dirichlet}(\alpha) \\ \varepsilon_i & \sim & \text{Multinominal}(\gamma) \\ z_i|\varepsilon_i, \theta & \sim & \text{Multinominal}(\theta_{\varepsilon_i}) \\ \varphi & \sim & \text{Dirichlet}(\beta) \\ w_i|z_i, \varphi & \sim & \text{Multinominal}(\varphi_{z_i}), \end{array}$$

where α and β are hyperparameters for the Dirichlet priors on θ and φ, respectively. ε is sampled from a multinominal distribution parameterized by γ. $\varepsilon_i \in E$ and $z_i \in Z$ are corresponding emotion and topic assignment for word w_i. Z here means the set of topics used for emotion-topic modeling. We will discuss the choice of topic numbers in detail in Section 4.2. Similar to previous work [28], α and β are set to symmetric Dirichlet priors with value of $50/|Z|$ and 0.1, respectively. The emotion frequency count γ is normalized and summed to 1 in this case, which allows it to serve as the parameters for the multinominal distribution.

According to this generative process, the joint probability of all the random variables for a document collection is

$$\begin{aligned} P(\gamma, \varepsilon, \mathbf{z}, \mathbf{w}, \theta, \varphi; \alpha, \beta) = P(\theta; \alpha)P(\varphi; \beta) \\ P(\gamma)P(\varepsilon|\gamma)P(\mathbf{z}|\varepsilon, \theta)P(\mathbf{w}|\mathbf{z}, \varphi). \end{aligned} \tag{5}$$

Because it is intractable to perform an exact inference for the emotion-topic model, we develop an approximate inference method based on Gibbs sampling. Specifically, for each word, we estimate the posterior distribution on emotion ε and topic z based on the following conditional probabilities, which can be derived by marginalizing the above joint probabilities in (5)

$$\begin{aligned} &P(\varepsilon_i = e|\gamma, \varepsilon_{-i}, \mathbf{z}, \mathbf{w}; \alpha, \beta) \\ &\quad \propto \frac{\alpha + nz_{-i}^{e,z_i}}{|Z|\alpha + \sum_z nz_{-i}^{e,z}} \times \frac{\gamma_{d_i,e}}{\sum_{e'} \gamma_{d_i,e'}}, \end{aligned} \tag{6}$$

$$\begin{aligned} &P(z_i = z|\mathbf{z}_{-i}, \gamma, \varepsilon, \mathbf{w}; \alpha, \beta) \\ &\quad \propto \frac{\alpha + nz_{-i}^{\varepsilon_i,z}}{|Z|\alpha + \sum_{z'} nz_{-i}^{\varepsilon_i,z'}} \times \frac{\beta + nw_{-i}^{z,w_i}}{|W|\beta + \sum_w nw_{-i}^{z,w}}, \end{aligned} \tag{7}$$

where, e and z are the candidate emotion and topic for sampling, respectively. $d_i \in D$ indicates the document from which current word w_i is sampled. $nz^{e,z}$ is the number of times topic z has been assigned to emotion e. Similarly, $nw^{z,w}$ means the number of times a word w has been assigned to topic z. The suffix $-i$ of nz and nw means the count that does not include the current assignment of emotion and topic for word w_i, respectively.

In more details, we start the algorithm by randomly assigning all the words to emotions and topics. Then, we repeat Gibbs sampling on each word in the document collection by applying (6) and (7) sequentially. This sampling process is repeated for N iterations when the stop condition (detailed in Section 4.2) is met.

With the sampled topics and emotions available, it is easy to estimate the distributions of ε, θ, and φ as follows: Appendix shows the detailed derivation process

TABLE 1
Social Emotion Assignment Statistics

Emotion	Training	Testing	All
Touched	48,938	9,665	58,603
Empathy	48,198	11,481	59,679
Boredom	74,073	22,321	96,394
Anger	131,744	44,737	176,481
Amusement	80,840	17,647	98,487
Sadness	53,643	14,330	67,973
Surprise	52,167	19,065	71,232
Warmness	22,859	7,466	30,325
Total	512,462	146,712	659,174

$$\varepsilon_{d,e} = \frac{\gamma_{d,e}}{\sum_{e'} \gamma_{d,e'}}, \tag{8}$$

$$\theta_{e,z} = \frac{\alpha + nz^{e,z}}{|Z|\alpha + \sum_{z'} nz^{e,z'}}, \tag{9}$$

$$\varphi_{z,w} = \frac{\beta + nw^{z,w}}{|W|\beta + \sum_{w'} nw^{z,w'}}. \tag{10}$$

With all the parameters derived above, we can apply the emotion-topic model to various applications. For instance, for the task of emotion prediction, we can estimate the probability of a word w given an emotion e by integrating out the latent topics z

$$P(w|e) = \sum_{z} \theta_{e,z}\varphi_{z,w}. \tag{11}$$

4 Experiments

This section presents the experimental results on both joint emotion-topic modeling and its application to emotion prediction.

4.1 Experiment Setup

We collected 2,858 most-viewed news articles between April 2007 and February 2008 from the society channel of Sina.[3] The online users have assigned these news articles to one of the eight social emotions, i.e., touched, empathy, boredom, anger, amusement, sadness, surprise, and warmness, for a total of 659,174 times. The document collection is split into two disjoint sets based on their release time. The earlier 2,286 documents are used for training, and the rest 572 documents are used for testing. Table 1 summarizes the detailed statistics for each social emotion on both data sets.

Before the modeling process, all news articles are preprocessed and cleaned using the following steps:

1. Extract the title and main body of each news article. Note that, in contrast to previous work [2] that mainly focuses on title information, the access of the main body of news articles provides the basis for modeling latent topics and helps alleviate the issue of data sparseness.
2. Apply named entity recognition to filter out person names from the documents, because we found that few of the person names occurring in news articles bear any consistent affective meanings.
3. Segment all the words for each article with Stanford Chinese Word Segmenter [29],[4] followed by removing the most general terms (i.e., the terms with a document frequency larger than 500) and the most sparse terms (i.e., the terms with a document frequency of 1).

TABLE 2
Term Statistics After Cleansing

# of Terms	Training	Testing	All
Distinct	22,564	13,901	23,817
Total	365,094	91,241	456,335

Table 2 shows the detailed statistics of the data set after cleansing.

