STC: A Joint Sentiment-Topic Model for Community Identification

Baoguo Yang and Suresh Manandhar(✉)

Department of Computer Science, University of York, York, UK
by550@york.ac.uk, suresh@cs.york.ac.uk

Abstract. Traditional methods for identifying communities in networks are based on direct link structures, which ignore the content information shared among groups of entities. Recently, community detection approaches by using both link and content have been studied. It is necessary to identify communities with different sentiment distributions based on corresponding topics, which cannot be identified by existing community discovery techniques. To directly detect the sentiment-topic level communities and to better explore the hidden knowledge within them, we propose to integrate social links, content/topics, and sentiment information to work out a novel community model. Experimental results on two types of real-world datasets demonstrate that our model can not only achieve comparable performance compared with a state-of-the-art community model, but also can identify communities with different topic-sentiment distributions.

1 Introduction

The rapid growth of social medias provide us more chance to contact with other people and share our interests and opinions online, such as Facebook, Myspace, Twitter, etc. Email is considered as another kind of communication tool, which brings us more convenience to send or receive messages. A huge amount of data are generated online every day. Discovering previously unknown knowledge and relationships among people is very useful and necessary for individuals and organizations.

Example 1 (Email Networks): Email is widely used in our daily life, especially in companies and universities. Email correspondence produces abundant social messages associated with social relations. For teachers, their email recipients can be students, colleagues, friends, family members, librarians, and book publishers, etc. To get a high-level overview of the emails in our mailboxes, it is very interesting and necessary to discover our social communities in an automatic way. In each community, we are interested in the topics we discussed, people we contacted with, and the sentiment on some topics. Such information is latent and unobservable.

Example 2 (Hotel Twitters): Twitter, a popular microblogging platform, is not only used by individuals, but also very popular in many organizations,

W.-C. Peng et al. (Eds.): PAKDD 2014 Workshops, LNAI 8643, pp. 535–548, 2014.
DOI: 10.1007/978-3-319-13186-3_48

such as companies, hotels, and online supermarkets. As we know many hotels have their own twitter accounts. The customers can send their tweets about opinions and reviews to the hotels, and can comment on other tweets about the environment, food, and service of the hotels. To make full use of the data, it is useful to automatically identify communities associated with this twitter account. The communities with obvious negative polarities should be considered firstly. The hotel managers can take actions to address the main issues these customers proposed, and then response to these groups of people about the quality improvement of the hotel to win more customers, and to avoid the negative information proliferation across communities. Note that if we only extract collections of tweets including same sentiment topics by using traditional sentiment analysis methods instead of mining communities, the important social links will be ignored.

Based on the above examples, it is demanding to devise an effective community discovery approach to tackle these issues. The research on communities has a long history, and it has been paid widely attention in the past decade. In [2,9], Girvan and Newman propose a popular divisive community detection algorithm based on the concept of betweenness. To improve the speed of the algorithm in [2], a modified algorithm is proposed by Tyler et al. in [15]. Also some overlapping community detection methods has been proposed, like [4,17]. In addition, dynamic community discovery has been studied in recent years [3,10], where communities are not static but evolve over time.

However, most of the existing community identification methods intend to learn the community structures just using links, which ignore the content information in social networks. In recent years, the research on community detection has attracted increasing attention and achieved great progress. Discovering communities by combining link and content has been proposed in the literature [12,14,18–20], however, these methods fail to consider the valuable sentiment information in social networks.

In this paper, we propose a novel ***S****entiment-****T****opic model for* ***C****ommunity discovery*, called STC, which is built by using social links, topics and sentiment in a unified way, where the sentiment is studied based on its corresponding topic. The main goal of this approach is to discover sentiment level communities, i.e., to find out some communities containing dominant sentiments on certain topics even though not all communities have dominant sentiment topics. In our model, we define a community as a collection of people who are directly or indirectly connected and share some sentiment topics with some members in this collection. Note that not all the topics are discussed by every member of the community, also not all the members have the identical sentiment towards a certain topic, and the connectivity among members is also a very important factor. In many cases, even if two groups of people have similar sentiment-topic distributions, they are not included in the same community when the two groups follow different user distributions.

