Originator or Propagator? Incorporating Social Role Theory into Topic Models for Twitter Content Analysis

Wayne Xin Zhao[1], Jinpeng Wang[1], Yulan He[2], Jian-Yun Nie[3] and Xiaoming Li[1]
[1]School of Electronic Engineering and Computer Science,Peking University, China
[2]School of Engineering & Applied Science, Aston University, UK
[3]Département d'informatique et de recherche opérationnelle,Université de Montréal, Canada
{batmanfly@gmail.com, JooPoo@pku.edu.cn, y.he@cantab.net, nie@iro.umontreal.ca, lxm@pku.edu.cn}

ABSTRACT

A large number of studies have been devoted to modeling the contents and interactions between users on Twitter. In this paper, we propose a method inspired from Social Role Theory (SRT), which assumes that a user behaves differently in different roles in the generation process of Twitter content. We consider the two most distinctive social roles on Twitter: originator and propagator, who respectively posts original messages and retweets or forwards the messages from others. In addition, we also consider role-specific social interactions, especially implicit interactions between users who share some common interests. All the above elements are integrated into a novel regularized topic model. We evaluate the proposed method on real Twitter data. The results show that our method is more effective than the existing ones which do not distinguish social roles.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*

Keywords

Social role theory, topic modeling, Twitter

1. INTRODUCTION

Social media such as Twitter are the object of intensive studies in recent years. One of the key problems is to understand how contents are generated by users and different models have been proposed for it. A model for content generation serves as a basis to many other research problems such as information extraction, search, and recommendation. Statistical topic models have been a privileged tool for this task due to its proven capability to model the content Many studies have focused on extending standard topic models by considering new characteristics of Twitter, such as geographical information or hashtag mechanism. While these characteristics can improve the resulting topic models for Twitter contents, an important factor that is missing from the previous investigations is the social role a user plays in communications on Twitter. No distinction was made between the content generation processes of two users in different roles.

Intuitively, one would expect that a user who expresses an original opinion on Twitter would use a different generation process than a user who merely propagates the opinion of another. The influence of social roles on the communication contents has been clearly demonstrated in sociology and social psychology in which Social Role Theory (SRT) has developed [1]. In online communities or social network studies, social roles identified include popular initiators, popular participants, joining conversationalists who have medium initiation and participation, information sources who post news and have a large number of followers, and information seekers or lurkers who post rarely [2]. This paper proposes a principled approach which incorporates SRT into the generative process of topic models. In particular, since we aim to model Twitter content generation, we will only focus on the two most common social activities on Twitter, posting status messages and retweeting or forwarding messages to others. Hence, the two social roles identified are "*originators*" who publish original tweets and "*propagators*" who retweet others' tweets.

An illustrative example of incorporating SRT into Twitter content generation process is shown in Figure 1, where there are four users a, b, c and d, referred to as *social actors*. Both a and b posted a tweet on the topic of "Gangnam Style", and the other two users forwarded these two tweets and reposted them in their individual Twitter homepages. In this example, both a and b published original contents and can be viewed as *originators*; while c and d replicated and spread the existing information and can be viewed as *propagators*. In the above example, "originators" and "propagators" are referred to as *social roles*. Furthermore, since retweeting can be understood as a means of participating in a diffuse conversation, this implies explicit or implicit *social interactions* arising between different social roles. For example, the retweeting of a's tweet by c can be viewed as an explicit interaction between c as a propagator and a as an originator. On the other hand, the fact that both c and d retweeted a's tweet indicates that there exists an implicit interaction between c and d where both are propagators of the same tweet: they tend to share some common interests. Such an implicit relation is also useful for modeling the content generation process. For example, knowing the retweets by c is useful to determine the content of retweets by d.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2505599.

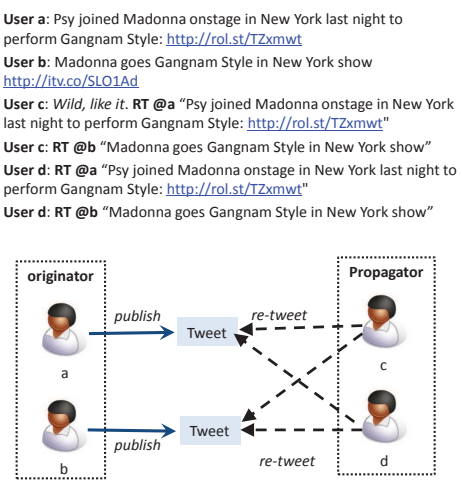


Figure 1: An illustrative example of social roles on Twitter.