4.2 Topic Number Selection

In this section, we focus on selecting the iteration number for Gibbs sampling and the size of latent topics based on document modeling performance, which is measured by conditional perplexity, a standard measure to compare probability models. In more details, we calculate the conditional perplexity on generation of word w from social emotion frequency γ as follows:

$$Perp(D) = \exp\left\{-\frac{\sum_{d\in D}\sum_{w_i\in d}\ln P(w_i|\gamma_d)}{\sum_{d\in D}\delta_d}\right\}. \tag{12}$$

For the emotion-term model, $P(w_i|\gamma_d)$ can be estimated as

$$P(w_i|\gamma_d) = \sum_{e\in E}\frac{\gamma_{d,e}}{\gamma_d}P(w_i|e). \tag{13}$$

For the emotion-topic model, $P(w_i|\gamma_d)$ can be estimated as

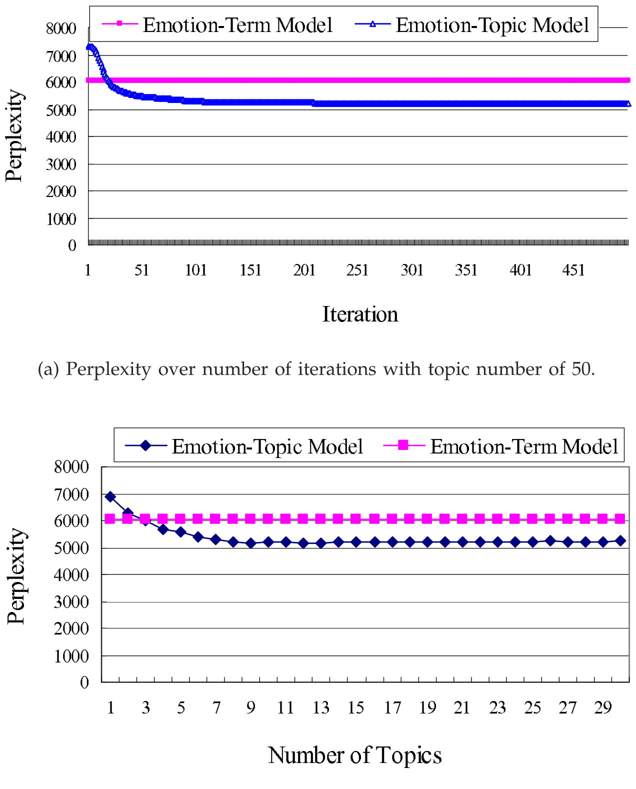
$$P(w_i|\gamma_d) = \sum_{e\in E}\sum_{z\in Z}\varepsilon_{d,e}\theta_{e,z}\varphi_{z,w_i}. \tag{14}$$

Fig. 4 depicts the perplexities for both emotion-term model and emotion-topic model.[5] Note that, because the emotion-term model does not consider any latent topics, its perplexity does not change with a growing number of topics and iterations. Fig. 4a shows the perplexity curves of the both models with 50 topics against the number of Gibbs sampling iterations. Our first observation is that the emotion-topic model converges to its asymptote in less than 200 iterations. After its convergence, the perplexity of emotion-topic model is noticeably less than that of the emotion-term model. Fig. 4b presents the perplexity curves with a growing number of topic numbers. We found that the emotion topic model perplexity again converges within a relative small number of topics, and becomes less than the baseline when the topic number is more than 4.

[3] While the proposed model is language independent, we use this Chinese data set for evaluation since we have not found similar services in English yet.

[4] It is a key preprocessing step for Chinese documents.

[5] The perplexity of LDA topic model is not discussed here since its topics has no connection with any emotions.



(a) Perplexity over number of iterations with topic number of 50.

(b) Perplexity over number of topics with iteration number of 300.

Fig. 4. Perplexity change history.

These results indicate the proposed emotion-topic model can capture the relationship between emotions and words more accurately. Based on above observations, we choose 20 topics and 300 iterations as the default setting unless otherwise specified.

4.3 Emotion-Topic Association

As we discussed before, emotion-term model cannot utilize the term cooccurrence information within document and cannot distinguish the general terms from the affective terms. On the other side, the traditional topic model can only discover the latest topics underlying the document set and cannot bridge the connections between social emotions and affective texts. The proposed emotion-topic model utilizes the complementary advantages of both emotion-term model and topic model.

In other words, the key advantage for the proposed emotion-topic model is its ability to uncover hidden topics that exhibit strong emotions. Fig. 5 shows the emotion distribution over the automatically discovered topics when the topic number is set to 20. It is easy to observe that these latent topics can be categorized into three types,

1. Empty topic, such as topic 18 and 19, which has little opportunities to be generated from any emotion.
2. Single-emotion topic, which is dominated by a specific kind of emotion. For instance, topic 1, 2, 3, 8 and 11 mainly convey the emotion of *amusement*, *surprise*, *anger*, *boredom*, and *warmness*.
3. Multiemotion topic, which shares a variety of emotions at the same time. These topics can be further categorized to two classes: 1) Emotion-background topic, such as topic 7 and 12, which is able to evoke all kinds of emotions. 2) Emotion-cluster topic, e.g., topic 4, which describes the people's livelihoods and mingles the feeling of *empathy*, *touched* in the same topic.

To offer a deeper understanding on topic discovery, Table 3 lists the top ranked terms for all the topics bearing strong social emotions. It can be found that, while most emotions like *sadness* and *anger* are associated with one significant topic in our corpus, the emotion of *surprise* is closely associated with two topics, i.e., Topic 2 and Topic 15. A closer analysis shows that Topic 2 is related to the news reports on exotic animals, while Topic 15 is related to the news on winning lottery awards. Although both these two topics share the similar *surprise* emotion, they belong to two independent events which are successfully identified by our emotion-topic model. The emotion of *empathy* also has the similar pattern, which can be generated by either Topic 4 that talks about the medical tragedies in persons' daily life, or topic 13 that reports criminal activities.

Table 4 lists the top ranked terms of 10 most salient topics learned by the LDA topic model with the same setting of topic numbers. While this two topic spaces share

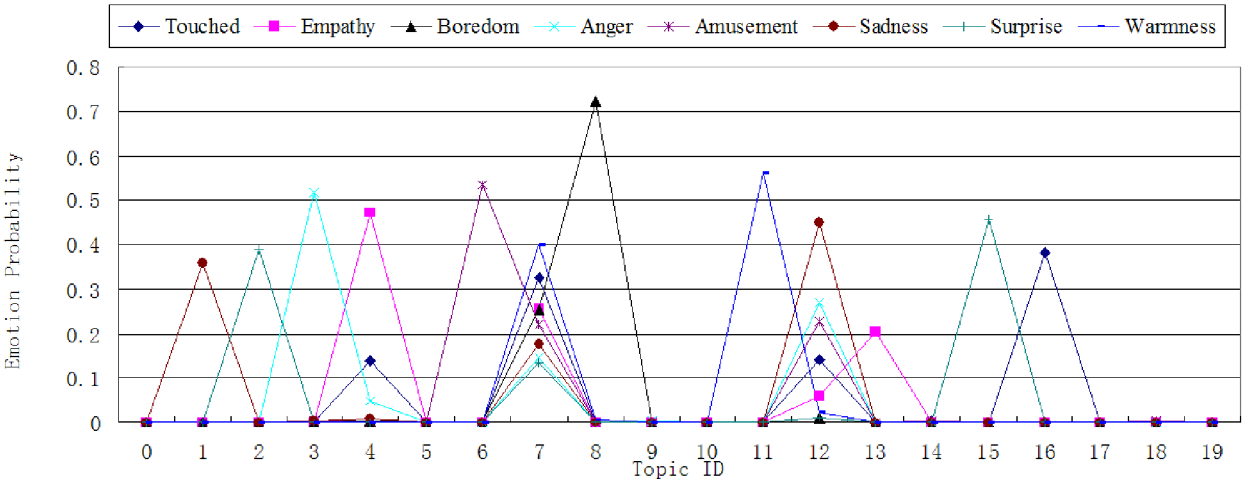


Fig. 5. Emotion distribution over topics.

