

Singapore Management University Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems

School of Information Systems

7-2012

Topic discovery from tweet replies

Bingtian DAI

Singapore Management University, btdai@smu.edu.sg


Ee Peng LIM

Singapore Management University, eplim@smu.edu.sg

Philips Kokoh PRASETYO

Singapore Management University, pprasetyo@smu.edu.sg

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the [Databases and Information Systems Commons](#), [Numerical Analysis and Scientific Computing Commons](#), and the [Social Media Commons](#)

Citation

DAI, Bingtian; LIM, Ee Peng; and PRASETYO, Philips Kokoh. Topic discovery from tweet replies. (2012). *Proceedings of the 10th Workshop on Mining and Learning with Graphs (MLG-2012), Edinburgh*. Research Collection School Of Information Systems.

Available at: https://ink.library.smu.edu.sg/sis_research/3160

This Conference Proceeding Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

7-2012

Topic Discovery from Tweet Replies

Bingtian DAI

Singapore Management University, btdai@smu.edu.sg

Ee Peng LIM

Singapore Management University, eplim@smu.edu.sg

Philips Kokoh Prasetyo

Singapore Management University, pprasetyo@smu.edu.sg

Follow this and additional works at: http://ink.library.smu.edu.sg/sis_research_smu

Citation

DAI, Bingtian; LIM, Ee Peng; and Prasetyo, Philips Kokoh, "Topic Discovery from Tweet Replies" (2012). *Research Collection School Of Information Systems (SMU Access Only)*. Paper 21.

http://ink.library.smu.edu.sg/sis_research_smu/21

Available at: http://ink.library.smu.edu.sg/sis_research_smu/21

This Conference Proceedings Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems (SMU Access Only) by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

Topic Discovery from Tweet Replies

Bing Tian Dai

Living Analytics Research Centre, Singapore Management University

BTDAI@SMU.EDU.SG

Ee-Peng Lim

Living Analytics Research Centre, Singapore Management University

EPLIM@SMU.EDU.SG

Philips Kokoh Prasetyo

Living Analytics Research Centre, Singapore Management University

PPRASETYO@SMU.EDU.SG

Abstract

Twitter¹ is a popular online social information network service which allows people to read and post messages up to 140 characters, known as “tweets”. In this paper, we focus on the tweets between pairs of individuals, i.e., the tweet replies, and propose a generative model to discover topics among groups of twitter users. Our model has then been evaluated with a tweet dataset to show its effectiveness.

1. Introduction

Twitter has become popular over the recent years, since it has provided a platform for individuals to share their moods and opinions openly on the Internet. Mining Twitter users and their tweets has attracted much interests from the social networks and data mining research communities (Weng et al., 2010; O’Connor et al., 2010; Michelson & Macskassy, 2010). In particular, some people study Twitter users’ topical interests by modeling the set of tweets they post. For example, if someone likes gadgets, he or she would probably tweets or re-tweets anything that is related to new IT gadgets and their reviews. His topical interests can thus be summarized from the set of tweets and retweets generated by him.

Unless a user protects his Twitter account, his tweets, retweets and the associated topical interests are often general and public. These topics are known to repre-

¹<http://twitter.com>

Appearing in *Proceedings of the Workshop on Mining and Learning with Graphs (MLG-2012)*, Edinburgh, Scotland, UK, 2012. Copyright 2012 by the author(s)/owner(s).

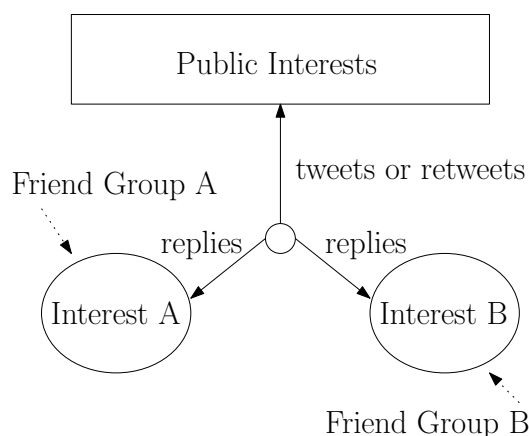


Figure 1. Public interests are discovered from the tweets and the retweets, more specific interests are discovered from the replies

sent the user’s *public interests*. There are, on the other hand, some topics that a user may only wish to share with a certain friend or a certain circle of friends. As shown in Figure 1, a user represented by the center node has public interests discovered from his tweets and retweets. He however may have the other two interests, which may show up only when he interacts with the two friend groups respectively.