As we can see from the above example, SRT provides a very interesting explanation of the generative process of Twitter content. However, SRT does not provide a framework ready to be implemented computationally. We have to take a comprehensive consideration of various elements in SRT, including social actors, social roles and social interactions. The main contribution and novelty of this paper is that we propose a novel regularized topic model that is flexible enough to capture the main ideas of SRT and reflect the key elements in SRT. We perform extensive experiments on two real Twitter data sets. Our results show that our model outperforms several baseline topic models that do not consider users' social roles or social interactions. The key features of our approach are the following: 1) We consider that a user can play multiple social roles, and each social role serves to fulfill different tasks and is associated with a user-specific and role-driven distribution over latent topics. 2) We formally model both explicit and implicit interactions with involved users' roles as context through the regularization factors.

2. PRELIMINARIES

2.1 Social Role Theory

Social Role Theory is a perspective in sociology and in social psychology that predominantly concerns characterizing behavior patterns or roles and explains roles by presuming that persons are members of social positions and hold expectations for their own behaviors and those of other persons [1]. Each person is a *social actor*, who acts according to some characterizing behavior patterns or *social roles*. Each *social role* is a set of rights, duties, expectations, norms and behaviors that a person has to face and fulfill[1]. Social actors can interact or collaborate with each other in a process called *social interaction*, which may influence involved users.

On Twitter, each user can be viewed as a social actor who is associated with a set of social roles. In social network studies, the social roles identified include popular initiators, popular participants, joining conversationalists, information sources, information seekers, and lurkers [2]. In our paper here, we focus on social activities which are related to the generation of online contents on Twitter. Hence, we only consider two types of social role: (1) *originators* who publish original content; (2) *propagators* who forward and spread content of others. Although such a definition of social roles is simple, it naturally captures the two most important aspects of Twitter content growth: the generation of new ideas and the spread of existing contents. As will be seen in Section 3, our proposed topic modeling approach can be easily extended to incorporate other social roles. A social actor is free to choose any role whenever she wants to engage in the process of information generation on Twitter. Although roles imply expected behaviors for social actors, a user can selectively contribute more information on the topics that she is more interested in. Furthermore, a user can explicitly interact with another user by forwarding her tweets; or implicitly interact with others by contributing contents to the same topics. During interactions, a user is influencing and being influenced by those who interact with her. Therefore the involved users tend to have similar topical interests. This will be modeled as regularization factors in our approach.

2.2 Notations

We first define a set of notations used in this paper before presenting our proposed role-based topic models.

Topics: A topic is a semantically coherent theme. We assume that there are a set of topics $\mathcal{T}$ over the document collection $\mathcal{C}$. We use variable θ_t to denote a topic model represented by a multinomial distribution $\theta_t = \{P(w|t)\}_{w \in \mathcal{V}}$ where $P(w|t)$ is the probability of word w given topic t according to the topic model θ_t, and $\mathcal{V}$ is the vocabulary.

Social actors: A user is a social actor who generates online content on Twitter. We use u or v to denote an individual user and $\mathcal{U}$ to represent a set of Twitter users (social actors).

Documents: A tweet is a document, which can be either a retweet or an originally-written tweet. We use $d_{u,i}$ to denote the ith tweet generated by user u. $\mathcal{C}$ is the entire collection of tweets generated by all users.

Social roles: We assume that there are a set of social roles $\mathcal{R}$ given a user u, and denote her as $u_{(r)}$ when she plays the role of r. A user will have a preference distribution to select roles, i.e., $\{P(r|u)\}_{r \in \mathcal{R}}$. We further assume that user u is associated with an interest distribution over topics when she plays the role of r, i.e., $\{P(t|u_{(r)})\}_{t \in \mathcal{T}}$. Note that all social roles will share a common set of topics, and the topic distribution of $\{P(t|u_{(r)})\}_{t \in \mathcal{T}}$ is both *user-* and *role-specific*. As we only define two social roles here, a user u can only have two possible roles $u_{(o)}$ (*originator*) and $u_{(p)}$ (*propagator*).