TABLE 3
The Top Ranked Terms Discovered by the Emotion-Topic Model on the Subset of Topics with Strong Emotions
(All the Terms Are Translated from Chinese to English)

ID	Top 6 Terms in Each Topic	Emotion
1	suicide, accident, incident, salvage, corpse, stress	Sadness
2	award, kilogram, bonuses, snake, weight, cats	Suprise
3	suspect, judgments, public security, authority, rape, intentionally	Anger
4	work, sister, brother, disease, illness, serious	Empathy
6	even, divorce, website, young lady, woman, actually	Amusement
7	feeling, all, so, back, body, then	Mixed
8	networks, college students, net, relationships, women, media	Boredom
11	bride, wedding, groom, sisters, happy, new couple	Warmness
12	early morning, bus, received, stop, causing, alarm	Mixed
13	hometown, head, black bear, regulations, euthanasia, surrender	Empathy
15	lotteries, study, award, win, phase,record	Suprise
16	save, touched, insist, transplant, story, thanks	Touched

TABLE 4
The Top Ranked Terms Discovered by the LDA Model
(All the Terms Are Translated from Chinese to English)

ID	Top 6 Terms in Each Topic	Weight
1	suspect, police, get, knife, alarm, capture	0.0905
2	salvage, accompany, train, first-aid, urgency, timely	0.0905
3	bed, feet, drink, old partner, grandma, madame	0.0733
4	education, female student, undergraduate, study, elementary school, teacher	0.0647
5	save, fire protection, woman, children, rain, river	0.0560
6	tourist, program,ceremony, perform, media, station	0.0474
7	organization, intent, rape, sentence, crime, girl	0.0388
8	district, suicide, heard, body, door, cause	0.0388
9	cat, animal, them, kilogram, host, weight	0.0388
10	sister, brother, work, couple, exam, study	0.0302
11	bus, accident, passenger, transportation, stop, ride	0.0302
12	net, network, write, female, professor, site	0.0302

some similarity, it is easy to find the advantages of emotion-topic model as follows:

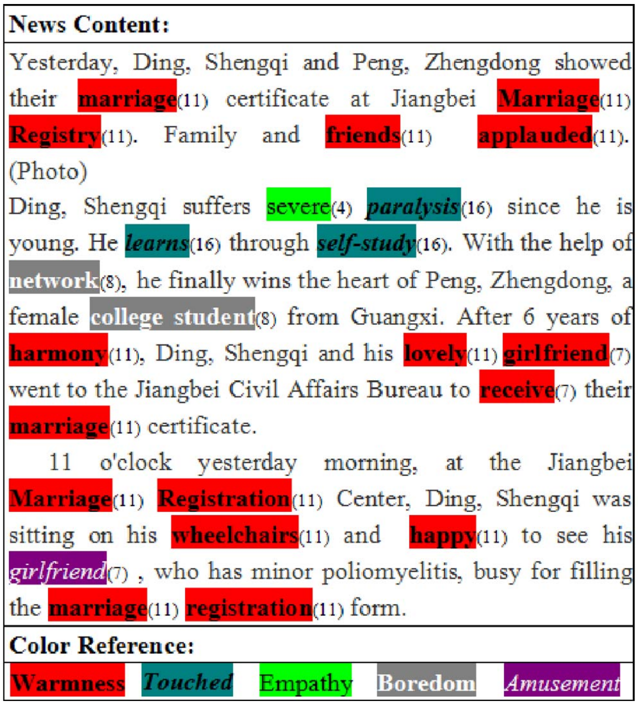
1. Topics with Emotion Focus. For example, both topic 16 in Table 3 and Topic 5 in Table 4 are start with topic term *save* and related to some reports of urgency. However, topic 16 in Table 3 is closely related to several touching reports, such as *organ donation and transplant*, indicated by several emotional terms such as *touched* and *insist*.
2. Merged Topics with Similar Emotions. Topic 3 in Table 3 with *anger* emotion is closely related to both topic 1 and 7 in Table 4. However, with the help of emotion guidance, these two topics in Table 4 are merged together, as shown in Table 3.
3. Additional Emotional Topics. Besides the topics with overlap, there are several isolated topics. For example, Topic 11 about *wedding* in Table 3 and Topic 4 about *education* in Table 4. That's to say, based on the content itself, the *wedding* topic is not salient enough. However, with the guidance of emotions, it becomes clear and emerges as a dominant topic of *warmness*. In contrast, the *education* topic is a salient content topic but not significant enough to serve as an emotional topic.

Above results confirm that the emotion-topic model is more effective in discovering one or more emotion-dependent topics for each emotion.

To illustrate the ability of mixing emotions on the document generation process, we give a case study in Fig. 6, which visualizes the emotion modeling result of a document entitled "A man with severe paralysis married a female college student."[6] One hundred and seventy nine web users submitted their emotions to this paper, where the most popular emotions are *warmness* (56 votes) and *touched* (51 votes). The affective terms with strong emotions are overlaid with distinctive fonts and colors based on the estimated emotion labels. The corresponding topic number is annotated as the number in the bracket following the word. As can be seen, the emotion-topic model successfully reveals a set of reasonable social emotions on a per term basis, e.g., *marriage* are assigned with emotion *warmness*, and *severe* are assigned with emotion *empathy*. It is worth pointing out that, even the same term can have different emotion assignments. One example is the term *girlfriend* which is connected with both *amusement* and *warmness*.

In Fig. 6, we can find that "college student" is assigned with emotion "boredom" with the context term "network." To further illustrate the model's capability of handling collections with terms that have ambiguous contexts, we visualize three more "college students" related snippets with

[6] http://news.sina.com.cn/s/2007-05-12/024612968284.shtml.



News Content:

Yesterday, Ding, Shengqi and Peng, Zhengdong showed their marriage(11) certificate at Jiangbei Marriage(11) Registry(11). Family and friends(11) applauded(11). (Photo)

Ding, Shengqi suffers severe(4) *paralysis*(16) since he is young. He *learns*(16) through *self-study*(16). With the help of network(8), he finally wins the heart of Peng, Zhengdong, a female college student(8) from Guangxi. After 6 years of harmony(11), Ding, Shengqi and his lovely(11) girlfriend(7) went to the Jiangbei Civil Affairs Bureau to receive(7) their marriage(11) certificate.

11 o'clock yesterday morning, at the Jiangbei Marriage(11) Registration(11) Center, Ding, Shengqi was sitting on his wheelchairs(11) and happy(11) to see his *girlfriend*(7) , who has minor poliomyelitis, busy for filling the marriage(11) registration(11) form.