Modeling topical interests is a challenging task from tweets and retweets. Consider the previous example, someone who has strong interests in IT gadgets may be an expert working in the IT sector. He may discuss with another IT expert or his colleague more about the features of the IT gadgets, which are likely to involve some technical details. This topical interest is more specific than the public interest. On the other hand, he may be a big fan of wines, but he does not share this openly by tweeting about wines. Instead, he

has a circle of friends who are also interested in wines. When one of his friends tweets about wines, he may gladly join the discussion. This demonstrates a different topical interests summarized from tweets with some circle of friends.

Therefore, to capture users' topical interests from just tweets and re-tweets is not sufficient. In **Twitter**, it happens often that a tweet from one's friend triggers a list of replies on a different topic. Therefore, tweet replies among users can provide more specific topical interests, or topics different from their public interests.

In this paper, we focus on discovering topics from tweet replies among **Twitter** users, as to contribute a brand new approach to study personal topical interests. There are at least two major challenges to discover topical interests from tweet replies.

First, among **Twitter** users, as the number of pairs is much more than the number of users themselves, grouping of users is necessary to describe the topics discovered from pairs of users. In conventional topic discovery, users are grouped based on their public interests, i.e., users from the same group are expected to have similar topical interests. Grouping of users on their pairwise topical interests is similar but more complicated. Given user A_1, A_2 from group A and user B_1, B_2 from group B , the topical interests between pair $\langle A_1, B_1 \rangle$ and pair $\langle A_2, B_2 \rangle$ are expected to be similar.

On the other hand, the grouping of users and topic discovery are correlated. If we assign the topical interests from one group to another, the change in a user's group membership has an impact on his friends' membership as they are related by the topics of their tweet replies. Grouping of users can be viewed as a compression from users to groups, whereas topic discovery can be considered as a summary from tweet replies to topics. They are expected to be modeled simultaneously, which is our second major challenge.

To solve the two major challenges, we will propose a model called **Uni-Topical Blockmodels (UTBM)**, as our paper's main contribution. In Section 2, we will give an overview on topic modeling, and topical blockmodels which our UTBM is built upon. Next, we will elaborate why we consider the uni-topic models, i.e., the mixture of unigrams, and propose our UTBM in Section 3. Technical details about UTBM will then be described in Section 4. Experimental studies in Section 5 then evaluate the UTBM model using a set of tweet replies we collected. The discovered topics are validated with general ground truth. Last but not least, we will conclude our model and discuss about its

generality in Section 6.

2. Related Works

Topic discovery has been well studied in the literature (Hofmann, 1999; Newman et al., 2009). One of the most well-known work is Latent Dirichlet Allocation (LDA), proposed by Blei, Ng and Jordan (Blei et al., 2003). To generate a document with LDA model, a multinomial distribution over the topics is first generated by a Dirichlet prior. For each word in the documents, it generates a topic according to the multinomial distribution, and then generates a word according to that generated topic. LDA has been shown to be effective on discovering topics from documents with multiple topics, e.g., news articles. However, tweets are short, i.e., less than 140 characters, which makes people wonder if LDA works on such short documents. Labeled LDA (Ramage et al., 2010) extended the original LDA by incorporating supervisions in order to deal with short and informal documents like tweets.

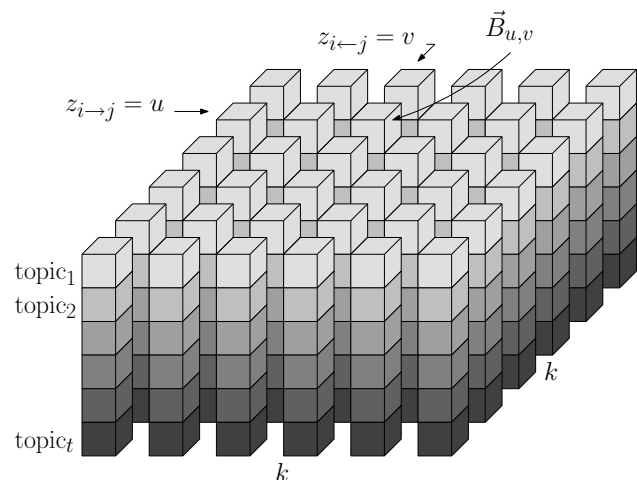


Figure 2. An illustration on topical GSBM

As tweet replies are between two **Twitter** users, it is thus necessary to model the relationships between pairs of individuals. Blockmodels (Wasserman & Faust, 1994; Doreian et al., 2005) have been proposed to study pairwise relationships in sociology. (Airoldi et al., 2008) proposed the mixed membership stochastic blockmodels (MMB) to model the existence of links between two individuals. The generalized stochastic blockmodels (GSBM) was then proposed by (Dai et al., 2012) to model relationships between two individuals. Unlike MMB, GSBM handles multiple relations simultaneously. These relations can be social connections or sets of interactions, and they are possibly correlated.