The rest of this paper is organized as follows: Sect. 2 introduces the related work. We present our community discovery model, the generative process and

parameter estimation in Sect. 3. In Sect. 4, we present and discuss the experimental results on two real-world datasets, the comparison with an up-to-date model is also reported. We give short discussion in Sect. 5, and the conclusions with future work are presented in Sect. 6.

2 Related Work

Traditional algorithms are focused on identifying disjoint communities [2,9], while in many real-world networks communities are allowed to overlap to some degree, where an entity can be included in multiple communities. The clique percolation method proposed by Palla et al. [11] is an early technique for overlapping community detection. Later, many algorithms have been proposed to improve the performance of the detection methods, such as OSLOM [4], SLPA [17], etc.

The above mentioned community identification methods ignore the content of social interactions in social networks. An early framework for community discovery using link and content elements is proposed in [19], the authors proposed two community-user-topic (CUT) models based on joint user and topic distributions. In [18], Yang et al. propose to integrate a popularity-based conditional link model with a discriminative content model into a unified framework to discover communities. For maximum likelihood inference, a novel two-stage optimization algorithm is proposed.

CART (Community-Author-Recipient-Topic) [12], a Bayesian generative model, is proposed to integrate link and content information in the social network for discovering communities, which is an extension of the Author-Recipient-Topic (ART) model [7]. It is assumed that the authors and recipients are generated from a latent group. Another novel method for detecting communities in social networks using links and content is proposed in [14]. In such method, the discussed topics, social links, and interaction types are all used to build several generative community models, namely, TUCM (Topic User Community Model), TURCM-1 and TURCM-2 (Topic User Recipient Community Models) and full TURCM model. More recently, a community profiling model, Collaborator Community Profiling (COCOMP), has been proposed by Zhou et al. in [20] to identify the communities of each user and their relevant topics and groups. In COCOMP, both the social links and topics between users are also considered. In [8,13], content and links are also learnt together to identify communities.

However, the above methods fail to consider the sentiment information of topics, which is an important factor when discovering more meaningful communities on a level of sentiment. The joint sentiment/topic model (JST) [6], an extension of the traditional Latent Dirichlet Allocation (LDA) model [1], is proposed to detect document-level sentiment and topic from documents. In [5], Li et al. introduce two probabilistic joint topic and sentiment models, namely, Sentiment-LDA and Dependency-Sentiment-LDA. Sentiments are related to topics in both of the models. However, JST, Sentiment-LDA, and Dependency-Sentiment-LDA are not proposed for community discovery.

To overcome the above problems and identify more meaningful communities, we propose our community model, STC, using topic, sentiment and user interactions in a unified way, which takes the topic-sentiment into consideration.

3 Our Community Discovery Model

The graphical representation of our proposed community model, STC, is shown in Fig. 1. There are mainly two different variables in this model, the latent variables and the observable ones:

- The latent (hidden) variables: Community assignment c ($c = 1, 2, \cdots, M$); Topic assignment z ($z = 1, 2, \cdots, K$); Sentiment label assignment l ($l = 1, 2, \cdots, S$).
- The observable variables: Word w (the word in the document); Person u (the person who is sharing the document).

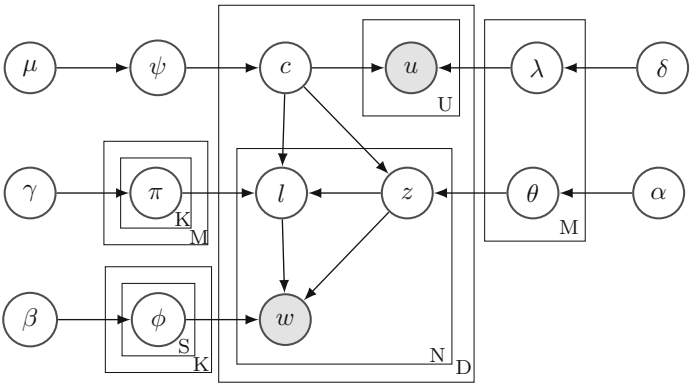


Fig. 1. Graphical notation of our proposed model.

3.1 Generative Process

Suppose there are K latent topics and S sentiment polarities, for each topic, and for each sentiment, we have: $\phi_{k,s}|\beta \sim Dir(\beta)$, where ϕ is the topic-sentiment distribution over words.