Color Reference:

Warmness *Touched* Empathy Boredom *Amusement*

Fig. 6. Case study of document generation. The emotion and number in the bracket indicate the emotion and topic id assigned by emotion-topic model.

different contexts in Fig. 7. Snippet 1 comes from a document entitled "two college students jump to death within one week in a college located in TsingDao." Snippet 2 comes from an article describing that a 52-aged person sells his company and become a college student by passing the entrance examination. Snippet 3 tells us an interesting love story where the student's confession is expressed by turning the dormitory's light on or off to form the character of "$I\heartsuit U$." As we can see from the figure, the evoked emotions of "college student" change according to topical contexts, which differentiates our model from the term-based one.

4.4 Social Emotion Prediction

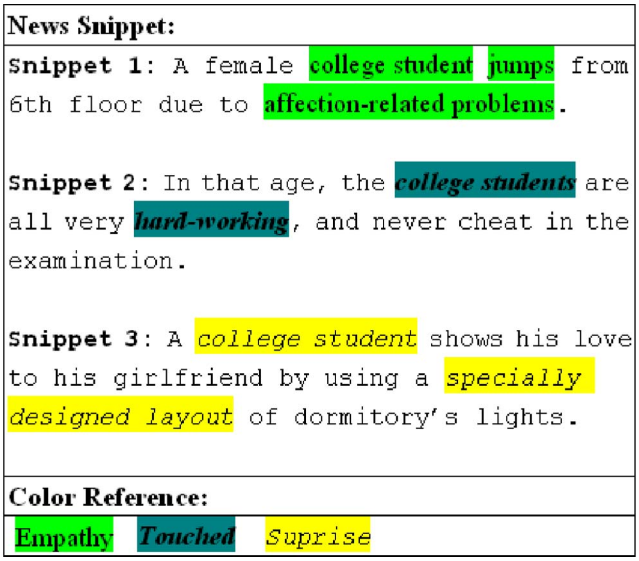
Because both the emotion-term and emotion-topic model can be applied to emotion prediction by estimating the probability of $P(e|d)$, we evaluate their prediction performance in this section. The evaluation metric is the accuracy at top N ($Accu@N, N = 1, 2, 3$). Given a document d, an truth emotion set $E_{topN@d}$ including the N top-ranked emotions, and the top ranked predicted emotion e_p, $Accu_d@N$ is calculated as

$$Accu_d@N = \begin{cases} 1 & \text{if } e_p \in E_{topN@d} \\ 0 & \text{else.} \end{cases} \quad (15)$$

Thus, the $Accu@N$ for the entire collection is

$$Accu@N = \sum_{d\in D} Accu_d@N/|D|. \quad (16)$$

Fig. 8 compares the emotion prediction results on the testing set. The emotion-topic model consistently outperforms the baseline emotion-term model in terms of Accuracy@1, 2 and 3. A T-Test shows that all these improvements are statistically significant (P-Value < 0.001). In particular, when using the most important metric of $Accu@1$, the emotion-topic model achieves an accuracy of 0.514 as compared with the baseline accuracy of 0.338. This can be translated into a relative improvement of 34.3 percent. Taking the complexity of human emotions in account, an accuracy of 0.514 on predicting eight emotions can be considered a relatively high score. In fact, according to the manual annotation study of SemEval, the average inter-annotator agreement measured by Pearson correlation measure is only 53.67 percent.



News Snippet:

Snippet 1: A female college student jumps from 6th floor due to affection-related problems.

Snippet 2: In that age, the *college students* are all very *hard-working*, and never cheat in the examination.

Snippet 3: A *college student* shows his love to his girlfriend by using a *specially designed layout* of dormitory's lights.

Color Reference:

Empathy *Touched* *Suprise*

Fig. 7. Case study of document generation on the term "college student."

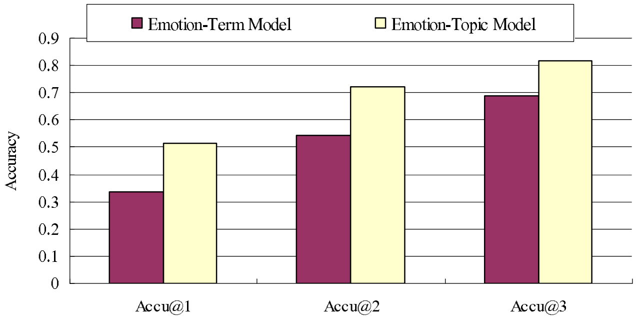


Fig. 8. Comparison of emotion prediction accuracy on Accuracy@1, 2, 3.

Besides emotion-term model, we introduce a new baseline which labels each document with the most popular emotion and treats the emotion prediction as a multiclass classification problem. More specifically, we use the C-SVC method of LibSVM[7] to train the classification model and then predict the most probable emotion of testing documents. The Kernel of Radial Basis Function (RBF) is adopted here and two most important parameters *gamma* and *cost* are tuned in a wide range.

Fig. 9 shows the accuracy of multiclass SVM over text terms with *gamma* ranging between 1 and 1/10,000, and *cost* ranging between 1 and 100. As we can see from the test data, even if we find the best parameter, i.e., $cost = 20$ and $gamma = 0.001$, it achieves an $Accu@1$ of 0.4755 which is 8.10 percent less than the proposed emotion-topic model. Fig. 10 shows the accuracy of multiclass SVM over 20 topics learned by the LDA topic model, with similar parameter ranges. Similarly, even we find the best parameter, i.e.,

[7] http://www.csie.ntu.edu.tw/cjlin/libsvm/.

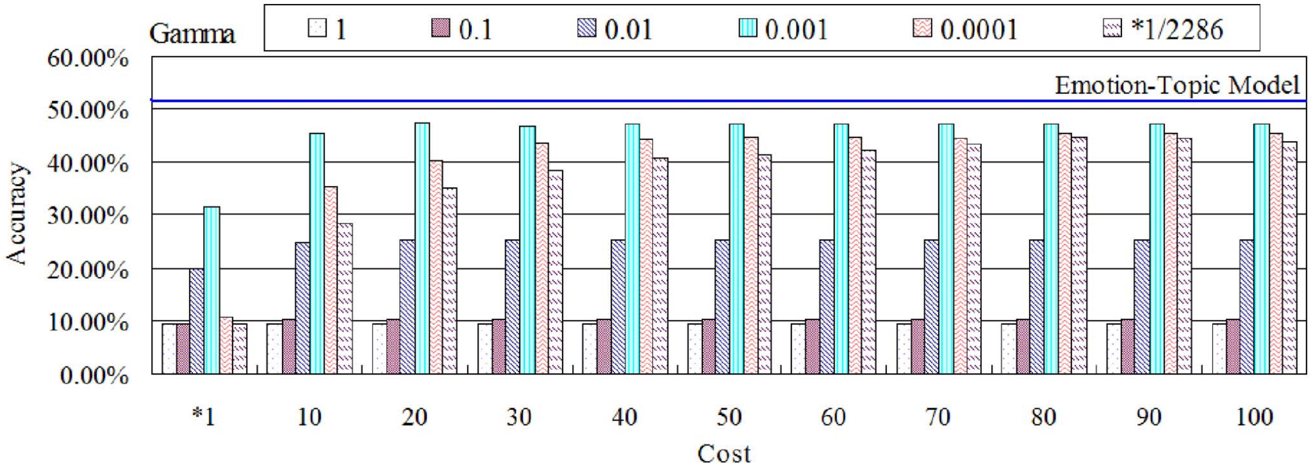


Fig. 9. Comparison of emotion prediction on Accuracy@1 between emotion-topic model and multiclass SVM on text terms with various parameters. The asterisk (*) in the figure indicates the default parameter used in LibSVM.