In GSBM, relationships from individuals in one block to individuals in another block are generated by the same multivariate probability distribution function associated to the role block.

For tweets, we can certainly model each word as an individual relation, and then apply GSBM directly. However, the number of words are in thousands, making it infeasible to compute a model within a reasonable time frame. Therefore, words have to be summarized to the topic level.

As shown in Figure 2, if each topic is regarded as a relation in GSBM, we get a topical GSBM, where each role block is described by parameters associated to topics. For example, when user x_i talks to user x_j , and they take positions u and v respectively, the topics between x_i and x_j is modeled by a probability distribution on topics, which is associated to the role block $\vec{B}_{u,v}$.

Note that GSBM can handle asymmetrical relationships, i.e., the relationship from x_i to x_j can be different from the relationship from x_j to x_i . In topic modeling, the topics between two individuals are generally symmetrical, therefore, a topical GSBM is symmetrical, i.e., $\vec{B}_{u,v} = \vec{B}_{v,u}$. The model we are going to propose in Section 3 makes use of such topical GSBM to tackle the first challenge discussed in Section 1.

Prior to our work, McCallum, Wang and Corrada-Emmanuel (McCallum et al., 2007) have proposed the Author-Recipient-Topic (ART) model to learn topics between pairs of individuals in an email network. In ART model, a Dirichlet prior generates a topic distribution for each pair of individuals. The words in the emails between a pair of individuals are then determined by their topic distribution, like what LDA does for a document. As an extension of ART model, they have also proposed the Role-Author-Recipient-Topic (RART) model which makes use of blockmodels to study roles and topics simultaneously. However, we cannot compare our model with RART as there is no implementation given for RART.

3. Uni-Topical Blockmodels (UTBM)

3.1. Mixtures of Unigrams

Tweets are short, since they are restricted to 140 characters. One question to ask is that, if there exist multiple topics in a tweet? The advantage of LDA (Blei et al., 2003) over its earlier techniques, e.g., the mixture of unigrams model, is that it models multiple topics in one document. This is reasonable since a news article may drift from one topic to another between sentences. However, such drift in topics is hardly ob-

served in tweets, as they are too short to cover more than one topic. Hence we are going to adopt a uni-topical model which assumes only one topic in each tweet.

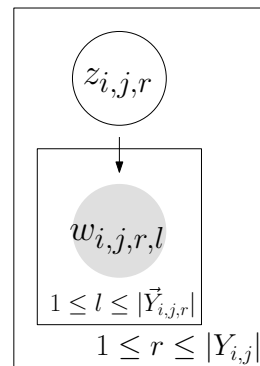


Figure 3. Mixture of Unigrams

Before we described our proposed model, we first present an earlier uni-topical model that assumes one topic in each document. The mixture of unigrams model, proposed by Nigam *et al.* (Nigam et al., 2000), considers just one topic for each document, and the words in that document are all generated with the same topic. We illustrate the mixture of unigrams model in Figure 3.

Let x_i and x_j be two **Twitter** users, and $Y_{i,j}$ be the set of tweet replies between x_i and x_j , i.e., each $Y_{i,j,r}$ is a tweet reply, where $1 \leq r \leq |Y_{i,j}|$. There is then one and only one topic $z_{i,j,r}$ associated to the tweet reply $Y_{i,j,r}$, which is a sequence of words of length $|Y_{i,j,r}|$. Thus, for each $1 \leq l \leq |Y_{i,j,r}|$, topic $z_{i,j,r}$ generates a word $w_{i,j,r,l}$, which is observed.

With the mixture of unigrams model, the probability of observing the tweet reply $Y_{i,j,r}$ is

$$p(Y_{i,j,r}) = \sum_{t=1}^T p(z_{i,j,r} = t) \prod_{l=1}^{|Y_{i,j,r}|} p(w_{i,j,r,l} | z_{i,j,r} = t)$$

Next, we see how to combine the mixture of unigrams model with the topical GSBM aforementioned in Section 2 to discover topics in tweet replies.

3.2. Combining with Topical GSBM

In this paper, we propose *Uni-Topical Blockmodels (UTBM)* as a combination of the topical blockmodels, i.e., topical GSBM, and the mixture of unigrams models, to discover topics in tweet replies. The UTBM model is shown in Figure 4. There are three model parameters, defined below.