Let M be the number of communities, each community is related to three key parameters: (1) user participant mixture λ; (2) topic mixture θ; (3) sentiment mixture π. Specifically, in each community m ($m = 1, 2, ..., M$), θ_m is the topic mixture (proportion) for the community m, which follows a Dirichlet distribution $Dir(\alpha)$, λ_m is the user participant mixture with respect to community m, which has a Dirichlet distribution with hyperparameter δ. And $\pi_{m,k}$ is the sentiment mixture for topic k of community m. Note that the sentiments are studied

based on topics, it is not reasonable to study sentiments without considering the corresponding topics. For example, given two topics "laptop" and "weather", the sentiment words "nice" and "bad" can be used to describe both topics. It is not clear which topic is discussed by people with a sentiment word "nice" if the topic is not provided.

$$\theta_m|\alpha \sim Dir(\alpha), \quad \lambda_m|\delta \sim Dir(\delta), \quad \pi_{m,k}|\gamma \sim Dir(\gamma).$$

We define a community proportion ψ based on the whole corpus, $\psi|\mu \sim Dir(\mu)$. In this model, α, β, δ, γ, μ are the hyperparameters of Dirichlet distributions.

Then the generative process for each document d, $d = 1, 2, ..., D$ is shown as follows: Choose a community assignment c_d for a document d: $c_d|\psi \sim Mult(\psi)$.

Assume there are U_d people sharing a document d. For each person $u_{d,p}$ $(p = 1, 2, ..., U_d)$ associated with document d, the generative process is: Choose a user $u_{d,p}$ from the participant mixture of community c_d: $u_{d,p}|\lambda, c_d \sim Mult(\lambda_{c_d})$.

Suppose there are N_d word tokens in a document d, For each word token $w_{d,n}$ $(n = 1, 2, ..., N_d)$ in document d. The generative process is:

(1) Choose a topic assignment $z_{d,n}$ from the topic mixture of community c_d:

$$z_{d,n}|\theta, c_d \sim Mult(\theta_{c_d}).$$

(2) Choose a sentiment label $l_{d,n}$ from the c_d-th community's sentiment mixture:

$$l_{d,n}|c_d, z_{d,n}, \pi \sim Mult(\pi_{c_d,z_{d,n}}).$$

(3) Choose a word $w_{d,n}$ from the distribution $\phi_{k,s}$ over words defined by the topic $z_{d,n}$ and sentiment label $l_{d,n}$: $w_{d,n}|z_{d,n}, l_{d,n}, \phi \sim Mult(\phi_{z_{d,n},l_{d,n}})$.

From the graphical representation shown in Fig. 1, the joint probability for the proposed model can be written as Eq. 1.

$$\begin{aligned} &P(\mathbf{u}, \mathbf{c}, \mathbf{z}, \mathbf{l}, \mathbf{w}, \lambda, \psi, \theta, \pi, \phi|\delta, \mu, \alpha, \gamma, \beta) \\ &= P(\mathbf{u}|\mathbf{c}, \lambda)P(\mathbf{c}|\psi)P(\mathbf{z}|\mathbf{c}, \theta)P(\mathbf{l}|\mathbf{c}, \mathbf{z}, \pi)P(\mathbf{w}|\mathbf{z}, \mathbf{l}, \phi) \\ &\quad P(\lambda|\delta)P(\psi|\mu)P(\theta|\alpha)P(\pi|\gamma)P(\phi|\beta). \end{aligned} \tag{1}$$

3.2 Model Inference and Parameter Estimation

In this model, a document belongs to a single community rather than multiple communities. Each document is shared by at least two people (i.e., an author and at least one recipient) to make sure there is at least one link associated with a document. Once the sender (or the author) of the document is known, the user links associated with this document will be displayed. For inference, the statistics and variables are described in Table 1.

Let $t = (d, n)$, the conditional posterior probability of c_d, z_t, and l_t can be written as follows.

Table 1. List of statistics and variables.

