Incorporating Social Role into Topic Models for Social Media Content Analysis

Wayne Xin Zhao, Jinpeng Wang, Yulan He, Jian-Yun Nie, Ji-Rong Wen and Xiaoming Li

Abstract—In this paper, we explore the idea of Social Role Theory (SRT) and propose a novel regularized topic model which incorporates SRT into the generative process of social media content. We assume that a user can play multiple social roles, and each social role serves to fulfil different duties and is associated with a role-driven distribution over latent topics. In particular, we focus on social roles corresponding to the most common social activities on social networks. Our model is instantiated on microblogs i.e. *Twitter* and community question-answering (cQA) i.e. *Yahoo! Answers*, where social roles on Twitter include "originators" and "propagators", and roles on cQA are "askers" and "answerers". Both explicit and implicit interactions between users are taken into account and modeled as regularization factors. To evaluate the performance of our proposed method, we have conducted extensive experiments on two Twitter datasets and two cQA datasets. Furthermore, we also consider multi-role modeling for scientific papers where an author's research expertise area is considered as a social role. A novel application of detecting users' research interests through topical keyword labeling based on the results of our multi-role model has been presented. The evaluation results have shown the feasibility and effectiveness of our model.

Index Terms—Topic models, Social Role Theory, Social Media

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. With the excellent performance of statistical topic models on traditional document collections (e.g., scientific publications) [1], researchers have developed various topic models to perform deep content analysis of online social networks by considering new characteristics of these social websites, such as geographical information [2] on Twitter.

While such side information can improve the topic model performance for social media analysis, what has been ignored in previous studies is that users often play different social roles in social networks and their online contents generated are influenced by their social roles. The influence of social behaviors and activities on the communication contents has been clearly demonstrated in sociology and social psychology in which the Social Role Theory (SRT) has been developed [3]. In this paper, we propose to incorporate social role into social media content analysis and provide a novel perspective to understand the underlying content generation process.

Taking Twitter as an example, there are two most common social activities, posting status messages and retweeting or forwarding messages to others. The two social roles corresponding to the two activities are "*originators*" who publish original tweets and "*propagators*" who retweet others' tweets. Intuitively, one would expect that a user who expresses an original opinion on Twitter would follow a different content generation process compared to a user who merely propagates others' tweets. An illustrative example of incorporating SRT into Twitter content generation process is shown in Fig. 1, where there are four users a, b, c and d, referred to as *social actors*. Both a and b posted a original tweet on the topic of "Gangnam Style" and can be viewed as *originators*; and the other two users forwarded these two tweets and re-posted them in their individual Twitter home pages and can be viewed as *propagators*. Here, "originators" and "propagators" are referred to as *social roles*. Furthermore, since retweeting can be understood as a means of participating in a diffuse conversation [4], 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 infer the contents of retweets by d.

As we can see from the above example, SRT provides a very interesting explanation of the generative process of Twitter contents. However, it is not straightforward to model SRT for content generation. We have to take a comprehensive consideration of various elements in SRT, including social actors, social roles and social interactions. The major contribution and novelty of this paper is that we propose a novel regularized topic model that is flexible enough to cap-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2014.2359672, IEEE Transactions on Knowledge and Data Engineering

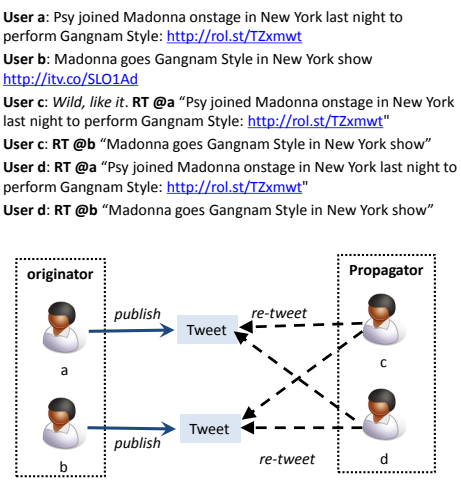


Fig. 1. An illustrative example of social roles on Twitter.

ture the key elements in SRT. 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 [5], [6]. Different from them, in this paper, we mainly focus on roles corresponding to common social activities, which are easy to identify, explain and understand. We perform extensive experiments on Twitter datasets, community question-answering datasets and scientific paper dataset. 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 assume that a user can play multiple social roles, and each social role serves to fulfil different duties and is associated with a user-specific and role-driven distribution over latent topics. We also incorporate a base interest distribution to model a user's persistent topical interest which is independent of roles the user plays. 2) We formally model both explicit and implicit interactions with involved users' roles as context through the regularization factors. In particular, we will show that the implicit interactions are very useful in modelling online contents. 3) Our approach provides a novel perspective to understand and analyze online social media, which is equally applicable to model other online social networks, such as Facebook and MySpace, where users also play different roles.

The rest of this paper is organized as follows. An overview of SRT and the notations used in the paper are given in Section 2. A novel topic model with social roles incorporated, called rPLSA, is presented and subsequently extended, called rrPLSA, by adding the modeling of social interactions as regularization factors in Section 3. We instantiate rrPLSA in the Twitter setting in Section 4. Experimental setup and results on Twitter datasets, cQA datasets and scientific paper dataset are discussed in Section 5, 6 and 7 respectively. Finally, the related work and conclusions are given in Section 8 and 9 respectively.

2 Preliminaries

2.1 Social Role Theory

Social Role Theory is a perspective in sociology and social psychology that predominantly concerns characterizing 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 [3]. 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.

A social actor is free to choose any role whenever she wants to engage in the process of information generation on social networks. 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 e.g., forwarding her tweets on Twitter; 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.

2.2 Notations

We first define a set of notations 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 to denote a topic model represented by a multinomial distribution $\theta = \{P(w|t)\}_{w \in \mathcal{V}}$ where $P(w|t)$ is the probability of word w given topic t according to the topic model θ, and $\mathcal{V}$ is the vocabulary.

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

Documents: A document consists of a bag of tokens published by a social actor on social networks. We use d to denote a document.