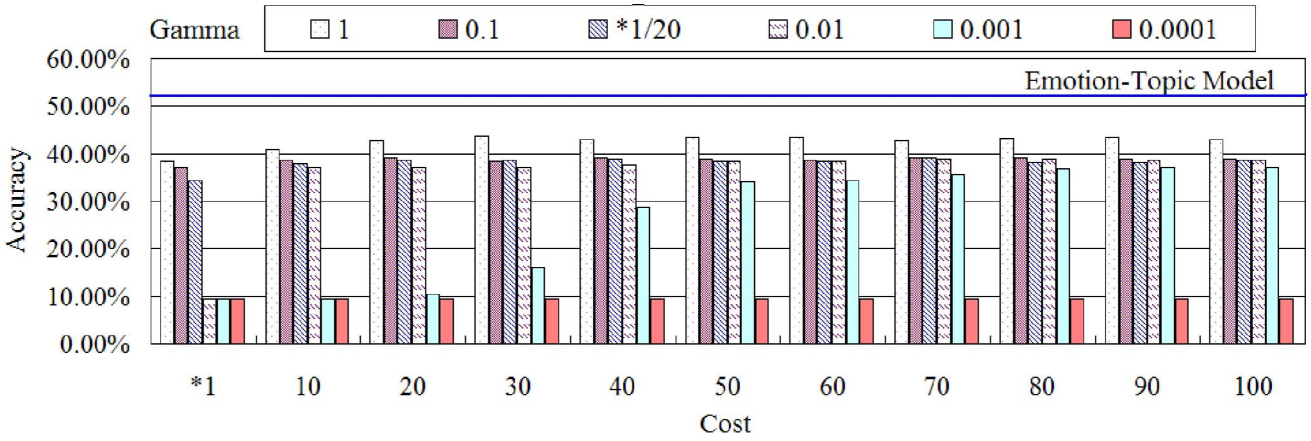


Fig. 10. Comparison of emotion prediction on Accuracy@1 between emotion-topic model and multiclass SVM on LDA topics with various parameters. The asterisk(*) in the figure indicates the default parameter used in LibSVM.

$cost = 50$ and $gamma = 1$, it achieves an $Accu@1$ of 0.4353 which is 15.31 percent less than the proposed emotion-topic model. A T-Test shows that the improvement of emotion-topic model over multiclass SVM is statistically significant (P-Value < 0.05) too. Other than the RBF kernel, we have also tried other kernels such as Linear Kernel and Polynomial Kernel with various parameters, however, the result is not as good as the best one achieved by the kernel of RBF.

Fig. 11 shows the learning curve of the emotion-topic model with different topic numbers. It is instructive to observe that the prediction accuracy also converges as the number of topics grows, and as expected, the convergence topic number is consistent with the convergence number of perplexities as shown in Fig. 4b.

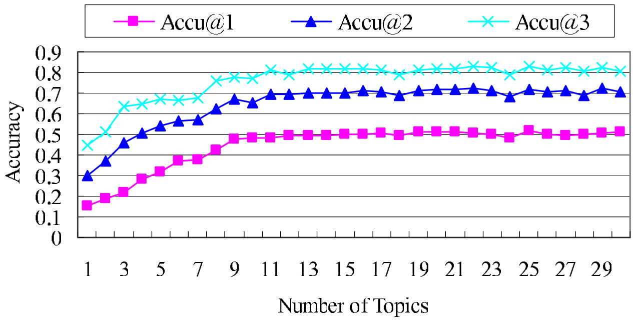


Fig. 11. Accuracy changes over different number of topics.

Another interesting application of this work is to study social perceptions associated with documents that record a real-world event. We apply our model to a time-stamped document collection about the great east Japan earthquake[8] and the prediction results successfully reveal the social emotions at different stages of the earthquake. Fig. 12 shows the predicted emotions of three documents released at different stages. News-1[9] is the first news article about the happening of this great earthquake and is dominated by emotion of "surprise." News-2[10] is a news article describing one of the most severe disasters in this earthquake, i.e., the explosion of Fukushima nuclear power plant following the earthquake, and our model assign a primary emotion of "empathy." News-3,[11] which is assigned with emotion "touched," reports the event that the emperor and empress of Japan will visit Saitama to extend regards to survivors.

To further evaluate the data set other than news articles, we test the proposed model on a set of forum data where the

[8] http://roll.news.sina.com.cn/s_japanearthquake0311_all/index.shtml.

[9] http://news.sina.com.cn/w/2011-03-11/135822095303.shtml.

[10] http://news.sina.com.cn/w/2011-03-14/152522112771.shtml.

[11] http://news.sina.com.cn/w/2011-04-07/175922252627.shtml.

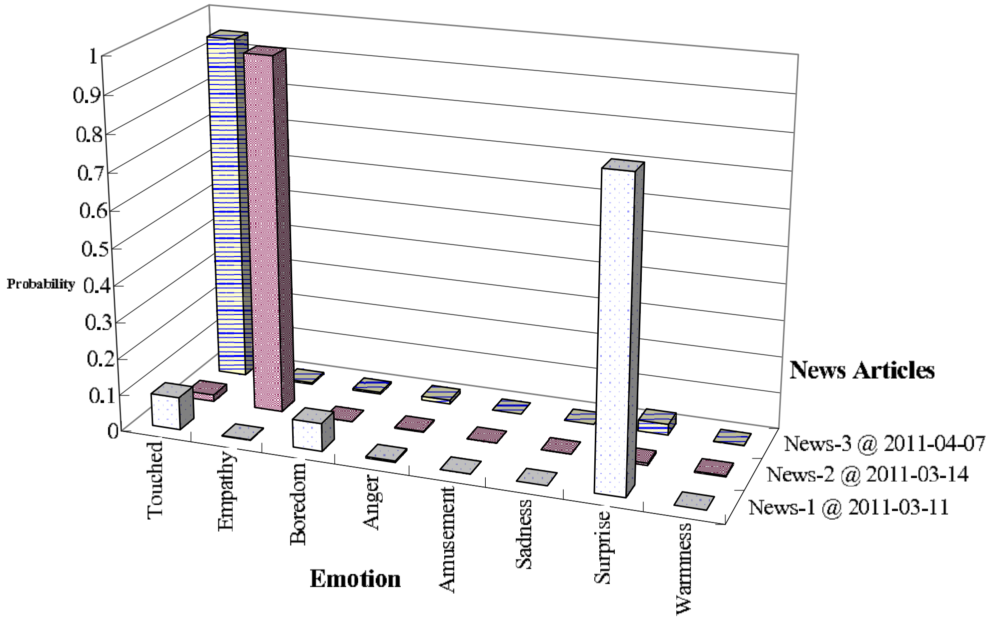
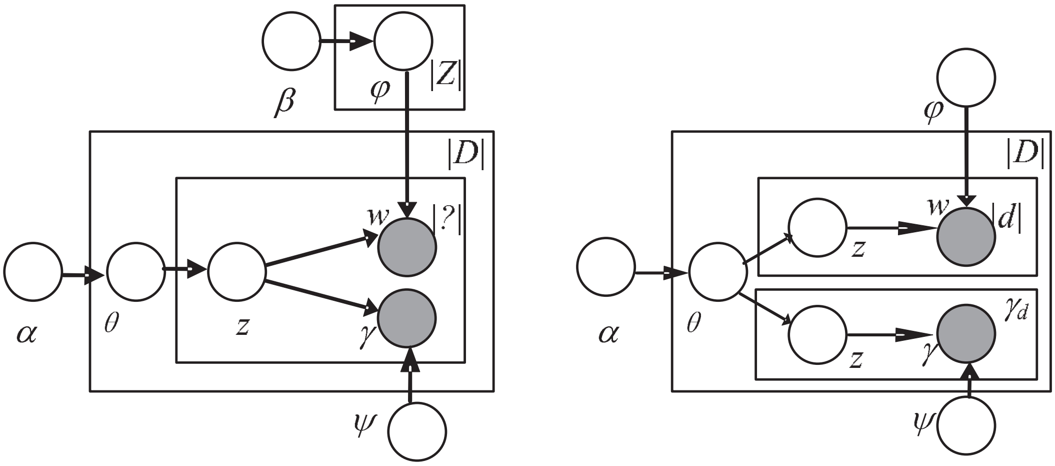