Statistic/Variable	Description
D_m	the number of documents assigned to community m
D	the total number of documents
$n_{m,k}$ ($n_{m,k}^{-d}$)	the number of times word tokens in the documents of community m are assigned to topic k (excluding document d)
$n_{m,k,s}$ ($n_{m,k,s}^{-d}$)	the number of times word tokens in the documents of community m are assigned to topic k and sentiment label s (excluding document d)
n_m (n_m^{-d})	the total number of words in the documents of community m (excluding those in document d)
$n_{k,s,v}$ ($n_{k,s,v}^{-t}$)	the number of times a word v is assigned to topic k and sentiment label s (excluding the word in position t)
$n_{k,s}$ ($n_{k,s}^{-t}$)	the number of times words are assigned to topic k with sentiment label s (excluding the word in position t)
$f_{d,k}$	the number of word tokens in document d associated with topic k
f_d	the total number of words in document d
$f_{d,k,s}$	the number of word tokens in document d associated with topic k and sentiment label s
$n_{c_d,k}^{-t}$	the number of times word tokens in community c_d are assigned to topic k excluding the word in position t
$n_{c_d,k,s}^{-t}$	the number of times word tokens in community c_d are assigned to topic k and sentiment label s excluding the word in position t
$n_{c_d}^{-t}$	the total number of words in the documents of community c_d excluding the word in position t
$g_{m,p}$ ($g_{m,p}^{-d}$)	the number of times a person p is involved in the documents of community m (excluding document d)
g_m (g_m^{-d})	the number of times persons are involved in the documents of community m (excluding document d)
$e_{d,p}$	the number of times a person p is involved in the document d
e_d	the number of persons who are sharing the document d
$\mathbf{l}_{d_{(k)}}$	the sentiment set of topic k in document d
$\mathbf{z}_d$	the topic set of document d
$\mathbf{u}_d$	the person set of document d

$$
\begin{aligned}
&P(c_d = m|\mathbf{c}_{-d}, \mathbf{u}, \mathbf{z}, \mathbf{l}, \mathbf{w}) \\
&\propto \frac{D_m^{-d} + \mu_m}{\sum_{j=1}^{M} \mu_j + D - 1} \times \frac{\prod_{k \in \mathbf{z}_d} \prod_{i=0}^{f_{d,k}-1} (\alpha_k + n_{m,k}^{-d} + i)}{\prod_{i=0}^{f_d - 1} (\sum_{k=1}^{K} \alpha_k + n_{m,k}^{-d} + i)} \\
&\times \prod_{k \in \mathbf{z}_d} \frac{\prod_{s \in \mathbf{l}_{d_{(k)}}} \prod_{i=0}^{f_{d,k,s}-1} (\gamma_s + n_{m,k,s}^{-d} + i)}{\prod_{i=0}^{f_{d,k}-1} (\sum_{s=1}^{S} \gamma_s + n_{m,k,s}^{-d} + i)} \times \frac{\prod_{p \in \mathbf{u}_d} (\delta_p + g_{m,p}^{-d})}{\prod_{i=0}^{e_d - 1} (\sum_{p=1}^{P} \delta_p + g_m^{-d} + i)}.
\end{aligned} \tag{2}
$$

When the community assignment c_d for document d is obtained, for simplicity, the posterior distribution of z_t and l_t can be derived as follows.

$$P(z_t = k, l_t = s|\mathbf{w}, \mathbf{z}_{-t}, \mathbf{l}_{-t}, c_d) \propto \frac{n_{c_d,k}^{-t} + \alpha_k}{\sum_{k=1}^{K} n_{c_d,k}^{-t} + \alpha_k} \times \frac{n_{c_d,k,s}^{-t} + \gamma_s}{\sum_{s=1}^{S} n_{c_d,k,s}^{-t} + \gamma_s} \times \frac{n_{k,s,v}^{-t} + \beta_v}{\sum_{v=1}^{V} n_{k,s,v}^{-t} + \beta_v}. \quad (3)$$

The updated parameters are represented as follows:

$$\psi_m = \frac{D_m + \mu_m}{\sum_{m=1}^{M} \mu_m + D}, \quad \lambda_{m,p} = \frac{g_{m,p} + \delta_p}{\sum_{p=1}^{P} g_{m,p} + \delta_p}, \quad \theta_{m,k} = \frac{n_{m,k} + \alpha_k}{\sum_{k=1}^{K} n_{m,k} + \alpha_k},$$

$$\pi_{m,k,s} = \frac{n_{m,k,s} + \gamma_s}{\sum_{s=1}^{S} n_{m,k,s} + \gamma_s}, \quad \varphi_{k,s,v} = \frac{n_{k,s,v} + \beta_v}{\sum_{v=1}^{V} n_{k,s,v} + \beta_v}.$$