Social roles: We assume that there are a set of social roles $\mathcal{R}$ given a user u, and denote the user as $u_{(r)}$ when she plays the role of r. A user will

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

have a preference distribution to select roles, i.e., $\{P(r|u)\}_{r\in\mathcal{R}}$. All social roles will share a common set of topics, and a user u is associated with an interest distribution over topics when she plays the role of r, i.e., $\{P(t|u_{(r)})\}_{t\in\mathcal{T}}$, which are both *user-* and *role-specific*.

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 between u and v with roles r_u and r_v respectively based on their social interaction patterns. 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 topical interests.

Base interest: Apart from the aforementioned role-specific topical interests, users might also have their persistent topical interests which are less likely to be influenced by social interactions and are thus independent of different roles that they play. In order to characterize users' persistent topical interests, we assume that a user u is associated with a base interest distribution represented by a multinomial distribution over topics, i.e. $\{P(t|u_{(B)})\}_{t\in\mathcal{T}}$, where $u_{(B)}$ denotes user u's persistent interest or base interest. Under a specific role, a user can generate the content with her role-specific interest or her base interest. A role-specific weight parameter $0 < \eta_{u_{(r)}} < 1$ is used to control the trade-off between the role-specific interest and the base interest.

3 The Model Framework

3.1 Role Specific Topic Models

With the notations introduced above, we now present our proposed topic model which is based on probabilistic latent semantic analysis (PLSA) [7] with users' social roles incorporated. The generative story of our model is as follows. When a user wants to post a document, she first selects a social role according to her role preference. Then, for each word, she chooses a topic based on either her role-specific interest or her base interest, and subsequently generates the word according to the selected topic. 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.

Modeling social roles

In the above generative story, we assume that each document is associated with a specific role and the role-specific weight parameter $\eta_{u_{(r)}}$ controls the trade-off between the role-specific interest and the base interest in the generative process. By summing over the latent variables, users' social roles r and topics t, the conditional probability of the document d given the user u can be defined as

$$P(d|u) = \sum_{r\in\mathcal{R}} P(r|u)\Big\{\prod_{w\in d}\sum_{t\in\mathcal{T}} P(w|t) \Big(\eta_{u_{(r)}} P(t|u_{(r)}) + (1-\eta_{u_{(r)}}) P(t|u_{(B)})\Big)\Big\}.$$

The above formula defines a general model for role-based topic modeling and can be applied to various scenarios involving different social roles. Here we make an assumption that the roles align to specific social activities and we can identify the role of a user by the activity that she performs when publishing a document. For example, on Twitter, we can consider the two most distinctive activities, i.e., posting an originally-written tweet or forward an existing tweet from others, which correspond to the two roles *originator* and *propagator* respectively. Similarly, in online question-answering communities, we can easily identify two activities, i.e., posting a question or answering a question. If a user posts a question, her role is naturally a *asker* while a user who provides an answer plays the role of *answerer*.

Therefore, we assume that given a document, a user will play a single role that can be identified by the activity she has performed. With this assumption, we can rewrite our model as follows

$$P(d|u) = \prod_{w\in d}\sum_{t\in\mathcal{T}} P(w|t)\Big(\eta_{u_{(r_{d,u})}} P(t|u_{(r_{d,u})}) + (1-\eta_{u_{(r_{d,u})}}) P(t|u_{(B)})\Big).$$

where $r_{d,u}$ is the role that user u plays when she posts document d. Given a document, the model associates the user with a specific role and each word in the document is generated from either the role-specific interest distribution or the user's based interest distribution. With some mathematical manipulations, the log likelihood function $L(\mathcal{C})$ for the entire corpus can be written as

$$L(\mathcal{C}) = \sum_{r\in\mathcal{R}}\sum_{u\in\mathcal{U}}\sum_{w\in\mathcal{V}} n_r(u,w)\log\Big\{\sum_{t\in\mathcal{T}} P(w|t)\Big(\eta_{u_{(r)}} P(t|u_{(r)}) + (1-\eta_{u_{(r)}}) P(t|u_{(B)})\Big)\Big\}, \quad (1)$$

where $n_r(u,w)$ is the frequency of w in the documents where user u plays the role of r. We refer to this model as role PLSA (**RolePLSA**). It provides a principled way to model social roles and user interests.

Modeling social interactions

Apart from social roles, social interaction is another important aspect to consider in SRT. On social network sites, users can interact with each other either explicitly or implicitly, and users' interests may be influenced by such interactions. Social interactions can be viewed as one type of social connections, and previous studies have shown that social connections are important evidence to reveal user interest similarities [8]. Therefore, 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 role specific social interactions through regularization factors, which have been widely used to model social connections [9], [10], [11].

Specially, if user u with role r_u has made a considerable number of interactions with user v with role r_v, the topic distribution of u with role r_u should be similar to the topic distribution of v with role r_v. To simplify the notations we use $P_{u_{(r_u)}}$ and $P_{v_{(r_v)}}$ to denote $\{P(t|u_{(r_u)})\}_{t\in\mathcal{T}}$ and $\{P(t|v_{(r_v)})\}_{t\in\mathcal{T}}$ respectively. We can formally model this assumption as follows

$$R(\mathcal{U}) = \sum_{u,v\in\mathcal{U}} \sum_{(r_u,r_v)\in\mathcal{R}^2} \lambda_{r_u,r_v} s(u_{(r_u)}, v_{(r_v)}) D\big(P_{u_{(r_u)}}, P_{v_{(r_v)}}\big) \tag{2}$$

where $s(u_{(r_u)}, v_{(r_v)})$ is the interest similarity between u and v with roles r_u and r_v respectively measured based on their social interactions, λ_{r_u,r_v} is the weight of type (r_u, r_v) interaction, $D\big(\cdot,\cdot\big)$ is a general distance function which measures the distance between two users' role-specific interest distributions and can be instantiated by various distance metrics. In this paper we mainly consider two types of distance metrics