Fig. 12. Case study of emotion prediction on the great east Japan earthquake.



(a) Topic-Over-Emotion Model

(b) Mixed-Membership Model

Fig. 13. Sample alternatives of emotion topic model, where ψ_z is a multinomial distribution of emotion specific to topic z.

web users can freely share their comments on line. More particularly, we take 2009 Nobel prize as an example, most of the web users showed "surprise" to the inventors. However, we also find that quite a number of negative comments, which evoke emotions like "boring" and "amusement," challenging Obama's peace Nobel prize, since the nominations for the prize had to be postmarked by February 1, only 12 days after Obama took office. Just as Obama himself said: he viewed the decision less as a recognition of his own accomplishments and more as "a call to action."[12]

5 Discussion

While the proposed emotion topic model works well by adding an additional layer of emotions into LDA, there are several alternatives in modeling the similar process. Since our work is motivated by modeling the "user-generated social emotion" for online documents, a straightforward way is to model the intrinsic emotions in the document. Fig. 13a gives a generation model which has been successfully employed in other applications. To name a few, Topic-Over-Time model [25] associated a time-stamp with each token in the document d, and User-Question-Answer model [30] associated category information with each token in the document. However, it is not applicable here since the amounts of emotion votings and terms are not identical, and of course, there is no one-one mapping between the ith emotion voting and ith term in a document. Another intuitive way is to model the generation of emotions and words separately with shared document specific topic distribution, e.g., Mixed-Membership model proposed in [31], which is named as Link-LDA model in [32]. Fig. 13b gives its graphical model illustration. While this model does

[12] http://edition.cnn.com/2009/WORLD/europe/10/09/nobel.peace.prize/index.html.

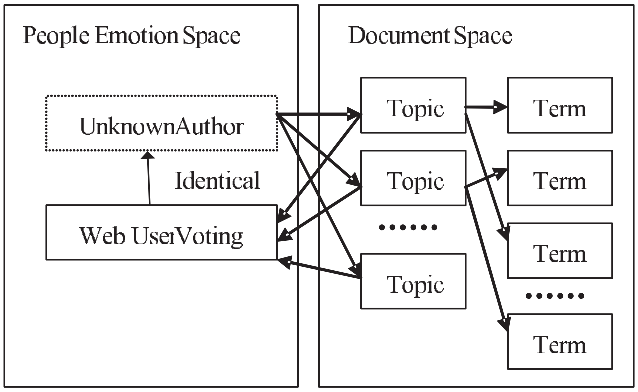


Fig. 14. Role illustration between people emotion space and document space.

not require observed one-one mapping between emotion and word, it fails to explicitly model the topical relationship between the emotion and text. As a result, it cannot mix the emotion in the generation of document text such as the one illustrated in Fig. 6.

Different from above models, the emotion topic model proposed in this paper can not only work without one-one mapping between emotion votings and terms, but also explicitly model the joint distribution of emotions and topics. However, it is not as intuitive as the two models above, since our basic idea is to generate the topics based on the emotions, which are in fact provided by the users after they read the document. The rationality that our model still works well in social affective text mining is that, we employ a hidden assumption that there is an unknown author for each document, who shares the similar emotion distribution with the document readers. Fig. 14 gives an illustration. Now, the generation process of our model becomes clear. Each document is first started with an emotion topic e of an unknown author, and then a general topic z is chosen from the emotion topic and finally the words w are generated from topic z. For instance, a document about general topic of *disaster* is most probably to be generated from emotion topic of *sympathy*, then the document is further generated from the specific topic *disaster* with various words, such as *hurricane*, *earthquake*, and *avalanche*.

6 Conclusion

In this paper, we present and analyze a new problem called social affective text mining, which aims to discover and model the connections between online documents and user-generated social emotions. To this end, we propose a new joint emotion-topic model by augmenting Latent Dirichlet Allocation with an intermediate layer for emotion modeling. Rather than emotion-term model that treats each term in the document individually and LDA topic model that only utilizes the text cooccurrence information, emotion-topic model allows associating the terms and emotions via topics which is more flexible and has better modeling capability. Experimental results on an online news collection show that the model is not only effective in extracting the meaningful latent topics, but also significantly improves the performance of social emotion prediction compared with the baseline emotion-term model and multiclass SVM.

As for future work, we are planning to evaluate our model with a larger scale of online document collections, and apply the model to other applications such as emotion-aware recommendation of advertisements, songs, and so on.

Appendix

Here, we present the derivation of the Gibbs sampling equations for emotion-topic model. Given the probability joint distribution of the probabilistic variables,

$$P(\varepsilon, z, w, \theta, \varphi; \alpha, \beta, \gamma), \tag{17}$$

we first integrate out θ and φ:

$$\begin{aligned} &P(\varepsilon, z, w; \alpha, \beta, \gamma) \\ &\quad= \int_\theta \prod_{e=1}^{|E|} P(\theta_e; \alpha) \prod_{\delta=1}^{|D|} \prod_{i=1}^{|d_\delta|} P(z_i|\varepsilon_i; \theta) d\theta \\ &\quad\times \int_\varphi \prod_{k=1}^{|Z|} P(\varphi_k|\beta) \prod_{\delta=1}^{|D|} \prod_{i=1}^{|d_\delta|} P(w_i|z_i; \varphi) d\varphi \\ &\quad\times \prod_{\delta=1}^{|D|} \prod_{i=1}^{|d_\delta|} P(\varepsilon_i; \gamma_d), \end{aligned} \tag{18}$$

where $|D|$ is the number of documents, and $|d_\delta|$ is the length of the δth document.