4 Experiment and Result Analysis

4.1 Experiment Setup

In the experiments, **two** types of datasets, the email dataset and the twitter microblog dataset are used. For Enron dataset[1], we randomly select five user folders, one of them called '*arnold-j*' is used for the experiment of individual user's perspective (denoted as arnold-j), and the other four folders, namely, *ermis-f*, *shively-h*, *whalley-g* and *zipper-a* are used together as a whole dataset (denoted as EnronFourUsrs). We conduct series of preprocessing work for arnold-j and EnronFourUsrs[2], like the initial duplicated email removal and the basic text mining preprocessing (stopwords removal, stemming, etc.). The second type of dataset is a twitter corpus[3], which includes 5513 tweets, covering 4 main topics, namely, Apple, Google, Microsoft, and Twitter. We kept the tweets belonging to one of the three sentiments (i.e., positive, negative and neutral), then the empty tweets and the ones without recipients are all removed. Some screen names are extracted from the text of tweets as the recipients, we also preprocess it to make the final document format the same as the Enron datasets. As for the four main topics in original twitter dataset, in fact, each main topic can be divided into several subtopics. The final preprocessed datasets for our experiments are shown in Table 2.

As the work in [5,6], we also use the subjectivity lexicons as prior information for model learning. Specifically, we use MPQA[4] [16] as the sentiment prior knowledge.

In our model, the initial values of the symmetric hyperparameters are set as: $\alpha = 50/K$, $\beta = \delta = \gamma = \mu = 0.1$. The collapsed Gibbs sampling algorithms are

[1] http://www-2.cs.cmu.edu/~enron/
[2] Note that we will use Enron to represent EnronFourUsrs in the following sections.
[3] http://www.sananalytics.com/lab/twitter-sentiment/
[4] http://www.cs.pitt.edu/mpqa/

Table 2. Basic information for the final datasets in the experiments.

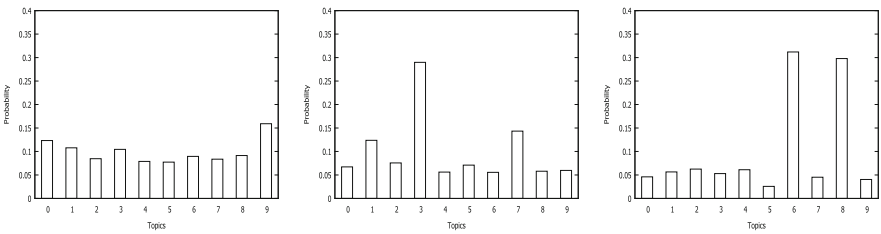
Dataset	# Docs	# Links	# Users
EnronFourUsrs	3804	38597	5623
arnold-j	2441	11474	2550
twitter	2247	3459	3460

executed 500 iterations to estimate the parameters in the models. The datasets are divided into two parts, 80 % of which are used for model training, and the rest are considered as held-out test set.

4.2 Analysis for Distributions Within Communities

In our model, each community has multiple topics, and each topic has multiple sentiment polarities, we studied the distributions within communities on different datasets.

Figure 2 gives the distribution of topics in individual communities. It can be seen from Fig. 2(a) that the topics are almost even within a single community 9 on Enron dataset. We also report selected communities on twitter dataset, in Fig. 2(b) and 2(c), some topics are dominant obviously in the communities. In Fig. 2(b), topic 3 (google android) is the dominant topic in community 1. In community 13, topic 6 (apple use) and topic 8 (iphone service) have large proportions, which are all the subtopics of "apple". These distributions imply that in some communities, people are only very interested in certain number of topics, which is in accordance with our main goal and community definition.



(a) Distribution of topics in community 9 of Enron dataset.

(b) Distribution of topics in community 1 of twitter dataset.

(c) Distribution of topics in community 13 of twitter dataset.

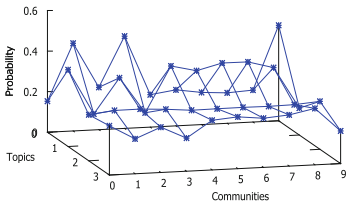
Fig. 2. Distribution of topics in individual communities, $M = 20$, $K = 10$.