- Symmetric Kullback-Leibler divergence (SKL). The Kullback-Leibler divergence[2] is a non-symmetric measure of the difference between two probability distributions P and Q, i.e. $D_{KL}(P||Q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$. Here we use the symmetric version to compute the distance

$$\begin{aligned} & D_{SKL}\big(P_{u_{(r_u)}}, P_{v_{(r_v)}}\big) \\ = \; & \frac{1}{2} D_{KL}(P_{u_{(r_u)}}||P_{v_{(r_v)}}) + \frac{1}{2} D_{KL}(P_{v_{(r_v)}}||P_{u_{(r_u)}}). \end{aligned}$$

- Residual Sum of Squares (RSS). We sum the squares of pointwise residual

$$D_{RSS}\big(P_{u_{(r_u)}}, P_{v_{(r_v)}}\big) = \frac{1}{2} \sum_{t\in\mathcal{T}} (P(t|u_{(r_u)}) - P(t|v_{(r_v)}))^2.$$

Particularly, with RSS distance metric, $R(\mathcal{U})$ can be rewritten in the form of a graph harmonic function

$$R(\mathcal{U}) = \sum_{t\in\mathcal{T}} \big(\mathbf{f}^t\big)^\top \mathbf{\Delta} \mathbf{f}^t, \tag{3}$$

where $\mathbf{f}^t$ is a L-dimensional ($L = |\mathcal{U}| \times |\mathcal{R}|$) vector of the probabilities on topic t for a user with a role of r, i.e. $f^t_{(u,r)} = P(t|u_r)$. $\mathbf{\Delta}$ is the graph Laplacian matrix and $\mathbf{\Delta} = \mathbf{A} - \mathbf{S}$, where $\mathbf{S}$ is a L-by-L matrix of weighted "edge" weights (i.e. similarities) and $S_{(u,r,v,r')} = \lambda_{r,r'} \times s(u_r, v_{r'})$. $\mathbf{A}$ is a L-by-L diagonal matrix and $A_{(u,r,u,r)} = \sum_{v\in\mathcal{U}} \sum_{r'\in\mathcal{R}} \lambda_{r,r'} \times s(u_r, v_{r'})$.[3] The major novelty of our work is the incorporation of social roles, for two users u and v, we have multiple regularization factors between them, i.e., a pair (r, r') indexes a regularization factor for $\{P(t|u_{(r)})\}_{t\in\mathcal{T}}$ and $\{P(t|v_{(r')})\}_{t\in\mathcal{T}}$. Graph Laplacian regularization is usually used to model the node similarities on the graph defined on feature space [12]. By casting our RSS loss function into the graph harmonic function, it leads to an intuitive explanation of social interaction: social interaction can be understood as links between user nodes on the role-specific interaction graph and users with large weight edges tend to share similar role-specific interests.

To incorporate both the social role based topic models and the regularization factors, we define a regularization framework by subtracting the regularization term from the log-likelihood of rPLSA as follows

$$L(\mathcal{C}, \mathcal{U}) = L(\mathcal{C}) - \mu R(\mathcal{U}), \tag{4}$$

where $\mu \geq 0$. When $\mu = 0$, it becomes 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 **RegRolePLSA**.

3.2 Parameter Estimation with a Generalized EM Algorithm

In this section, we discuss how to estimate model parameters for both rPLSA and rrPLSA. Our parameters include the topics and the role-specific distributions of topics, which are denoted as $\theta = \{P(w|t)\}$, $\phi = \{P(t|u_{(r)})\}$ respectively. In addition, we need to estimate the role-specific weights, $\{\eta_{u_{(r)}}\}$, which controls the trade-off between role-specific interests and base interests. We use Ψ to denote all the parameters. We will start with rPLSA and then extend the method to rrPLSA by using a generalized EM algorithm.

3.2.1 Standard EM algorithm for rPLSA

We adopt Expectation Maximization (EM) algorithm for parameter estimation of rPLSA, which degenerates to the log-likelihood function without the regularization terms. In the E-step, we determine the topic assignment of each word. Formally, we use the hidden variables $z_{u_{(r)},w}$ and $z^r_{u_{(B)},w}$ to indicate the

2. http://en.wikipedia.org/wiki/Kullback-Leibler_divergence

3. For convenience, we use a pair (u, r) to index the $(u\times|\mathcal{R}|+r)^{th}$ entry of a vector; similarly, we use a quadruple (u, r, v, r') to index the entry in the $(u\times|R|+r)^{th}$ row and the $(v\times|R|+r')^{th}$ column of a L-by-L matrix, where $L = |\mathcal{U}| \times |\mathcal{R}|$.

$$P(z_{u_{(r)},w}=t)=\frac{\eta_{u_{(r)}}P(w|t)P(t|u_{(r)})}{\sum_{t'}P(w|t')\{\eta_{u_{(r)}}P(t'|u_{(r)})+(1-\eta_{u_{(r)}})P(t'|u_{(B)})\}}.$$

$$P(z^r_{u_{(B)},w}=t)=\frac{(1-\eta_{u_{(r)}})P(w|t)P(t|u_{(B)})}{\sum_{t'}P(w|t')\{\eta_{u_{(r)}}P(t'|u_{(r)})+(1-\eta_{u_{(r)}})P(t'|u_{(B)})\}}.$$

$$P(t|u_{(r)})=\frac{\sum_{w}n_r(u,w)P(z_{u_{(r)},w}=t)+\alpha}{\sum_{t',w'}n_r(u,w')P(z_{u_{(r)},w'}=t')+|\mathcal{T}|\alpha}.$$

$$P(t|u_{(B)})=\frac{\sum_{r,w}n_r(u,w)P(z^r_{u_{(B)},w}=t)+\alpha}{\sum_{r',t',w'}n_r(u,w')P(z^{r'}_{u_{(B)},w'}=t')+|\mathcal{T}|\alpha}.$$

$$P(w|t)=\frac{\sum_{u,r}n_r(u,w)\{P(z_{u_{(r)},w}=t)+P(z^r_{u_{(B)},w}=t)\}+\beta}{\sum_{u,r',w'}n_{r'}(u',w')\{P(z_{u_{(r')},w'}=t)+P(z^{r'}_{u_{(B)},w'}=t)\}+|\mathcal{V}|\beta}.$$

$$\eta_{u_{(r)}}=\frac{\sum_{t,w}n_r(u,w)P(z_{u_{(r)},w}=t)}{\sum_{t',w'}n_r(u,w')\{P(z_{u_{(r)},w'}=t')+P(z^r_{u_{(B)},w'}=t')\}}.$$