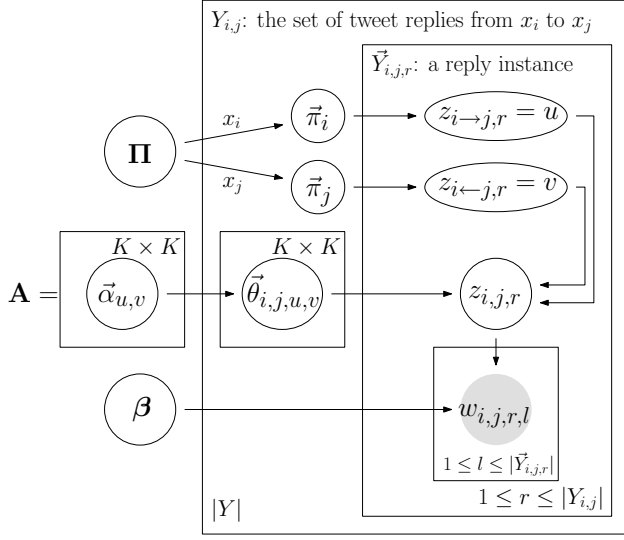


Figure 4. Uni-Topical Blockmodels (UTBM)

- $\Pi = \{\vec{\pi}_i : 1 \leq i \leq N\}$. Each $\vec{\pi}_i$ is a K -dimensional vector, representing the membership distribution of individual x_i over the K blocks.
- $\mathbf{A} = \{\vec{\alpha}_{u,v} : 1 \leq u, v \leq K\}$. Each $\vec{\alpha}_{u,v}$ is a T -dimensional Dirichlet prior, associated to the role block (u, v) , where T is the number of topics.
- $\beta = \{\beta_{t,m} : 1 \leq t \leq T, 1 \leq m \leq M\}$. The t^{th} row vector of β , $\vec{\beta}_t$, is the word probabilistic distribution for the t^{th} topic. M is the number of distinct words in all tweet replies.

In UTBM, there is one Dirichlet prior associated with each role block. For role block (u, v) , the Dirichlet prior $\vec{\alpha}_{u,v}$ determines the topic distributions when the two individuals take the positions u and v respectively. The generative process for the set of tweet replies between x_i and x_j , i.e. $Y_{i,j}$, is as follows:

1. Generate the topic distribution $\vec{\theta}_{i,j,u,v} \sim \text{Dir}(\vec{\alpha}_{u,v})$ for x_i taking position u and x_j taking position v .
2. For each tweet reply $\vec{Y}_{i,j,r} \in Y_{i,j}$:
 - (a) Generate latent variable $z_{i \rightarrow j, r}$ as the position of x_i , and $z_{i \leftarrow j, r}$ as the position of x_j , from $\vec{\pi}_i$ and $\vec{\pi}_j$ respectively. Let $z_{i \rightarrow j, r} = u$ and $z_{i \leftarrow j, r} = v$.
 - (b) Generate a topic $z_{i,j,r}$ by the topic distribution $\vec{\theta}_{i,j,u,v}$ since x_i and x_j take positions u and v when generating this reply, i.e., $z_{i,j,r} \sim \text{Multinomial}(\vec{\theta}_{i,j,u,v})$.

- (c) Generate word $w_{i,j,r,l}$ from $\vec{\beta}_{z_{i,j,r}}$, the multinomial probability distribution conditioned on the topic $z_{i,j,r}$, for all $1 \leq l \leq |\vec{Y}_{i,j,r}|$.

Note that the generations for the number of tweet replies $|Y_{i,j}|$ and the length of a tweet reply $|\vec{Y}_{i,j,r}|$ can be modeled by any reasonable discrete probability distributions, e.g., Poisson distributions.

We denote the model parameters by \mathcal{M} , i.e., $\mathcal{M} = \{\Pi, \mathbf{A}, \beta\}$, and the observed data in this model are the set of tweet replies between pairs of individuals, i.e., $\mathcal{D} = \{Y_{i,j} : i \neq j\}$. The log likelihood of model \mathcal{M} given data \mathcal{D} is therefore

$$l(\mathcal{M}|\mathcal{D}) \triangleq \sum_{Y_{i,j} \in \mathcal{D}} l(\mathcal{M}|Y_{i,j})$$

In the UTBM model diagram in Figure 4, there are four latent variables, $\theta_{i,j,u,v}$ for each $Y_{i,j}$ and $Z_{i,j,r} = \{z_{i \rightarrow j, r}, z_{i \leftarrow j, r}, z_{i,j,r}\}$ for each $\vec{Y}_{i,j,r}$. The log likelihood of \mathcal{M} given each observed set of tweet replies $Y_{i,j}$ is

$$\begin{aligned} l(\mathcal{M}|Y_{i,j}) &\triangleq \log p(Y_{i,j}|\mathcal{M}) \\ &= \log \int_{\theta_{i,j}} \sum_{Z_{i,j}} p(Y_{i,j}, \theta_{i,j}, Z_{i,j}|\mathcal{M}) d\theta_{i,j} \end{aligned} \quad (1)$$