Social interactions: Generally speaking, social interaction is a kind of action that occurs as two or more users have an effect upon one another. In this paper, we do not consider each individual interaction but the overall interactive patterns between two users at a macro level. As we mentioned earlier, social interactions take place between two users with certain social roles and they drive users to have similar role-specific interests. Formally, we introduce a similarity function $s(u_{(r_u)}, v_{(r_v)})$ which measures the similarity of interests between u and v with roles r_u and r_v respectively. For the two social roles considered here, there are four possible forms for $s(u_{(r_u)}, v_{(r_v)})$, namely

[1] http://en.wikipedia.org/wiki/Role_theory

$s(u_{(p)}, v_{(p)})$, $s(u_{(o)}, v_{(p)})$, $s(u_{(p)}, v_{(o)})$ and $s(u_{(o)}, v_{(o)})$. A large value of $s(u_{(r_u)}, v_{(r_v)})$ indicates that u and v with roles r_u and r_v interact more often and hence are more likely to have similar interests. We consider involved users' roles when modeling social interactions, i.e., role-specific interactions.

3. THE TOPIC MODELING APPROACH - ROLE PLSA

With the notations introduced above, we now present our proposed topic model which is based on probabilistic latent semantic analysis (PLSA) [3] with users' social roles incorporated. The generative story of our model is as follows. When a user wants to post a tweet, she first selects a social role according to her role preference. Then, for each word, she chooses a topic based on her role-specific interest and subsequently generates the word according to the interest-related topic model under the specific role. Meanwhile, each user's role-specific interests are also influenced through social interactions.

In what follows, we start with a basic model without social interactions and then further extend it by incorporating the interactions as regularization factors.

In this paper, we assume that the prior distribution of users $\{P(u)\}_{u\in\mathcal{U}}$ follows a uniform distribution and we do not explicitly model $P(u)$. By summing over the latent variables, users' social roles r and topics t, the conditional probability of the ith tweet $d_{u,i}$ given the user u can be defined as

$$P(d_{u,i}|u) = \prod_{w\in d_{u,i}} \Big\{ \sum_{r\in\mathcal{R}} \Big(\sum_{t\in\mathcal{T}} P(w|t)P(t|u_{(r)}) \Big) P(r|u) \Big\}.$$

The above formula defines a general model for role-based topic modeling and can be applied to various scenarios involving different social roles in addition to the two roles we defined for Twitter here. The key challenges are how to align the learnt roles to the considered roles, i.e., originators and propagators, and how to relate a tweet to a specific role of a user. Instead of explicitly learning $P(r|u)$, our solution is to incorporate prior knowledge by making use of the retweeting conventions on Twitter to differentiate user roles. For example, tweets containing "RT" or "via" and followed by "@username" are considered as retweets and hence their authors' social role would be *propagator*, i.e., u_r = "*propagator*". Some retweets contain text before "RT" or "via", the user may play the roles of propagator and originator simultaneously. One may determine the probability of a specific role according to the proportion of the texts before and after "RT" or "via". However, we observe that in most cases, the text before "RT" or "via" is usually very short, or the inserted text is mainly to comment on the original post. Therefore, we will simply consider these cases as retweeting and the role of the user as "propagator". Otherwise, we consider their authors' social role as *originator*, i.e., u_r = "*originator*".

The log likelihood function $L(\mathcal{C})$ for the entire corpus can be written as

$$\begin{aligned} L(\mathcal{C}) &= \sum_{u\in\mathcal{U}}\sum_{w\in\mathcal{V}} n_O(u,w)\log\Big(\sum_{t\in\mathcal{T}} P(w|t)P(t|u_{(o)})\Big) \\ &\quad + \sum_{u\in\mathcal{U}}\sum_{w\in\mathcal{V}} n_R(u,w)\log\Big(\sum_{t\in\mathcal{T}} P(w|t)P(t|u_{(p)})\Big), \quad (1) \end{aligned}$$

where $n_O(u,w)$ is the frequency of w in the originally-written tweets by u while $n_R(u,w)$ is the frequency of w in the retweets by u.