Fig. 2. EM updating formulae for rPLSA.

topic assignment of word w according to the role-specific interest and the base interest respectively in the documents where u plays role r. Let $P(z_{u_{(r)},w}=t)$ and $P(z^r_{u_{(B)},w}=t)$ denote the posterior probabilities of word w generated by topic t according to role-specific interest and base interest respectively when u plays the role r. In the M-step, we first write the complete expected log likelihood of the whole dataset as follows (i.e., Q-function in EM algorithm)

$$Q_{\mathcal{C}}=\sum_{u,r,t,w}n_r(u,w)P(z_{u_{(r)},w}=t)\log\left(\eta_{u_{(r)}}p(w|t)P(t|u_{(r)})\right)$$
$$+\sum_{u,r,t,w}n_r(u,w)P(z^r_{u_{(B)},w}=t)\log\left((1-\eta_{u_{(r)}})p(w|t)P(t|u_{(B)})\right)$$

Then we maximize the $Q_{\mathcal{C}}$ function with respect to different parameters, i.e., θ and ϕ. It is worth noting that in order to avoid zero probabilities, we have applied Laplace smoothing[4] by adding a small value of α when estimating $P(t|u_{(r)})$, and a small value of β when estimating $P(w|t)$. The updating formulas of the EM algorithm are given in Figure 2.

3.2.2 A Generalized EM Algorithm for rrPLSA

Above, we have shown how to use to learn the model parameters of rPLSA when $\mu=0$. When $\mu\neq 0$, the case is more complicated and cannot be solved by the standard EM algorithm. As such, we adopt the generalized EM (GEM) algorithm to find the solution. GEM does not perform a maximization of $Q(\Psi;\Psi_n)$; instead it tries to find Ψ_{n+1} which increases the Q-function: $Q(\Psi_{n+1})>Q(\Psi_n)$. We rewrite the Q-function as $Q_{\mathcal{C},\mathcal{U}}=Q_{\mathcal{C}}-\mu R(\mathcal{U})$, where $R(\mathcal{U})$ is the regularized factor defined by Eq. 2.

To solve the above Q-function, we have the same E-step for the hidden variables and the M-step for $\{P(w|t)\}$ and $\{\eta_{u_{(r)}}\}$. The major obstacle is that we cannot obtain a closed form solution of $\phi=\{P(t|u_{(r)})\}$ with the incorporation of the regularization factors. The main idea of the GEM algorithm can be summarized as follows. In the $(n+1)$th M-step, we first find $\phi^{(0)}_{n+1}$ using the standard M-step for rPLSA in Figure 2, which maximizes $Q(\mathcal{C})$. The obtained solutions do not necessarily lead to the optimal values of $Q(\mathcal{C},\mathcal{U})$.So the idea is that we start from $\phi^{(0)}_{n+1}$ and decrease $R(\mathcal{U})$ using the Newton-Raphson method.

Given a function $f(x)$(twice differentiable) and the initial value x_t, the Newton-Raphson updating formula to decrease $f(x)$ is defined as: $x_{k+1}=x_k-\delta\frac{f'(x_k)}{f''(x_k)}$, where $0\leq\delta\leq 1$ is the step parameter. We need to run the above iterative algorithm which decreases $R(\mathcal{U})$ by updating $\{P(t|u_{(r)})\}$ in every M-step. Let k be the inner iteration number for minimization of $R(\mathcal{U})$ and n the outer iteration number of the EM algorithm. We repeatedly update ϕ_n and ϑ_n using the corresponding updating equations until $Q_{\mathcal{C},\mathcal{U}}(\phi^{(k)}_{n+1})>Q_{\mathcal{C},\mathcal{U}}(\phi_n)$. If such a stopping condition is met, then we set $(\phi_{n+1}\leftarrow(\phi^{(k)}_{n+1})$. Otherwise, we set $(\phi_{n+1})\leftarrow(\phi_n)$ and continue to the next E-step.

Note the above GEM algorithm is applicable with various distance functions for modeling social interactions. We present the updating equations for RSS and SKL respectively in Figure 3. For RSS function, it is easy to see that for each user $u\in\mathcal{U}$ we have $\sum_t P(t|u_{(r)})^{(k+1)}_{n+1}=1$ and $P(t|u_{(r)})^{(k+1)}_{n+1}>0,\forall t\in\mathcal{T},r\in\mathcal{R}$. And the updating formula in Figure 3 for RSS has intuitive explanations. The new role-specific topic distribution of a user is the old distribution smoothed by the weighted topic distributions of her "neighbors" who interact with her. Furthermore, the neighbors can be divided into different groups corresponding to the value of r_v, i.e., the role that the neighbor plays. While for SKL, the case becomes more complicated. We cannot guarantee that $\sum_t P(t|u_{(r)})^{(k+1)}_{n+1}=1$ and have to normalize it at the end of each inner iteration. We have compared the performance of our models using either RSS or SKL and did not notice any significant difference. As such, we adopt the RSS function to model regularization factors in the remainder of the paper.

4 Instantiation of the Framework on Twitter

In the previous section, we have introduced a general framework for joint modeling roles and topics. In what follows, we will study how to instantiate the framework on Twitter.

On Twitter, each user can be viewed as a social actor who is associated with a set of social roles and a tweet a user published can be viewed as a document of the user. We consider two types of social roles corresponding to two most common social activties on Twitter: (1) *originators* who publish original content; (2) *propagators* who forward and spread content of others, which naturally captures the two most important aspects of Twitter content growth: the generation of new ideas and the spread of existing content.