where $\theta_{i,j} = \{\vec{\theta}_{i,j,u,v} : 1 \leq u, v \leq K\}$ and $Z_{i,j} = \{Z_{i,j,r} : 1 \leq r \leq |Y_{i,j}|\}$. By the dependencies among the variables in Figure 4, the probability can be further factorized into²

$$\begin{aligned} &p(Y_{i,j}, \theta, Z|\mathcal{M}) \\ &= p(\theta|\mathcal{M}) \cdot \prod_{r=1}^{|Y_{i,j}|} [p(Z_r|\theta, \mathcal{M}) \cdot p(\vec{Y}_{i,j,r}|Z_r, \mathcal{M})] \\ &= p(\theta|\mathcal{M}) \cdot \prod_{r=1}^{|Y_{i,j}|} [p(z_{\rightarrow, r}|\mathcal{M}) \cdot p(z_{\leftarrow, r}|\mathcal{M}) \\ &\quad \cdot p(z_r|z_{\rightarrow, r}, z_{\leftarrow, r}, \theta, \mathcal{M}) \cdot p(\vec{Y}_{i,j,r}|z_r, \mathcal{M})] \end{aligned}$$

However, as z_r is dependent on other latent variables, there is no closed form for the log likelihood in Equation 1. In Section 4, a variational model will be presented to learn the UTBM model.

4. Variational Inference

As shown in Figure 5, the four latent variables in this variational model are separated and controlled by different variational parameters. For a pair of individuals, x_i and x_j , each topic distribution $\vec{\theta}_{u,v}$

²Indices i and j are dropped to save space.

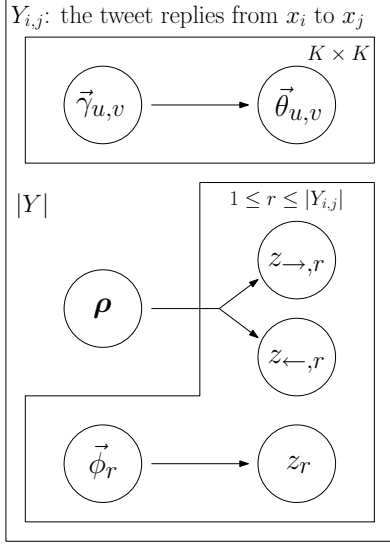


Figure 5. Graphical model representation of the variational inference

($1 \leq u, v \leq K$) for tweet replies when x_i and x_j take positions u and v respectively, is determined by an individual Dirichlet prior $\vec{\gamma}_{u,v}$. For each tweet reply $\vec{Y}_{i,j,r}$ from x_i to x_j , ρ first determines $z_{\rightarrow,r}$ and $z_{\leftarrow,r}$ simultaneously, and then $\vec{\phi}_r$ determines z_r , the topic for all words $w_{r,l} \in \vec{Y}_{i,j,r}$. Let $\mathcal{M}'_{i,j}$ denote the variational model, i.e., $\mathcal{M}'_{i,j} = \{\vec{\gamma}_{u,v}, \rho, \vec{\phi}_r : Y_{i,j} \in Y, 1 \leq u, v \leq K, 1 \leq r \leq |Y_{i,j}|\}$; and the latent variables follow a distribution $q(\theta, \mathbf{Z} | \mathcal{M}'_{i,j})$. The variational inference is elaborated as below. By Equation 1,

$$\begin{aligned} \log p(Y_{i,j} | \mathcal{M}) &= \log \int_{\theta} \sum_{\mathbf{Z}} p(Y_{i,j}, \theta, \mathbf{Z} | \mathcal{M}) d\theta \\ &= \log \int_{\theta} \sum_{\mathbf{Z}} q(\theta, \mathbf{Z} | \mathcal{M}'_{i,j}) \frac{p(Y_{i,j}, \theta, \mathbf{Z} | \mathcal{M})}{q(\theta, \mathbf{Z} | \mathcal{M}'_{i,j})} d\theta \\ &\geq \int_{\theta} \sum_{\mathbf{Z}} q(\theta, \mathbf{Z} | \mathcal{M}'_{i,j}) \log p(Y_{i,j}, \theta, \mathbf{Z} | \mathcal{M}) d\theta \\ &\quad - \int_{\theta} \sum_{\mathbf{Z}} q(\theta, \mathbf{Z} | \mathcal{M}'_{i,j}) \log q(\theta, \mathbf{Z} | \mathcal{M}'_{i,j}) d\theta \\ &= \mathbb{E}_q[\log p(Y_{i,j}, \theta, \mathbf{Z} | \mathcal{M})] - \mathbb{E}_q[\log q(\theta, \mathbf{Z} | \mathcal{M}'_{i,j})] \end{aligned}$$