We refer to this model as role PLSA (**rPLSA**). It provides a principled way to incorporate social roles and user interests into the model. In the next section, we will discuss how to formally model social interactions.

4. A REGULARIZED FRAMEWORK WITH SOCIAL INTERACTIONS

In addition to social roles and role-specific interests, social interaction is another important aspect to consider in SRT. As we mentioned earlier, on Twitter, users can interact with each other either explicitly or implicitly, and users' interests may be influenced during such interactions. Usually, social interactions tend to suggest that the interests of involved users are similar. We can derive from social interactions the similarity measurements, i.e., $s(u_{(r_u)}, v_{(r_v)})$, which indicates the similarity degree between users' role-specific interests. We model social interactions through regularization factors.

4.1 Modeling Explicit Interactions

On Twitter, one of the most prominent interactions is the forwarding mechanism, a.k.a. retweet. We adopt the retweet mechanism to measure explicit interactions. Specially, if user a has forwarded a considerable number of tweets from user b, the topic distribution of a as a propagator should be similar to the topic distribution of b as an originator. We can formally model this assumption by the following topical difference between the two users:

$$R_1 = \sum_{a,b\in\mathcal{U}} s(a_{(p)}, b_{(o)}) \Big\{ \sum_{t\in\mathcal{T}} (P(t|a_{(p)}) - P(t|b_{(o)}))^2 \Big\}, \quad (2)$$

where $s(a_{(p)}, b_{(o)})$ is the similarity between a and b as an originator and a propagator respectively. We measure $s(a_{(p)}, b_{(o)})$ as

$$s(a_{(p)}, b_{(o)}) = \frac{n_{a,b}}{n_{a_{(p)}} + n_{b_{(o)}} - n_{a,b}}, \quad (3)$$

where $n_{a,b}$ is the number of retweets forwarded by a from b, $n_{a_{(p)}}$ is the number of retweets of a and $n_{b_{(o)}}$ is the number of tweets written originally by b.

4.2 Modeling Implicit Interactions

Sometimes, users do not explicitly but implicitly interact with the others. For example, if both a and b are very interested in the song of "Gangnam Style" and publish originally-written tweets on this topic, we say a and b, both as originators, interact with each other implicitly. They reveal similar interests as originators and contribute new information on the same topic. Similarly, c and d, both as propagators, interact with each other implicitly since they replicate existing tweets to spread information on the same topic.

Compared with explicit interactions, it is more difficult to discover and model implicit interactions. We identify implicit interactions through users' forwarding behaviors. As the illustrative example in Figure 1 shows, the tweets of a and b are forwarded by common users c and d. It indicates that a and b might have similar interests as originators due to the fact that they interact with common propagators.

Similarly, c and d might also have similar interests as propagators since they interact with common originators. The above two types of implicit interactions can leverage latent similarities of user interests. Let us consider them in more detail.

Type I: an originator $\leftrightarrow$ common propagators $\leftrightarrow$ another originator. This type of implicit interactions exists between two originators who are retweeted by some common propagators. Intuitively, if the tweets of two users a and b have been forwarded by a considerable number of common users, the topic distribution of a as an originator should be similar to the topic distribution of b as another originator.

We can formally model this assumption as the following dissimilarity of topical distributions

$$R_2 = \sum_{a,b\in\mathcal{U}} s(a_{(o)}, b_{(o)}) \Big\{ \sum_{t\in\mathcal{T}} (P(t|a_{(o)}) - P(t|b_{(o)}))^2 \Big\}, \quad (4)$$

where $s(a_{(o)}, b_{(o)})$ is the similarity between a and b as originators. Each originator is represented as a vector where each of its elements corresponds to one of her propagators weighted by the number of tweets forwarded by the propagator. We use the cosine function to compute the similarity

$$s(a_{(o)}, b_{(o)}) = \sum_{c\in\mathcal{U}} \frac{n_{c,a} n_{c,b}}{\sqrt{(\sum_{c'} n_{c',a}^2)(\sum_{c'} n_{c',b}^2)}} \quad (5)$$

where $n_{c,a}$ and $n_{c,b}$ denote the number of retweets forwarded by c from a and b respectively.