For the two social roles considered here, $u_{(o)}$ (*originator*) and $u_{(p)}$ (*propagator*), there are four

4. http://en.wikipedia.org/wiki/Additive_smoothing

$$\text{With RSS: } P(t|u_{(r_u)})_{n+1}^{(k+1)} = (1-\delta)P(t|u_{(r_u)})_{n+1}^{(k)} + \delta \frac{\sum_{v\in\mathcal{U}}\sum_{r_v\in\mathcal{R}}\left(s(v_{(r_v)},u_{(r_u)})P(t|v_{(r_v)})_{n+1}^{(k)}\right)}{\sum_{v\in\mathcal{U}}\left(\sum_{r_v\in\mathcal{R}} s(v_{(r_v)},u_{(r_u)})\right)}$$

$$\text{With SKL: } P(t|u_{(r_u)})_{n+1}^{(k+1)} = (1-\delta)P(t|u_{(r_u)})_{n+1}^{(k)} + \delta \frac{\sum_{v\in\mathcal{U}}\sum_{r_v\in\mathcal{R}}\left(s(v_{(r_v)},u_{(r_u)})\left(1+\log\frac{P(t|u_{(r_u)})_{n+1}^{(k)}}{P(t|v_{(r_v)})_{n+1}^{(k)}}\right)\right)}{\sum_{v\in\mathcal{U}}\left(\sum_{r_v\in\mathcal{R}} s(v_{(r_v)},u_{(r_u)})\left(1+\frac{P(t|v_{(r_v)})_{n+1}^{(k)}}{P(t|u_{(r_u)})_{n+1}^{(k)}}\right)\right)} P(t|u_{(r_u)})_{n+1}^{(k)}$$

Fig. 3. Newton-Raphson updating formulas for ϕ 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.

possible forms for our defined similarity function $s(u_{(r_u)}, v_{(r_v)})$, namely $s(u_{(p)}, v_{(p)})$, $s(u_{(o)}, v_{(p)})$, $s(u_{(p)}, v_{(o)})$ 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.

4.1 A two-role topic model

The first problem is how to relate a tweet to a specific role of a user. 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., $u_r =$ "$propagator$". 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

$$L(\mathcal{C}) = \sum_{r\in\{o,p\}}\sum_{u\in\mathcal{U}}\sum_{w\in\mathcal{V}} n_r(u,w)\log\Bigg\{\sum_{t\in\mathcal{T}} P(w|t)\left(\eta_{u_{(r)}}P(t|u_{(r)}) + (1-\eta_{u_{(r)}})P(t|u_{(B)})\right)\Bigg\}, \quad (5)$$

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

4.2 Incorporating social interactions

We describe how to model both explicit and implicit social interactions on Twitter through regularization factors in our proposed general framework below.

4.2.1 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 as follows

$$R_1 = \sum_{a,b\in\mathcal{U}} s(a_{(p)}, b_{(o)})\Big\{\sum_{t\in\mathcal{T}}(P(t|a_{(p)}) - P(t|b_{(o)}))^2\Big\}, \quad (6)$$

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

$$s(a_{(p)}, b_{(o)}) = \frac{n_{a,b}}{n_{a_{(p)}} + n_{b_{(o)}} - n_{a,b}}, \quad (7)$$

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.

4.2.2 Modeling Implicit Interactions

Sometimes, users do not explicitly but implicitly interact with one another. 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 to 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 shown 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 and are described as follows.

Type I: an originator $\leftrightarrow$ common propagators $\leftrightarrow$ another originator (co-retweeted). This type of implicit interactions exists between two originators who are retweeted by some common propagators.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2014.2359672, IEEE Transactions on Knowledge and Data Engineering

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 follows

$$R_2 = \sum_{a,b \in \mathcal{U}} s(a_{(o)}, b_{(o)}) \Big\{ \sum_{t \in \mathcal{T}} (P(t|a_{(o)}) - P(t|b_{(o)}))^2 \Big\}, \quad (8)$$

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

$$s(a_{(o)}, b_{(o)}) = \sum_{c \in \mathcal{U}} \frac{n_{c,a} n_{c,b}}{\sqrt{(\sum_{c'} n_{c',a}^2)(\sum_{c'} n_{c',b}^2)}} \quad (9)$$

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 (co-retweet). 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

$$R_3 = \sum_{a,b \in \mathcal{U}} s(a_{(p)}, b_{(p)}) \Big\{ \sum_{t \in \mathcal{T}} (P(t|a_{(p)}) - P(t|b_{(p)}))^2 \Big\}. \quad (10)$$

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

$$s(a_{(p)}, b_{(p)}) = \sum_{c \in \mathcal{U}} \frac{n_{a,c} n_{b,c}}{\sqrt{(\sum_{c'} n_{a,c'}^2)(\sum_{c'} n_{b,c'}^2)}} \quad (11)$$

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

4.2.3 Integrating the Model with Regularization Factors

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

$$R(\mathcal{U}) = \lambda_1 R_1 + \lambda_2 R_2 + \lambda_3 R_3, \quad (12)$$

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

With both the likelihood function and regularization factors, we can combine them through the Equation 4 in Section 3. The functions of $s(\cdot, \cdot)$ in Eq. 7, Eq. 9 and Eq. 11 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., $K = 30$.

TABLE 1
Statistics of the two Twitter datasets.

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

5 Evaluation of Regularized Two-Role RolePLSA on Twitter Datasets

5.1 Construction of the Datasets

We evaluate our proposed models on two datasets sampled from the Twitter data shared by Kwak et al. [13] 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 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.

Our proposed rrPLSA model is highly motivated by the Social Role Theory. Hence, we would like to gain some insights from social role analysis on the Twitter data. In particular, we want to seek for answers to the following question:

Q: Are there any topical difference between different roles for the same user? How does topical difference vary across different users?