We denote the last line of the above equation as auxiliary function $\mathcal{L}(\vec{Y}_{i,j}, \mathcal{M}'_{i,j}, \mathcal{M})$, i.e.,

$$\mathcal{L}(\vec{Y}_{i,j}, \mathcal{M}'_{i,j}, \mathcal{M}) = \mathbb{E}_q[\log p(Y_{i,j}, \theta, \mathbf{Z} | \mathcal{M})] - \mathbb{E}_q[\log q(\theta, \mathbf{Z} | \mathcal{M}'_{i,j})] \quad (2)$$

As latent variables $\vec{\theta}_{u,v}$, $(z_{\rightarrow,r}, z_{\leftarrow,r})$ and z_r are conditionally independent given $\mathcal{M}'_{i,j}$, thus by expanding

Equation 2,

$$\begin{aligned} &\mathcal{L}(Y_{i,j}, \mathcal{M}'_{i,j}, \mathcal{M}) \quad (3) \\ &= \sum_{u,v} \left\{ \log \Gamma\left(\sum_{t=1}^T \alpha_{u,v,t}\right) - \sum_{t=1}^T \log \Gamma(\alpha_{u,v,t}) \right. \\ &\quad \left. + \sum_{t=1}^T (\alpha_{u,v,t} - 1) [\Psi(\gamma_{u,v,t}) - \Psi(\sum_{t'=1}^T \gamma_{u,v,t'})] \right\} \\ &+ |Y_{i,j}| \sum_{u=1}^K \left(\sum_{v=1}^K \rho_{u,v} \right) \log \pi_{i,u} + |Y_{i,j}| \sum_{v=1}^K \left(\sum_{u=1}^K \rho_{u,v} \right) \log \pi_{j,v} \\ &+ \sum_{u,v} \rho_{u,v} \sum_{t=1}^T [\Psi(\gamma_{u,v,t}) - \Psi(\sum_{t'=1}^T \gamma_{u,v,t'})] \sum_{r=1}^{|Y_{i,j}|} \phi_{r,t} \\ &+ \sum_{r=1}^{|Y_{i,j}|} \sum_{t=1}^T \phi_{r,t} \sum_{l=1}^{|\vec{Y}_{i,j,r}|} \log \beta_{t,w_{r,l}} \\ &- \sum_{u,v} \left\{ \log \Gamma\left(\sum_{t=1}^T \gamma_{u,v,t}\right) - \sum_{t=1}^T \log \Gamma(\gamma_{u,v,t}) \right. \\ &\quad \left. + \sum_{t=1}^T (\gamma_{u,v,t} - 1) [\Psi(\gamma_{u,v,t}) - \Psi(\sum_{t'=1}^T \gamma_{u,v,t'})] \right\} \\ &- |Y_{i,j}| \sum_{u,v} \rho_{u,v} \log \rho_{u,v} - \sum_{t=1}^T \sum_{r=1}^{|Y_{i,j}|} \phi_{r,t} \log \phi_{r,t} \end{aligned}$$

By solving Equation 3 with respect to variational model $\mathcal{M}'_{i,j}$, we have

$$\begin{aligned} \rho_{u,v} &\propto \pi_{i,u} \pi_{j,v} \exp\left[\frac{1}{|Y_{i,j}|} \sum_{t=1}^T \Psi(\gamma_{u,v,t}) \sum_{r=1}^{|Y_{i,j}|} \phi_{r,t}\right] \\ \gamma_{u,v,t} &= \alpha_{u,v,t} + \rho_{u,v} \sum_{r=1}^{|Y_{i,j}|} \phi_{r,t} \\ \phi_{r,t} &\propto \prod_{l=1}^{|\vec{Y}_{i,j,r}|} \beta_{t,w_{r,l}} \exp\left\{\sum_{u,v} \rho_{u,v} [\Psi(\gamma_{u,v,t}) - \Psi(\sum_{t'=1}^T \gamma_{u,v,t'})]\right\} \end{aligned}$$

Variational inference iterates through the above three equations, and the variational parameters for all $Y_{i,j} \in Y$ then update UTBM model parameters as

$$\begin{aligned} \pi_{i,u} &\propto \sum_{Y_{i,j} \in Y} |Y_{i,j}| \sum_{v=1}^K \rho_{i,j,u,v} + \sum_{Y_{j,i} \in Y} |Y_{j,i}| \sum_{v=1}^K \rho_{j,i,v,u} \\ \beta_{t,m} &\propto \sum_{Y_{i,j} \in Y} \sum_{r=1}^{|Y_{i,j}|} \phi_{i,j,r,t} \sum_{l=1}^{|\vec{Y}_{i,j,r}|} \delta(w_{i,j,r,l}, w_m) \end{aligned}$$

where $\delta(w_{i,j,r,l}, w_m) = 1$ if $w_{i,j,r,l} = w_m$, and $\delta(w_{i,j,r,l}, w_m) = 0$ otherwise. The updating formulae

for each $\vec{\alpha}_{u,v}$ are similar to the $\vec{\alpha}$ updating formulae in (Blei et al., 2003).