Type II: a propagator $\leftrightarrow$ common originators $\leftrightarrow$ another propagator. Similarly, if two users a and b have similar forwarding behaviors, i.e., co-forwarding many tweets from common users, then the topic distribution of a as a propagator should be similar to the topic distribution of b as a propagator.

We can formally model this assumption as follows

$$R_3 = \sum_{a,b\in\mathcal{U}} s(a_{(p)}, b_{(p)}) \Big\{ \sum_{t\in\mathcal{T}} (P(t|a_{(p)}) - P(t|b_{(p)}))^2 \Big\}. \quad (6)$$

where $s(a_{(p)}, b_{(p)})$ is the similarity between a and b as propagators. We represent each propagator as a vector of originators weighted by the number of forwarding tweets between them, and then we use the cosine function to compute the similarity

$$s(a_{(p)}, b_{(p)}) = \sum_{c\in\mathcal{U}} \frac{n_{a,c} n_{b,c}}{\sqrt{(\sum_{c'} n_{a,c'}^2)(\sum_{c'} n_{b,c'}^2)}} \quad (7)$$

where $n_{a,c}$ (and likewise $b_{b,c}$) denotes the number of retweets forwarded by a from c.

4.3 Integrating the Model with Regularization Factors

After defining the three regularization factors, we combine them into a unified regularized formula

$$R(\mathcal{U}) = \lambda_1 R_1 + \lambda_2 R_2 + \lambda_3 R_3, \quad (8)$$

where $\lambda_1, \lambda_2, \lambda_3 > 0$, and $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

The functions of $s(\cdot,\cdot)$ in Eq. 3, Eq. 5 and Eq. 7 provides a way to measure the interest similarities between two users with specific roles. Given a user u with the role r, i.e., $u_{(r)}$, we can find her K most similar originators and K most similar propagators respectively, referred to as *neighbors* of $u_{(r)}$. To make our algorithm efficient, for $u_{(r)}$, we only keep at most 30 neighbors in each role, i.e., $K = 30$.

To incorporate both the social role based topic models and the regularization factors, we define a regularization framework by adding the (negative) regularization term to the log-likelihood of rPLSA as follows

$$L(\mathcal{C},\mathcal{U}) = L(\mathcal{C}) - \mu R(\mathcal{U}), \quad (9)$$

where $\mu \geq 0$. When $\mu = 0$, it becomes the rPLSA that we introduced before; when $\mu > 0$, the whole likelihood is a trade-off between text based likelihood and the regularization loss. We refer to this model as **rrPLSA**.

$$\begin{aligned}
\pi_{u_{(o)},t,w} &= P(z=t|u_{(o)},w) = \frac{P(w|t)P(t|u_{(o)})}{\sum_{t'} P(w|t')P(t'|u_{(o)})}.\\
\pi_{u_{(p)},t,w} &= P(z=t|u_{(p)},w) = \frac{P(w|t)P(t|u_{(p)})}{\sum_{t'} P(w|t')P(t'|u_{(p)})}.\\
P(t|u_{(o)}) &= \frac{\sum_w n_O(u,w)\pi_{u_{(o)},t,w} + \alpha}{\sum_{t',w'} n_O(u,w')\pi_{u_{(o)},t',w'} + |\mathcal{T}|\alpha}.\\
P(t|u_{(p)}) &= \frac{\sum_w n_R(u,w)\pi_{u_{(p)},t,w} + \alpha}{\sum_{t',w'} n_R(u,w')\pi_{u_{(p)},t',w'} + |\mathcal{T}|\alpha}.\\
P(w|t) &= \frac{\sum_u \left(n_O(u,w)\pi_{u_{(o)},t,w} + n_R(u,w)\pi_{u_{(p)},t,w} \right) + \beta}{\sum_{u',w'} \left(n_O(u',w')\pi_{u'_{(o)},t,w'} + n_R(u',w')\pi_{u'_{(p)},t,w'} \right) + |\mathcal{V}|\beta}.
\end{aligned}$$

Figure 2: EM updating formulae for rPLSA.