We study the question on $\mathcal{D}_{random}$. We divide users into four groups according to the number of followers they have. Given a user, we further divide all the tweets she posted into two clusters with one cluster consisting of originally-written tweets and the other cluster consisting of retweets. Then we compute the intra-similarity within each of these two clusters (denoted as *O-sim* and *R-sim* respectively) and the inter-similarity between these two clusters (denoted as *OR-sim*).[5] Finally, we average these similarity values over the users in each group. It can be seen from Table 2 that the average inter-similarity is much smaller than average intra-similarities, which indicates that there is indeed a significant topical difference when users play different roles.

5. We represent a tweet as a vector of terms weighted using standard *tf-idf* method, then compute cosine similarities between tweet vectors.

TABLE 2
Difference of topical interests of dual roles on Twitter datasets.

#followers	O-sim	R-sim	OR-sim
<100	0.240	0.296	**0.177**
[100,1000]	0.099	0.164	**0.062**
[1000,10000]	0.042	0.123	**0.014**
>10000	0.036	0.171	**0.007**

5.2 Experimental Setup

We have proposed two models. One is RolePLSA which only considers social roles but ignores social interactions. The other is RegRolePLSA which models both social roles and social interactions. For simplicity, in what follows, RolePLSA and RegRolePLSA are shortened as *rPLSA* and *rrPLSA* respectively. We compare our models with the following topic models:

• *Author-Topic* (AT) Model [14]. We aggregate all the tweets of the same user into one document and run the AT model on such aggregated documents. It has been shown that the AT model is more effective than the standard LDA model on modeling short tweets [15]. We use it as a comparison of rPLSA to examine the impact of social roles.

• *Simple-role PLSA* (srPLSA) Model [16]. For a user, we aggregate her originally-written tweets and retweets respectively into two documents. Then we run the AT model on such aggregated documents and each user will have two topic distributions. srPLSA does not consider the correlation of role-specific interests for the same user, and it is mainly used as a comparison of rPLSA by examining the incorporation of role-specific interest correlations.

• *Enhanced NetPLSA*. Mei et al. [17] proposed the NetPLSA which extends the AT model by incorporating explicit social networks. Although the original work for NetPLSA does not consider implicit links, we incorporate both explicit links and implicit links into NetPLSA in order to better examine the effect of social roles. To allow a fair comparison, we also applied Laplace smoothing with the same smoothing parameters. We called it enhanced NetPLSA.

We now discuss how to set the parameters in our models. The first parameter that requires tuning is the trade-off coefficient μ in Equation 4. μ is set to 1000 which gives the smallest perplexity when $\mu \in [500, 2000]$. Other parameters to set are λ_1,λ_2 and λ_3 in Eq. 12. Our experiments reveal that simply setting $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$ usually gives good performance. For simplicity, we set all λs to be the same for all the subsequent experiments. 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 $\beta = 1e-7$.

For all the mentioned models, we report the results averaged over 10 runs with different random initialization. By varying the number of topics from 10 to 100, we found that these models tend to generate redundant topics when the topic number is larger than 60 and generate too general topics with a topic number smaller than 20. So we only report the results with the number of topics varying between 20 and 60.

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 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 commonly used perplexity measure on held-out documents is adopted as the evaluation metrics of document modeling. A lower perplexity score indicates better generalization performance [18]. 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 $P(\mathbf{w}_d)$.

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. Retweet prediction is a very challenging problem and previous research has proposed a rather complicated model to solve this problem [19]. 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 in higher positions.

Given a user, our proposed topic models can learn the base, the originator-specific and the propagator-specific interest distributions. We use the interpolation

TABLE 3
Performance comparisons of retweet prediction on $\mathcal{D}_{random}$.

Metrics	AT	srPLSA	NetPLSA	rPLSA	rrPLSA
P@10	0.053	0.055	0.057	0.056	**0.062**
P@20	0.100	0.101	0.109	0.103	**0.120**
P@30	0.140	0.148	0.167	0.154	**0.172**
P@100	0.410	0.419	0.443	0.430	**0.467**
MRR	0.160	0.163	0.173	0.165	**0.182**

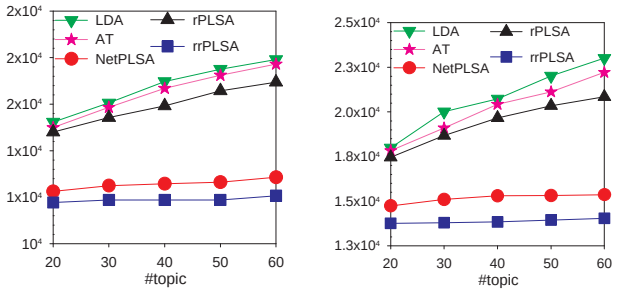
of the base interest and the propagator-specific interest distribution, i.e. $\eta_{u_{(r)}}P(t|u_{(r)}) + (1-\eta_{u_{(r)}})P(t|u_{(B)})$, for retweet prediction. Given a set of topic models $\{\theta_t\}_{t\in\mathcal{T}}$, we compute the conditional probability of topic t given a tweet d for each 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 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 $\mathcal{D}_{random}$.

Experimental results. The results of perplexity and retweet prediction are shown in Figure 4 and Table 3 respectively. It can be observed that in terms of perplexity results, srPLSA has better predictive power than AT by separating originally-written tweets from retweets. Furthermore, rPLSA outperforms srPLSA by explicitly modeling the persistent topical interests shared among multiple roles for a user. Enhanced NetPLSA gives superior performance compared to all the other baselines and also performs better than rPLSA, which shows the effectiveness of incorporating social interactions into the topic models. By additionally modelling both explicit and implicit social interactions as regularization factors, rrPLSA significantly outperforms all the other models by a large margin, including NetPLSA. The best results are achieved using rrPLSA, which improves over the Enhanced NetPLSA by 5.2% in `MRR`. These findings show the effectiveness of our proposed rrPLSA which models both social roles and interactions. In terms of retweet prediction, the performance of all the models is relatively low as revealed by Table 3 due to the fact that we only consider the semantic similarity measured based on topic distributions between users and tweets.