Specifically, we can write out the integration formulae more clearly:

$$\begin{aligned} &\int_\theta \prod_{e=1}^{|E|} P(\theta_e|\alpha) \prod_{\delta=1}^{|D|} \prod_{i=1}^{|d_\delta|} P(z_i|\varepsilon_i; \theta) d\theta \\ &\quad= \int_\theta \prod_{e=1}^{|E|} \int_{\theta_e} \frac{\Gamma(|Z|\alpha)}{\Gamma(\alpha)^{|Z|}} \prod_{k=1}^{|Z|} \theta_{e,k}^{n_{e,k,\cdot}} \\ &\quad= \prod_{e=1}^{|E|} \frac{\Gamma(|Z|\alpha)}{\Gamma(\alpha)^{|Z|}} \frac{\prod_{k=1}^{|Z|} \Gamma(\alpha + n_{e,k,\cdot})}{\Gamma(|Z|\alpha + \sum_k n_{e,k,\cdot})} \end{aligned} \tag{19}$$

$$\begin{aligned} &\int_\varphi \prod_{k=1}^{|Z|} P(\varphi_k|\beta) \prod_{\delta=1}^{|D|} \prod_{i=1}^{|d_\delta|} P(w_i|z_i; \varphi) d\varphi \\ &\quad= \prod_{k=1}^{|Z|} \int_{\varphi_k} \frac{\Gamma(|W|\beta)}{\Gamma(\beta)^{|W|}} \prod_{w=1}^{|W|} \varphi_{k,w}^{\beta-1} \varphi_{k,w}^{n_{\cdot,k,w,\cdot}} \\ &\quad= \prod_{k=1}^{|Z|} \frac{\Gamma(|W|\beta)}{\Gamma(\beta)^{|}W|} \frac{\prod_{w=1}^{|W|} \Gamma(\beta + n_{\cdot,k,w,\cdot})}{\Gamma(|W|\beta + \sum_w n_{\cdot,k,w,\cdot})} \end{aligned} \tag{20}$$

Now we derive the formulae for Gibbs Sampling. In the following, we list only the derivation process for the ε part, i.e., (6), since the derivation of the z part is almost the same.

$$
\begin{aligned}
&P(\varepsilon_i = \xi | \varepsilon^{-i}, z, w; \alpha, \beta, \gamma) \\
&\quad = \frac{P(\varepsilon_i = \xi, \varepsilon^{-i}, z, w; \alpha, \beta, \gamma)}{\sum_e P(\varepsilon_i = e, \varepsilon^{-i}, z, w; \alpha, \beta, \gamma)} \\
&\quad \propto P(\varepsilon_i = \xi, \varepsilon^{-i}, z, w; \alpha, \beta, \gamma) \\
&\quad = \prod_{e=1}^{|E|} \frac{\Gamma(|Z|\alpha)}{\Gamma(\alpha)^|Z|} \frac{\prod_{k=1}^{|Z|} \Gamma(\alpha + n_{e,k,\cdot})}{\Gamma(|Z|\alpha + \sum_k n_{e,k,\cdot})} \\
&\qquad \times \prod_{k=1}^{|Z|} \frac{\Gamma(|W|\beta)}{\Gamma(\beta)^{|W|}} \frac{\prod_{w=1}^{|W|} \Gamma(\beta + n_{\cdot,k,w,\cdot})}{\Gamma(|W|\beta + \sum_w n_{\cdot,k,w,\cdot})} \\
&\qquad \times \frac{\gamma_{d_i,\xi}}{\sum_\zeta \gamma_{d_i,\zeta}} \\
&\quad \propto \frac{\prod_{e,k,(e,k)\neq(\xi,z_i)} \Gamma\big(\alpha + n_{e,k,\cdot}^{-i}\big)\Gamma\big(\alpha + n_{\xi,z_i,\cdot}^{-i} + 1\big)}{\prod_{e\neq\xi} \Gamma\big(|Z|\alpha + \sum_k n_{i,k,\cdot}^{-i}\big)\Gamma\big(|Z|\alpha + \sum_k n_{\xi,k,\cdot}^{-i} + 1\big)} \\
&\qquad \times \frac{\gamma_{d_i,\xi}}{\sum_\zeta \gamma_{d_i,\zeta}} \\
&\quad \propto \frac{\prod_{e,k} \Gamma\big(\alpha + n_{e,k,\cdot}^{-i}\big)\big(\alpha + n_{\xi,z_i,\cdot}^{-i}\big)}{\prod_e \Gamma\big(|Z|\alpha + \sum_k n_{i,k,\cdot}^{-i}\big)\big(|Z|\alpha + \sum_k n_{\xi,k,\cdot}^{-i}\big)} \\
&\qquad \times \frac{\gamma_{d_i,\xi}}{\sum_\zeta \gamma_{d_i,\zeta}} \\
&\quad \propto \frac{\alpha + n_{\xi,z_i,\cdot}^{-i}}{|Z|\alpha + \sum_k n_{\xi,k,\cdot}^{-i}} \times \frac{\gamma_{d_i,\xi}}{\sum_\zeta \gamma_{d_i,\zeta}}
\end{aligned}
\tag{21}
$$

REFERENCES

[1] R. Cai, C. Zhang, C. Wang, L. Zhang, and W.-Y. Ma., "Musicsense: Contextual Music Recommendation Using Emotional Allocation," *Proc. 15th Int'l Conf. Multimedia,* pp. 553-556, 2007.

[2] C. Strapparava and R. Mihalcea, "Semeval-2007 Task 14: Affective Text," *Proc. Fourth Int'l Workshop Semantic Evaluations (SemEval '07),* pp. 70-74, 2007.

[3] C. Yang, K.H.-Y. Lin, and H.-H. Chen, "Emotion Classification Using Web Blog Corpora," *Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '07),* pp. 275-278, 2007.

[4] C.O. Alm, D. Roth, and R. Sproat, "Emotions from Text: Machine Learning for Text-Based Emotion Prediction," *Proc. Joint Conf. Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP '05),* pp. 579-586, 2005.

[5] C. Strapparava and R. Mihalcea, "Learning to Identify Emotions in Text," *Proc. 23rd Ann. ACM Symp. Applied Computing (SAC '08),* pp. 1556-1560, 2008.

[6] A. Esuli and F. Sebastiani, "Sentiwordnet: A Pub-Licly Available Lexical Resource for Opinion Mining," *Proc. Fifth Int'l Conf. Language Resources and Evaluation (LREC '06),* 2006.

[7] C. Strapparava and A. Valitutti, "Wordnet-Affect: An Affective Extension of Wordnet," *Proc. Fourth Int'l Conf. Language Resources and Evaluation (LREC '04),* 2004.

[8] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

[9] C.P. Robert and G. Casella, *Monte Carlo Statistical Methods,* second ed. Springer Publisher 2005.