4.4 Parameter Estimation with a Generalized EM Algorithm

We adopt the standard Expectation Maximization (EM) algorithm for parameter estimation of rPLSA, i.e. when $\mu = 0$. It is worth noting that in order to avoid zero probabilities, we have applied Laplace smoothing[2] by adding a small value of α when estimating $P(t|u_{(o)})$ and $P(t|u_{(p)})$, and a small value of β when estimating $P(w|t)$. We found that the model performance is relatively stable when $\alpha \in [1e-5, 1]$ and $\beta \in [1e-7, 1e-1]$. In all our experiments reported here, we set $\alpha = 1e-3$ and $\beta = 1e-7$. The updating formulas of the EM algorithm are given in Figure 2.

When $\mu \neq 0$, the case is more complex and cannot be solved by the standard EM algorithm. Therefore, we adopt the generalized EM (GEM) algorithm to find the solution, which has been described in details in [7]. Due to space limitation, we only present the Newton-Raphson updating formulas for $P(t|u_{(o)})$ and $P(t|u_{(p)})$ in Figure 3. Although the formulas in Figure 3 look complicated, it has intuitive explanations. The new role-specific topic distribution of a user is the old distribution smoothed by the topic distributons of her "neighbors" defined in Section 4.3. Furthermore, the neighbors can be divided into two groups, namely originators and propagators.

5. EXPERIMENTS

5.1 Construction of the Datasets

We evaluate our proposed models on two datasets sampled from the Twitter data shared by Kwak et al. [5] which spanned the second half of year 2009. For each dataset, we first select 30 seed users, and then perform breadth-first search for two iterations to add users by using the retweeting

[2] http://en.wikipedia.org/wiki/Additive_smoothing

$$P(t|u_{(o)})_{n+1}^{(k+1)} = (1-\delta)P(t|u_{(o)})_{n+1}^{(k)} + \delta \frac{\sum_{v\in\mathcal{U}}\left(s(v_{(p)},u_{(o)})P(t|v_{(p)})_{n+1}^{(k)} + s(v_{(o)},u_{(o)})P(t|v_{(o)})_{n+1}^{(k)}\right)}{\sum_{v\in\mathcal{U}}\left(s(v_{(p)},u_{(o)}) + s(v_{(o)},u_{(o)})\right)}$$

$$P(t|u_{(p)})_{n+1}^{(k+1)} = (1-\delta)P(t|u_{(p)})_{n+1}^{(k)} + \delta \frac{\sum_{v\in\mathcal{U}}\left(s(u_{(p)},v_{(o)})P(t|v_{(o)})_{n+1}^{(k)} + s(v_{(p)},u_{(p)})P(t|v_{(p)})_{n+1}^{(k)}\right)}{\sum_{v\in\mathcal{U}}\left(s(u_{(p)},v_{(o)}) + s(v_{(p)},u_{(p)})\right)}$$

Figure 3: Newton-Raphson updating formulas for $P(t|u_{(o)})$ and $P(t|u_{(p)})$ in the M-step of rrPLSA. The step parameter δ, empirically set to be 0.05, can be interpreted as a controlling factor of smoothing the role-based topic distribution via social interactions.

Table 1: Statistics of the two datasets.

	#users	#tweets	#retweet-links
$\mathcal{D}_{music}$	13,094	4,663,365	83,069
$\mathcal{D}_{random}$	12,498	4,302,784	92,712

links of these seed users (including both retweet in and out links). The first dataset is domain-specific with seed users selected from music celebrities. The second dataset has its seed users randomly selected from the users with most retweets. Hence it contains general tweets without specific topic focus. We collect all tweets of the users in August, 2009. Since we aim to study the effect of social interactions, we discard users with very few tweets or very few retweet in/out links. The statistics of the two datasets is summarized in Table 1.

5.2 Experimental Setup

For our proposed models, we have two variants: one is rPLSA which ignores social interactions and the other is rrPLSA with social interactions taken into account. We empirically set $\mu = 1000$ and $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$. We compare our models with the following topic models:

• *Latent Dirichlet Allocation* (LDA). We treat one tweet as one document and run the standard LDA model on our datasets.

• *Author-Topic* (AT) Model. We aggregate all the tweets of the same user into one document and run the AT model on such aggregated document datasets. It is worth noting that our model rPLSA will degenerate to the AT model when we set the number of roles to one. We use it as a comparison of rPLSA to examine *the impact of social roles*.