Apart from the analysis on the topic distribution within selected individual communities, we also investigated the topic distributions for all the communities, and the sentiment distribution for all the topics in an individual community. Figure 3(a) and 3(b) give the topic and sentiment distributions on twitter

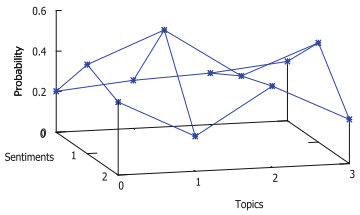
Table 3. Arnold-j's biggest community (community 4), $M = 5$, $K = 10$.

Topic ID	Topic	Positive	Negative	Neutral	people (denoted by the username of the enron email address)
4 (0.1337)	trading	0.3701	0.4498	0.1801	john.arnold (0.3746),
3 (0.1215)	power supply	0.5739	0.2403	0.1858	jennifer.fraser(0.0282),
5 (0.1167)	contract	0.3579	0.3363	0.3058	ina.rangel(0.0217)

dataset, respectively. It is obvious from Fig. 3(a) that different communities have nearly different topic distributions, although some topic distributions for some communities are a bit similar. As can be seen from Fig. 3(b) about the sentiment distribution for topics in community 0 that the sentiments for different topics can be different, which is common in real-world life that two communities may have different sentiment towards certain topics even if they have similar topic distributions (i.e., the two communities are talking about similar range of topics).



(a) Distribution of topics in all communities for twitter dataset.



(b) Distribution of sentiments of all topics in community 0 for twitter dataset.

Fig. 3. Distribution of topics within communities (sentiments for topics) for twitter dataset, $M = 10$, $K = 4$.

4.3 Community Analysis on Individual Users

We also studied the communities for a single user, arnold-j (*John Arnold*, a vice president in Enron company). Table 3 lists the largest community membership (community 4) for arnold-j, Column 1 and 2 show the main relevant topics and the corresponding probabilities within this community, columns 3–5 list the sentiment proportions for the corresponding topics, and the final column represents the top three active persons with high likelihoods in this community. It is obvious from Table 3 that the dominant sentiment polarity can vary with topics. Also we can see that *John Arnold* is the core people in this community.

In twitter dataset, we choose one entity with the screen name '@Apple' to study the hidden knowledge in its community. Table 4 shows the selected communities and sentiment topics that @Apple related to. Column 1 gives three selected participated communities, column 2 and 3 list the top two mainly discussed topics for each community with proportions, and the last three columns

Table 4. Selected communities of the user @Apple (ScreenName), $M = 20$, $K = 10$.

Community	Topic ID	Topic	Positive	Negative	Neutral
9	6 (0.3075)	iphone service	0.9152	0.0492	0.0356
	8 (0.2967)	apple use	0.9398	0.0335	0.0267
10	3 (0.2895)	google android	0.8445	0.0618	0.0937
	1 (0.1327)	twitter operation	0.6029	0.1972	0.1999
5	7 (0.1373)	microsoft	0.1595	0.7182	0.1223
	2 (0.1315)	twitter share	0.6311	0.2307	0.1382

describe the sentiment proportions for the corresponding topics. It is obvious from Table 4 that the mainly discussed topics among communities are different, which demonstrates that community 9, 10 and 5 are well identified, and also proves the effectiveness and feasibility of our model.

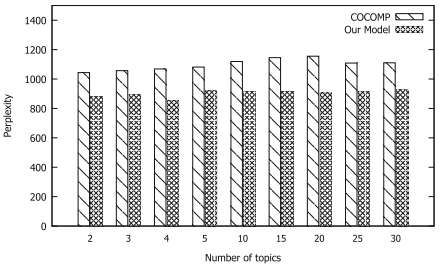
Based on the topics listed in Table 4, we show the top five words for each sentiment polarities of *topic* 1 and *topic* 6 in Table 5, each column lists a collection of highly ranked sentiment words and topic words. From these words, we can observe that topic 1 is about *twitter*, and topic 6 is about *apple*. It's a first attempt to detect sentiment-topic level communities via our STC model, while the sentiment information cannot be detected by the existing COCOMP model.

Table 5. Top ranked words for selected topics with different sentiments extracted by STC model.