Further analysis of social interactions. The above experiments have shown that rrPLSA performs much better than the other baselines by considering social roles and interactions. We further study the impact of social interactions. Recall we have three parameters to tune in Eq. 12, namely λ_1, λ_2 and λ_3. We consider the



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

Fig. 4. Performance comparisons of perplexity on Twitter datasets.

TABLE 4
The impact of explicit and implicit interactions on $\mathcal{D}_{random}$.

Methods	Perplexity (#topic)			Ret. predict.		Aver.
	20	40	60	P@10	MRR	#link
rPLSA	16081	18175	19157	0.056	0.165	-
rPLSA$_{+e}$	14682	16358	17261	0.057	0.167	6
rPLSA$_{+i,1}$	14376	15649	16246	0.058	0.170	13
rPLSA$_{+i,2}$	14167	15140	15537	0.059	0.174	20
rPLSA$_{+i,1,2}$	13604	14013	14751	0.059	0.176	31
rrPLSA	13210	13297	13571	0.060	0.182	37

following variants of our model: only with explicit interactions (rPLSA$_{+e}$: $\lambda_1 = 1$), only with type-I implicit interactions (rPLSA$_{+i,1}$: $\lambda_2 = 1$), only with type-II implicit interactions (rPLSA$_{+i,2}$: $\lambda_3 = 1$), with type-I/II implicit interactions (rPLSA$_{+i,1,2}$: $\lambda_2 = \lambda_3 = 0.5$), and with all types of interactions (rrPLSA: $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$). We present the results of the above variants on perplexity and retweet prediction in Table 4. We can see that both rPLSA$_{+e}$ and rPLSA$_{+i,1,2}$ perform better than the basic model rPLSA with the improvement obtained using rPLSA$_{+i,1,2}$ being more significant that that obtained using rPLSA$_{+e}$. The major reason is that in online social networks, explicit interactions or links are very sparse. Hence incorporating implicit social interactions seems to be more effective than only considering explicit interactions. In Table 4, we also report the average number of links (interactions) per user. We can see that only 5 explicit links can be used on average. On the other hand, we are able to derive much more implicit interactions between users, and the latter results in more significant performance improvement.These observations confirm that *implicit interactions* are indeed important to leverage for modeling social media content.

We further examine the specific types of implicit interactions. As shown in Table 4, type-II is more effective than type-I since it can capture more implicit links, and a combination of both is better than either one of them. Type I interactions are useful to capture the implicit relationship between popular or authoritative Twitter users. For example, it will be

TABLE 5
Examples of implicit links identified. The first two pairs were identified by type-I interactions while the last two pairs were identified by both type-I and II interactions.

User A	User B
Jörg Tauss (@tauss)	Piratenpartei (@Piratenpartei)
Jörg Tauss is a German politician in the Pirate Party of Germany	
Bonnie Burton(@bonniegrrl)	Star Wars(@starwars)
Bonnie Burton is a former Content Developer in StarWars.com.	
Baratunde (@baratunde)	Liza Sabater (@blogdiva)
Both are famous political bloggers.	
Susan Cooper (@BuzzEdition)	Reg Saddler (@zaibatsu)
Both are social media enthusiasts.	

uncommon to see: 1) Lady Gaga forwards a message from Justin Bieber; or 2) they retweet from common users, so there will be neither explicit nor type-II implicit interactions between them. On the contrary, it is natural to see their tweets be forwarded by some common users, thus type-I interactions can capture such implicit relationship between them through common propagators. Similarly, the type-II interactions can be viewed as endorsements of some common originators and hence implies similar user interests. We present four illustrative examples of user pairs which are identified by implicit interactions but not explicit links in Table 5. The first two pairs have the relation type of organization-member while users of the last two pairs are linked because they exhibit similar topical interests in the Twitter content published.

Indeed, our formulation of social interaction is closely related to two important concepts in social science [20]: *social influence* and *homophily*. Social influence refers to processes in which interactions with others causes individuals to conform, while homophily refers to processes of social selection, where individuals are more likely to form ties with similar others. Thus on one hand, explicit interactions drive users' interests to be similar due to the effect of social influence. On the other hand, the way that we construct virtual links between users based on their implicit interactions is an application of homophily effects, i.e. users tend to be friends with those who share similar interests. As shown in Table 5, the identified implicit links are effective to capture user pairs with similar interests but without explicit interactions, which can be used to improve the task of friend recommendation on online social networks.

6 Evaluation of Regularized Two-Role RolePLSA on cQA Datasets

6.1 Construction of the Datasets

We also evaluate our proposed models on data collected from online question-answering communities, a.k.a cQA. We use two datasets sampled from the *Yahoo! Answer* datasets shared by Mao et al. [21] which has a four-year time span between 2005 and 2008.

TABLE 6
Statistics of the two cQA datasets.

Datasets	#users	#docs	#reply-links
$\mathcal{D}_{random}$	14,713	684,039	27,687
$\mathcal{D}_{hardware}$	14,336	427,448	58,900

TABLE 7
Difference of topical interests of dual roles.

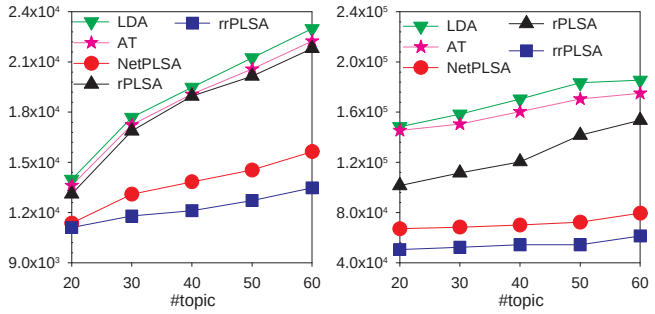
#best answer	Q-sim	A-sim	QA-sim
[0, 1)	0.294	0.362	**0.173**
[1, 5)	0.265	0.255	**0.128**
[5, 10)	0.249	0.187	**0.095**
[10, 100)	0.230	0.126	**0.062**
[100, ∞)	0.196	0.061	**0.030**

Since the cQA datasets come with category labels, we consider a domain-specific dataset in the category of "Hardware" $\mathcal{D}_{hardware}$ and a random dataset $\mathcal{D}_{random}$. For each user, we consider two types of different "documents": the question text she posted as a *question document* and the answer text she provided to questions from others as an *answer document*. In a question thread, an asker would post a question document while other engaged users would play the role of answerers to provide candidate answer documents. We refer to the reply relationship in a thread as a *reply link*. For the answerers to the same question in a question thread, they have *co-reply* links; while for the askers answered by the same answerers in different question threads, they have *co-replied* links. Since we aim to study the effect of social interactions, we discard users with very few question/answer documents or very few reply in/out links. The statistics of the two datasets is summarized in Table 6.