• *NetPLSA* [7] extends AT model by incorporating explicit social networks, which is a widely used model for text data with network links. We use it as a comparison of rrPLSA to examine *the impact of social roles and role-specific social interactions, especially the implicit interactions*. We construct the social network using the retweet links and set the link weight from vertex (or user) u to v as $s(u_{(p)}, v_{(o)})$ in Equation 3. To do a fair comparison, we also applied Laplace smoothing with the same smoothing parameters.

5.3 Predictive Power

We set up two evaluation tasks to evaluate models' predictive power on unseen data, namely document modeling and retweet prediction. All the models were trained on each of these two datasets summarized in Table 1 (data in August 2009), and then tested on a test set. We built the test set by first randomly selecting 5000 users from each of the training sets. For these users, we collected all their tweets posted in the first week of September 2009 for document modeling. We also collected the tweets of all the users they follow and kept the information about whether these testing users have forwarded the tweets or not for retweet prediction.

Document modeling. The first evaluation task aims to examine the overall generalization ability of modeling unseen data. The commonly used perplexity measure is adopted as the evaluation metrics of document modeling. A lower perplexity score indicates better generalization performance. In our experiments, a "document" is simply a tweet posted by a user. Given a test set $\mathcal{D}_{test}$, the perplexity is computed as:

$$perplexity(\mathcal{D}_{test}) = \exp\left\{-\frac{\sum_{d\in\mathcal{D}_{test}}\log P(\mathbf{w}_d)}{\sum_{d\in\mathcal{D}_{test}} N_d}\right\},$$

where d is a document in $\mathcal{D}_{test}$, $\mathbf{w}_d$ is the token stream of d, and N_d is the number of tokens in d. For all the models evaluated here, each of them has its own formula to compute $P(\mathbf{w}_d)$. It is worth mentioning that both rPLSA and rrPLSA will use different topic distributions respectively for originally-written tweets and retweets.

Retweet prediction. On Twitter, a user can browse all the tweets from the users in her following list and can decide to retweet some of the tweets to her own followers. In this part, we focus on evaluating models' capability on predicting whether a user will retweet a tweet from the users she follows. As we aim to test whether our proposed topic models are better than the other baselines, we simplify the retweet prediction task as follows. For each user, we only consider the tweets of the users she follows from whom she has at least forwarded one tweet in the first week of September, 2009. We compute the topic similarity between a candidate tweet and the topical interest of a user. Then we rank these tweets in a descending order. A better method should be able to rank those tweets that the user has actually forwarded at higher positions.

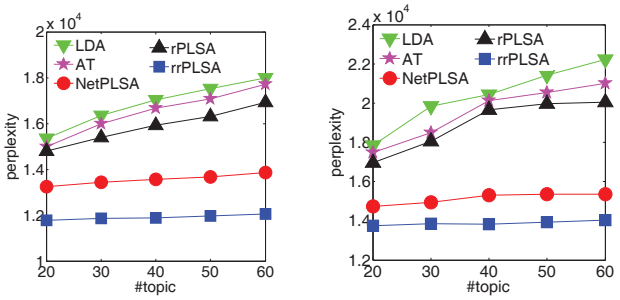
Given a user, the baseline topic models can only learn a single user interest distribution; while our proposed topic models can learn both the originator specific and the propagator specific interest distributions. We only use the propagator specific interest distribution for retweet prediction. Given a set of topic models $\{\theta_t\}_{t\in\mathcal{T}}$, we compute the conditional probability of topic t given a tweet d for each of $t \in \mathcal{T}$

$$P(t|d) = \frac{\prod_{w\in d} P(w|\theta_t)}{\sum_{t'\in\mathcal{T}}\prod_{w\in d} P(w|\theta_{t'})}.$$

Given a user and a set of tweets, we first compute the negative KL-divergence of the topic distributions of the user

Table 2: Performance comparisons of retweet prediction on $\mathcal{D}_{random}$.