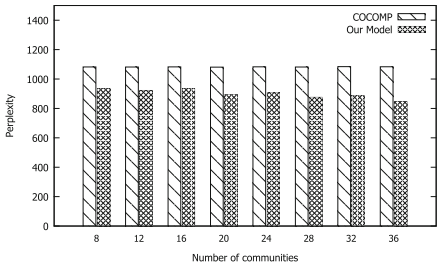
Topic 1 (Twitter Operation)			Topic 6 (Apple Use)		
Positive	Negative	Neutral	Positive	Negative	Neutral
twitter	wrong	yeah	appl	account	touch
win	poor	custom	steve	site	babi
tech	troubl	absolut	job	close	player
world	mark	move	great	longer	feel
good	damag	launch	love	brand	report

4.4 Comparing with COCOMP Model

Note that the ground-truth communities are usually unavailable, which make the evaluation challenging. To evaluate our model, we also analysed the perplexity value, and made comparison with the state-of-the-art COCOMP model [20], which is a topic-level community discovery model. Each word in our model is determined by two factors, namely topic and sentiment, while there is only one factor, topic, for the COCOMP model. In our STC model, to generate a target word, both the topic and sentiment should be correctly assigned, otherwise the perplexity value will get worse, while only a correct topic assignment is required



(a) Perplexity under varying number of topics, $M = 20$.



(b) Perplexity under varying number of communities, $K = 5$.

Fig. 4. Perplexity results comparison between COCOMP and our model for twitter dataset.

in COCOMP model. The computation equations for the perplexity of our model is shown in Eq. 4. The lower perplexity tends to have the better performance.

$$Perplexity(D_{test}) = \frac{\sum_{m=1}^{M} \log P(\tilde{\mathbf{w}}_m|\mathbf{w})}{\sum_{m=1}^{M} n_m}. \tag{4}$$

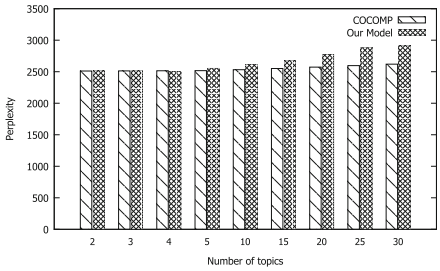
$$P(\tilde{\mathbf{w}}_m|\mathbf{w}) = \prod_{n=1}^{n_m} \sum_{k=1}^{K} \sum_{s=1}^{S} P(w_n = t|z_n = k, l_n = s)\, P(l_n = s|z_n = k, c_{w_n} = m) P(z_n = k|c_{w_n} = m) \tag{5}$$

$$= \prod_{t=1}^{V} \left(\sum_{k=1}^{K} \sum_{s=1}^{S} \phi_{k,s,t} \pi_{m,k,s} \theta_{m,k} \right)^{n_m^{(t)}}.$$

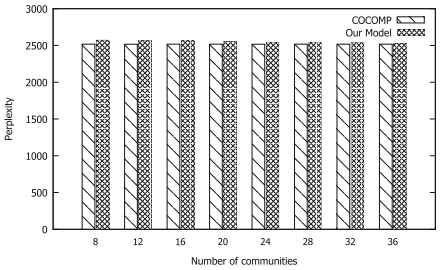
$$\log P(\tilde{\mathbf{w}}_m|\mathbf{w}) = \sum_{t=1}^{V} n_m^{(t)} \log(\sum_{k=1}^{K} \sum_{s=1}^{S} \phi_{k,s,t} \pi_{m,k,s} \theta_{m,k}). \tag{6}$$

In Eq. 4, D_{test} shows the held-out testing documents, $\tilde{\mathbf{w}}_m$ denotes the words from testing documents appeared in community m, $\mathbf{w}$ represents the words in the training documents. n_m is the number of words in community m. As for Eq. 5, $n_m^{(t)}$ is the number of times a term t observed in community m, and c_{w_n} represents the community that the word w_n appears in.

The perplexity results for the two datasets are shown in Figs. 4 and 5. In each figure we illustrated the values of perplexity for our STC model and COCOMP with varying number of topics and communities. As can be seen from Fig. 4(a) and 4(b), the perplexity values of our model are lower than the COCOMP model. Although in Fig. 5(a) and 5(b), the perplexity value are worse than the COCOMP to some extent, it is still comparable to the COCOMP. Enron email and Twitter are two different types of social networking sites, the former is more



(a) Perplexity under varying number of topics, $M = 20$.