$$\begin{aligned} \alpha_{u,v,t,\text{new}} &= \alpha_{u,v,t,\text{old}} - \frac{g_{u,v,t} - c_{u,v}}{h_{u,v,t}} \\ g_{u,v,t} &= \sum_{Y_{i,j} \in Y} [\Psi(\sum_{t'=1}^T \alpha_{u,v,t'}) - \Psi(\alpha_{u,v,t}) \\ &\quad + \Psi(\gamma_{i,j,u,v,t}) - \Psi(\sum_{t'=1}^T \gamma_{i,j,u,v,t'})] \\ h_{u,v,t} &= - \sum_{Y_{i,j} \in Y} \Psi'(\alpha_{u,v,t}) \\ c_{u,v} &= \frac{\sum_{t=1}^T \frac{g_t}{h_t}}{[\sum_{Y_{i,j} \in Y} \Psi'(\sum_{t=1}^T \alpha_{u,v,t})]^{-1} + \sum_{t=1}^T h_t^{-1}} \end{aligned}$$

In the next section, we will evaluate our UTBM model.

5. Experiments

5.1. Dataset Description

We collected our tweets through Palanteer³, a service that allows users to search for socio-political tweets generated by a set of **Twitter** users in Singapore. Palanteer starts from 69 seed **Twitter** users who are interested in current affairs, including political candidates, political parties or organizations, journalists, and bloggers. The set of **Twitter** users is further expanded by following the incoming and outgoing follow-links from the seed users. This expansion is done for two times, i.e., all users are at most 2-hops away from the 69 seed users. Note that only users who explicitly specify their location as Singapore in their profiles are included. In this experiment, we used the data collected by Palanteer in May 2011.

In tweets, words prefixed by “#” are called *hashtags*, which serve as keywords for the tweets. As there are many spam tweets containing a single hashtag only but no other text, we filtered away such tweets. The correlations among hashtags are still preserved as tweets with at least two hashtags are still included in the dataset. The filtered dataset contains 4,566 users with 6,378 tweet replies among them, and 5,672 hashtags.

There are two factors that make it challenging to conduct topical discovery on this dataset. Firstly, the social network is relatively sparse as tweet-to-user ratio is only about 1.4. Secondly, unlike the conventional document datasets with many words in each document, most of the tweet messages contain only two hashtags.

³<http://research.larc.smu.edu.sg/palanteer/>, developed by researchers and engineers at Living Analytics Research Centre, Singapore Management University

Nevertheless, we first examine the topics discovered by our UTBM model from this dataset.

5.2. Discovered Topics

There are two parameters in UTBM, the number of blocks (K) and the number of topics (T). With this particular dataset, we run our experiment with $K = 10$, and $T = 20$. A set of hashtags for each of the 20 topics is then given by UTBM. We selected 8 of these topics, gave them a label each, and listed them together with the associated hashtags in Table 1.

As the data was collected in May 2011, when the Singapore General Election took place, there are many tweets about politics, e.g., a tweet with hashtag “#xxrally” is probably sent out when the user is watching a rally organized by party xx. We therefore see the hashtags like “#sgelections”, “#ge2011”, “#wprally” appear together and represent the topic “Politics”.

Another obvious topic is “Zodiac”, where people tweet about some zodiac characteristics. There are 12 hashtags, each for one zodiac. These hashtags often appear together since one tweet may contain a rank list of a few zodiacs, or talk about the match-making among zodiacs. Hence we group the 12 hashtags, together with the two general ones, “#zodiacfacts” and “#zodiaczone”, under the topic of “Zodiac”.

There are several users interested in “Shopping”, “Photography” and “Tennis”. As the great Singapore sale always happens in the mid-year, the tweet replies about shopping contain hashtags like “#greatsingaporeale” and “#gss”. Note that the hashtag “#singapore” is shared among “Shopping”, “Photography” and some other groups of hashtags, as it is a general hashtag, and used with tweets about Singapore. Thus the topics of the hashtag “#singapore” is not clear, and get included by a few topics. This hashtag is popular because of the dataset itself. If we use another dataset, we probably do not see hashtag “#singapore”. There is another special tag “#awesome”, which is also general, but only appears with hashtags under “Tennis”. This is probably due to its high frequency with other tennis hashtags.