Metrics	LDA	AT	NetPLSA	rPLSA	rrPLSA
P@10	0.050	0.053	0.056	0.055	**0.059**
P@20	0.098	0.100	0.101	0.101	**0.117**
P@30	0.138	0.140	0.150	0.148	**0.166**
P@100	0.408	0.410	0.436	0.419	**0.453**
MRR	0.159	0.160	0.168	0.163	**0.181**



(a) $\mathcal{D}_{music}$ (b) $\mathcal{D}_{random}$

Figure 4: Performance comparisons of perplexity ($\times 10^4$).

and each of candidate tweets, and subsequently rank these tweets in a descending order. We adopt `precision@N` and `MRR` (Mean Reciprocal Rank) commonly used in information retrieval as our evaluation metrics, i.e., a retweet will be judged as a relevant "document". We set the topic number to 40, and only report the results on $\mathcal{D}_{random}$ due to the space limit.

Experimental results on Perplexity. The results of perplexity and retweet prediction are shown in Figure 4 and Table 2 respectively. It can be observed that in terms of perplexity results, rPLSA has better predictive power than AT and LDA by incorporating social roles, although it performs worse than NetPLSA. By additionally incorporating social interactions as regularization factors, rrPLSA significantly outperforms other models by a large margin.

Experimental results on Retweet prediction. Retweet prediction is a very challenging problem and previous research has proposed complex models to solve this problem [4], including content features, temporal features and social link features. Here we do not want to take into account all these features and want to focus only on the content features and the social roles. Therefore, the overall performance of all the models on retweet prediction is low, as revealed by Table 2. Nevertheless, rrPLSA is still better than NetPLSA in terms of all metrics. The best results are achieved using rrPLSA, which outperforms NetPLSA by 7.7% in terms of `MRR`. These results show the effectiveness of our proposed rrPLSA which incorporates social roles and interactions.

6. RELATED WORK

Several previous studies have included author or user information when modeling documents using topic models. For example, the Author-Topic (AT) model [8] models a document with multiple authors as a distribution over topics that is a mixture of the distributions associated with the authors. Built upon the AT model, there are a few studies related to our work: the Author-Recipient-Topic (ART) model [6] and Role-Author-Recipient-Topic (RART) model [6]. Our work is related to the aforementioned models but have the following differences or novelties: 1) In our proposed rrPLSA, the roles we consider align to the two most common social activities on Twitter. 2) The role-based interests are both role-specific and user-specific, i.e., ***each role of each user is modeled as a distribution over topics***. 3) We formally model both *explicit* and *implicit* interactions with involved users' roles as context through the regularization factors.

7. CONCLUSIONS

In this paper, we have proposed a novel topic model, called rrPLSA, which incorporates both social roles and social interactions into a unified framework. Our proposed model aims to explicitly capture the underlying generative process of Twitter contents in a new perspective, i.e., Social Role Theory, and it reflects the key elements of SRT. There are several possible directions to pursue for future work. In this paper, we incorporate domain knowledge to identify users' behaviorial social roles. It is worth to explore automatic learning methods for user role identification. It is also possible to extend our proposed approach to model other online social networks, such as Facebook, MySpace, online question-answering communities, where users play different roles. Another promising direction is to investigate the feasibility of incorporating other characteristics such as users' geographical regions into our current framework.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for the constructive comments. The work was partially supported by NSFC Grant 61272340 and 60933004. Yulan He was partially supported by the EPSRC grant EP/J020427/1. Xin Zhao was supported by Microsoft Research Asia Fellowship.

8. REFERENCES

[1] B.J. Biddle. Recent development in role theory. *Annual review of sociology*, pages 67–92, 1986.

[2] Jeffrey Chan, Conor Hayes, and Elizabeth M. Daly. Decomposing discussion forums and boards using user roles. In *ICWSM*, 2010.

[3] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, 1991.

[4] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. Predicting popular messages in twitter. In *WWW*, 2011.

[5] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue B. Moon. What is twitter, a social network or a news media? In *WWW*, 2010.

[6] Andrew McCallum, Andrés Corrada-Emmanuel, and Xuerui Wang. Topic and role discovery in social networks. In *IJCAI*, 2005.

[7] Qiaozhu Mei, Deng Cai, Duo Zhang, and Chengxiang Zhai. Topic modeling with network regularization. In *WWW*, 2008.

[8] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas L. Griffiths. Probabilistic author-topic models for information discovery. In *KDD*, 2004.