[10] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fuku-shinna, "Mining Product Reputations on the Web," *Proc. Eighth ACM SIGKDDInt'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02),* pp. 341-349, 2002.

[11] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," *Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '04),* pp. 168-177, 2004.

[12] A.-M. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," *Proc. Joint Conf. Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP '05),* pp. 339-346, 2005.

[13] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP '02),* pp. 79-96, 2002.

[14] R. Tokuhisa, K. Inui, and Y. Matsumoto, "Emotion Classification Using Massive Examples Extracted From The Web," *Proc. 22nd Int'l Conf. Computational Linguistics (Coling '08),* pp. 881-888, 2008.

[15] H. Liu, H. Lieberman, and T. Selker, "A model of Textual Affect Sensing Using Real-World Knowledge," *Proc. Int'l Conf. Intelligent User Interfaces (IUI '03),* 2003.

[16] A.J. Gill, D. Gergle, R.M. French, and J. Oberlander, "Emotion Rating from Short Blog Texts," *Proc. 26th Ann. SIGCHI Conf. Human Factors in Computing Systems (CHI '08),* pp. 1121-1124, 2008.

[17] G. Mishne, K. Balog, M. de Rijke, and B. Ernsting, "Moodviews: Tracking and Searching Mood-Annotated Blog Posts," *Proc. Int'l AAAI Conf. Weblogs and Social Media (ICWSM '07),* 2007.

[18] K. Balog and M. de Rijke, "How to Overcome Tiredness: Estimating Topic-Mood Associations," *Proc. Int'l AAAI Conf. Weblogs and Social Media (ICWSM '07),* 2007.

[19] K. Balog, G. Mishne, and M. Rijke, "Why Are They Excited? Identifying and Explaining Spikes in Blog Mood Levels," *Proc. Ninth Conf. European Chapter of the Assoc. for Computational Linguistics (EACL '06),* 2006.

[20] G. Mishne and M. de Rijke, "Capturing Global Mood Levels Using Blog Posts," *Proc. AAAI Spring Symp. Computational Approaches to Analysing Weblogs (AAAI-CAAW '06),* 2006.

[21] H. Liu, T. Selker, and H. Lieberman, "Visualizing the Affective Structure of a Text Document," *Proc. CHI '03 Extended Abstracts on Human Factors in Computing Systems Conf.,* 2003.

[22] T. Hofmann, "Probabilistic Latent Semantic Indexing," *Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '99),* 1999.

[23] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The Author-Topic Model for Authors and Documents," *Proc. 20th Conf. Uncertainty in Artificial Intelligence (UAI '04),* pp. 487-494, 2004.

[24] I. Titov and R. McDonald, "A Joint Model of Text and Aspect Ratings for Sentiment Summarization," *Proc. 46th Ann. Meeting of the Assoc. for Computational Linguistics (ACL '08),* June 2008.

[25] X. Wang and A. McCallum, "Topic over Time: A Non-Markov Continuous-Time Model of Topical Trends," *Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '06),* pp. 424-433, 2006.

[26] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs," *Proc. 16th Int'l World Wide Web Conf. (WWW '07),* 2007.

[27] W.-H. Lin, E. Xing, and A. Hauptmann, "A Joint Topic and Perspective Model for Ideological Discourse," *Proc. European Conf. Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD '08),* pp. 17-32, 2008.

[28] T. Griffiths and M. Steyvers, "Finding Scientific Topics," *Proc. Nat'l Academy of Sciences USA,* vol. 101, pp. 5228-5235, 2004.

[29] P.-C. Chang, M. Galley, and C. Manning, "Optimizing Chinese Word Segmentation for Machine Translation Performance," *Proc. Assoc. for Computational Linguistics (ACL) Third Workshop Statistical Machine Translation,* 2008.

[30] J. Guo, S. Xu, S. Bao, and Y. Yu, "Tapping on the Potential of q&A Community by Recommending Answer Providers," *Proc. ACM 17th Conf. Information and Knowledge Management (CIKM '08),* 2008.

[31] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed-Membership Models of Scientific Publications," *Proc. Nat'l Academy of Sciences USA,* vol. 101, pp. 5220-5227, 2004.

[32] R.M. Nallapati, A. Ahmed, E.P. Xing, and W.W. Cohen, "Joint Latent Topic Models for Text and Citations," *Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '08),* pp. 542-550, 2008.

Shenghua Bao received the PhD degree in computer science from Shanghai Jiao Tong University in 2008. He is a staff researcher at IBM Research-China. His research interests lie primarily in web search, data mining, machine learning, and related applications. He received the IBM PhD Fellowship in 2007. Currently, he serves as a program committee member of international conferences like EMNLP '10 and WSDM '10, and a reviewer of several journals, including, *IEEE Transactions on Knowledge and Data Engineering*, *ACM Transactions on Asian Language Information Processing*, and *IPM*.

Shengliang Xu received the MS degree in computer science from Shanghai Jiao Tong University in 2010. He is currently working toward the PhD degree focusing on database in the Department of Computer and Software Engineering, University of Washington, Seattle. His research interests span broadly in the data management area.

Li Zhang is a research staff member on the Information Analysis and Search team. She has been working on many projects including information retrieval, text analytics, knowledge management, and business intelligence. She received the IBM Technical Accomplishment Awards for research in 2005 and 2007 and an IBM Outstanding Technical Accomplishment Award for research in 2008.

Rong Yan received the PhD degree from the Language Technologies Institute, School of Computer Science at Carnegie Mellon University in October 2006. He is now working at Facebook as a research scientist, focusing on large-scale machine learning, ad-targeting and optimization, and social media. He worked at IBM TJ Watson Research Center as a research staff member from Nov. 2006 to Nov. 2009. His previous project was IBM Multimedia Analysis and Retrieval (iMARS), which won the 2004 Wall Street Journal Innovation Award (Multimedia category). His research interests include large-scale data mining, ad optimization, machine learning, multimedia information retrieval, video content analysis, and computer vision.

Zhong Su is a research staff member at the IBM China Research Lab (CRL) and manages the Information Analysis and Search team. He has been involved in many projects in CRL including text analytics, enterprise search, metadata management, and information integration. He leads both the Chinese Knowledge Management and Natural Language Processing project teams, which received the IBM Technical Accomplishment Awards for research in 2005, 2006, 2007, and 2009, respectively, and the IBM Outstanding Technical Accomplishment Award for research in 2008. He is an IBM Master Inventor.

Dingyi Han received the BS degree in computer science in 2002 and the PhD degree in computer science in 2007 from Shanghai Jiaotong University, PR China. He is currently an assistant professor on the faculty of computer science & engineering, Shanghai Jiaotong University, P.R. China. His research interests include web science, peer-to-peer technology and semantic technology.

Yong Yu is the vice president of the Department of Computer Science & Engineering at Shanghai Jiao Tong University (SJTU). He received several prizes for his scholarship and his distinguished teaching career. In addition, as the head coach of the SJTU ACM-ICPC team, he and his team won the 2002 and 2005 ACM ICPC Championships. Currently, his main research topic is data and knowledge management, including web search and mining, semantic search, and social search.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.