Justin Bieber’s fans contribute another topic about Justin Bieber. The hashtags are about Justin Bieber’s songs, e.g., “Never Say Never”, or his concert like “My World Tour”, or some activities like “#10millionbeliebers”.

The last two topics are more about advertisements. Job seekers may send out tweets containing both “#jobs” and the type of the job she wants to do.

Politics	Zodiacs	Shopping	Photography
#sgelections #sgelection #ge2011 #sgpolitics #voteforchange #sgrally #nsprally #wprally	#zodiacfacts #zodiaczone #aquarius #cancer #scorpio #sagittarius #libra #pisces	#greatsingaporesale #singapore #gss #shopping #food #checkitout #sgfood #swagapore	#sgig #iphoneography #singapore #iphoneasia #dogs #foodporn #pets #streetphotography
Tennis	Bieber	Jobs	Follow
#tennis #rolandgarros #ynwa #awesome #frenchopen #federer #wimbledon	#neversaynever #believe #dreambig #beast #10millionbeliebers #myworldtour #beliebers	#jobs #adminjobs #fail #marketing #cs #tech #rapture	#followback #teamfollowback #nowfollowing #follow4follow #follow #followfollowfollow #mustfollow

Table 1. Topics discovered from the tweet replies

We then see a topic with hashtags “#jobs”, “#adminjobs”, “#marketing” and “#tech”. Some people would like to get more people to follow them, so they broadcast with hashtags “#followback”, “#teamfollowback”, “#follow4follow” to ask others to follow them back as a return of following. These people are probably spammers. Therefore, by discovering the topics with the associated hashtags, we are able to identify groups of spam Twitter users by the spam hashtags.

5.3. Discovered Blocks

With the membership probabilistic distribution over the 10 blocks, a partition can then be obtained by assigning each individual to the principle block, i.e., the block with largest membership probability.

We then counted the list the hashtags which are frequently used by members in one block, and we did not discover a set of topics which is as interesting as the one listed in Table 1. Although the two topics, politics and zodiac, indeed show up, the other topics are not obvious. This result shows that, it is not sufficient to discover topics of tweet replies by simply treating them as documents. In the topic discovery of tweet replies, or any other kind of conversations, it is therefore a must to consider both the talking party (or the sender) and the listening party (or the recipient).

6. Conclusion

In this paper, we focus on discovering topics from tweet replies. Our proposed model, the uni-topical Blockmodels, assumes there is only one topic in each tweet, and applies the mixture of unigrams model on the Generalized Stochastic Blockmodels (GSBM). Experiments on a collected tweets dataset show that this model is effective in discover topics from tweet replies. Our model is able to be generalized to other social communications, as long as the assumption that there is only one topic in a message holds. Examples of such communications are instant messaging, short message service (SMS), etc.

Acknowledgments

This work is supported by Singapore’s National Research Foundation’s research grant, NRF2008IDM-IDM004-036.

We appreciate Hanghang’s effort in organizing this workshop and the comments from the reviewers.

References

Airoldi, Edoardo M., Blei, David M., Fienberg, Stephen E., and Xing, Eric P. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

- Dai, Bing Tian, Chua, Freddy Chong Tat, and Lim, Ee-Peng. Structural analysis in multi-relational social networks. In *SDM*, 2012.
- Doreian, Patrick, Batagelj, Vladimir, and Ferligoj, Anuška. *Generalized Blockmodeling*. Cambridge University Press, 2005. ISBN 0521840856.
- Hofmann, Thomas. Probabilistic latent semantic analysis. In *UAI*, pp. 289–296, 1999.
- McCallum, Andrew, Wang, Xuerui, and Corrada-Emmanuel, Andrés. Topic and role discovery in social networks with experiments on enron and academic email. *J. Artif. Intell. Res.*, 30:249–272, 2007.
- Michelson, Matthew and Macskassy, Sofus A. Discovering users’ topics of interest on twitter: a first look. In *AND*, pp. 73–80, 2010.
- Newman, David, Asuncion, Arthur U., Smyth, Padhraic, and Welling, Max. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10:1801–1828, 2009.
- Nigam, Kamal, McCallum, Andrew, Thrun, Sebastian, and Mitchell, Tom M. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2/3):103–134, 2000.
- O’Connor, Brendan, Krieger, Michel, and Ahn, David. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*, 2010.
- Ramage, Daniel, Dumais, Susan T., and Liebling, Daniel J. Characterizing microblogs with topic models. In *ICWSM*, 2010.
- Wasserman, Stanley and Faust, Katherine. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994. ISBN 0521387078.
- Weng, Jianshu, Lim, Ee-Peng, Jiang, Jing, and He, Qi. Twitterank: finding topic-sensitive influential twitterers. In *WSDM*, pp. 261–270, 2010.