(b) Perplexity under varying number of communities, $K = 5$.

Fig. 5. Perplexity results comparison between COCOMP and our model for Enron dataset.

formal than the latter. Generally, there are more sentiment information in tweets than in emails. It is not the main concerning about which model has better perplexity value as long as our model has closer performance with COCOMP. Our model is proposed to identify sentiment level communities, which is not considered by COCOMP and other community discovery methods.

5 Discussions

We build our community discovery model, STC, by using social links, topics and sentiment information in a unified way. Those three factors are very significant to the identification of the meaningful community structures. However, it is not indicating that the more additional information incorporated into the model, the better result we can get. When the information is not important, the redundant factors can make the model more complex and inefficient. Not all the communities have sentiment information, our model is proposed to identify communities that have a certain degree of sentiment polarities.

6 Conclusion and Future Work

Discovering communities from networks has been widely studied in recent years, which can help us to understand the latent knowledge and distributions within them. In this paper, we propose a novel community discovery model, STC, to explore communities with different topic-sentiment distributions. This model is built by combining content, links and sentiment words seamlessly, which can identify communities in a level of sentiment analysis. While most of existing methods for community identification fail to consider the valuable sentiment factor in the networks. Experimental results validated on two types of real-world datasets show that our model can detect sentiment-level communities and can achieve comparable performance, which might be applicable for the opinion analysis and decision making in large business and marketing service.

There are several future extensions to investigate for this work. The topic and sentiment words in our experiment are mixed together, it is interesting to separate them. In addition, discovering communities which have obvious sentiment differences on a certain topic is also very useful. Another direction is to investigate the evolution of communities with the change of users' sentiment topics.

References

1. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Girvan, M., Newman, M.: Community structure in social and biological networks. PNAS **99**(12), 7821–7826 (2002)
3. Kim, M., Han, J.: A particle-and-density based evolutionary clustering method for dynamic networks. VLDB Endowment **2**(1), 622–633 (2009)
4. Lancichinetti, A., Radicchi, F., Ramasco, J., Fortunato, S.: Finding statistically significant communities in networks. PloS One **6**(4), e18961 (2011)
5. Li, F., Huang, M., Zhu, X.: Sentiment analysis with global topics and local dependency. In: AAAI, pp. 1371–1376 (2010)
6. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: CIKM, pp. 375–384 (2009)
7. McCallum, A., Wang, X., Corrada-Emmanuel, A.: Topic and role discovery in social networks with experiments on enron and academic email. J. Artif. Intell. Res. **30**(1), 249–272 (2007)
8. Natarajan, N., Sen, P., Chaoji, V.: Community detection in content-sharing social networks. In: ASONAM, pp. 82–89 (2013)
9. Newman, M., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**(2), 026113 (2004)
10. Palla, G., Barabasi, A., Vicsek, T.: Quantifying social group evolution. Nature **446**(7136), 664–667 (2007)
11. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**, 814–818 (2005)
12. Pathak, N., DeLong, C., Banerjee, A., Erickson, K.: Social topic models for community extraction. In: The 2nd SNA-KDD Workshop, vol. 8 (2008)
13. Ruan, Y., Fuhry, D., Parthasarathy, S.: Efficient community detection in large networks using content and links. In: WWW, pp. 1089–1098 (2013)
14. Sachan, M., Contractor, D., Faruquie, T., Subramaniam, L.: Using content and interactions for discovering communities in social networks. In: WWW, pp. 331–340 (2012)
15. Tyler, J., Wilkinson, D., Huberman, B.: Email as spectroscopy: automated discovery of community structure within organizations. In: Communities and Technologies, pp. 81–96 (2003)
16. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: HLT-EMNLP, pp. 347–354 (2005)
17. Xie, J., Szymanski, B., Liu, X.: Slpa: uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In: ICDM Workshops, pp. 344–349 (2011)
18. Yang, T., Jin, R., Chi, Y., Zhu, S.: Combining link and content for community detection: a discriminative approach. In: KDD, pp. 927–936 (2009)

19. Zhou, D., Manavoglu, E., Li, J., Giles, C., Zha, H.: Probabilistic models for discovering e-communities. In: WWW, pp. 173–182 (2006)
20. Zhou, W., Jin, H., Liu, Y.: Community discovery and profiling with social messages. In: KDD, pp. 388–396 (2012)