Following a similar setup on the Twitter data (Table 2), we examine the topical difference between different roles for the same user on cQA datasets. We take $\mathcal{D}_{random}$ and divide users into four groups according to the number of best answers they have. Given a user, we further divide all the documents she posted into two clusters with one cluster consisting of question documents and the other answer documents. Then we compute the intra-similarity within each of these two clusters (denoted as *Q-sim* and *A-sim* respectively) and the inter-similarity between these two clusters (denoted as *QA-sim*). It can be seen from Table 7 that the average inter-similarity is much smaller than average intra-similarities, which indicates that there is indeed a significant topical difference when cQA users play different roles.

6.2 Experimental Setup

We can simply use the two-role rrPLSA introduced for Twitter in Section 4 by making the following mappings: (1) A question document $\rightarrow$ an originally-written tweet; (2) An answer document $\rightarrow$ a retweet; (3) A reply link $\rightarrow$ a retweet link; (4) A co-reply



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

Fig. 5. Performance comparisons of perplexity on cQA datasets.

link $\rightarrow$ a co-retweet link; (5) A co-replied link $\rightarrow$ a co-retweeted link. For the experiments on the cQA datasets, we need to set larger smoothing parameters with $\alpha = 1, \beta = 1e{-}3$. For other parameters, we follow a similar experimental setup for Twitter in Section 5.2 and omit the details here.

6.3 Quantitative evaluation

First, we randomly split the cQA datasets into two equal parts: a training set and a test set. Then we perform two quantitative evaluation tasks on the test collection, namely perplexity and question routing.

Perplexity. We follow the perplexity measure introduced in Section 5.3 to evaluate models' predictive power on unseen data. The results of perplexity are shown in Figure 5. It can be observed that in terms of perplexity results, rPLSA and srPLSA has better predictive power than AT by incorporating social roles. By additionally incorporating social interactions as regularization factors, rrPLSA significantly outperforms other models by a large margin. Different from Twitter datasets, all the methods tend to have larger perplexity on $\mathcal{D}_{random}$. One main reason is that questions in $\mathcal{D}_{random}$ (cQA) cover a wide variety of topics and therefore result in larger perplexities. Interestingly, rPLSA gives a much larger gain than AT on $\mathcal{D}_{random}$ (cQA), which indicates that incorporating roles is effective in reducing model perplexity on datasets with diverse topics.

cQA question routing. We also evaluate the performance of our proposed methods on question routing in cQA. Question routing aims to push the right questions to the right persons and thus enables askers obtain quick and high-quality answers [22]. Question routing could potentially depend on many other factors apart from topic similarity such as user experience [23]. Nevertheless, we only consider topic similarity here as we aim to compare our proposed model against other topic models. We simplify the question routing task as follows. We first build a list of candidate users. Then, given a question, we compute the topic similarity between the question and the topical interest of a candidate user from this list. Finally, we rank these users in a descending order. A better method should be able to rank those users who actually answered the question in higher positions. Here comes the question on how to build the list of candidate users for comparison. We could simply select all the users in our cQA datasets as candidates. But it will be very computationally expensive since we have over 14,000 users in our datasets. As such, we build our candidate user list in the following way. For each question, we choose the top 10 most similar questions asked before and add the answerers of these top 10 questions into our candidate list. Then, for the asker of this question, we select the users who have answered at least one question posted by the same asker previously. Finally, we randomly add another one hundred users who have never answered the questions posted by the asker before.

TABLE 8
Performance comparisons of question routing on $\mathcal{D}_{random}$.

Metrics	AT	srPLSA	NetPLSA	rPLSA	rrPLSA
P@10	0.135	0.138	0.140	0.140	**0.146**
P@20	0.185	0.187	0.188	0.189	**0.201**
P@30	0.223	0.227	0.234	0.228	**0.246**
P@100	0.599	0.604	0.619	0.611	**0.631**
MRR	0.155	0.158	0.171	0.159	**0.187**

Following the method used in retweet prediction (Section 5.3), given a question and a set of candidate users, we first compute the negative KL-divergence of the topic distributions of the question and each of candidate user, and subsequently rank these users in a descending order. We still adopt `precision@N` and `MRR` (Mean Reciprocal Rank) as our evaluation metrics. We set the topic number to 40, and only report the results on $\mathcal{D}_{random}$. We present the results of question routing in Table 8. We have similar findings as those in Table 3: 1) rPLSA has better performance than AT and srPLSA; 2) by incorporating social interactions as regularization factors, rrPLSA significantly outperforms other models by a large margin: rrPLSA improves over NetPLSA by 9.3% in terms of `MRR`. These findings show the superiority of our proposed rrPLSA over baseline topic models again.

7 Evaluation of Multiple-Role Role-PLSA on Scientific Paper Datasets

In the previous sections, we have presented the experimental results of applying the proposed two-role rrPLSA model to the Twitter and cQA data. In this section, we consider a more general application scenario where multi-role modeling is needed. Specifically, we build our evaluation dataset from scientific publications. Intuitively, a researcher is likely to be

interested in several research areas and publish papers in multiple conference venues, which somehow indicates the researcher's different expertise areas. If a research expertise area is treated as a role, then we could use multi-role modeling for researchers.

7.1 Dataset Construction

We use the Microsoft Academic Search API[6] to build our dataset, which consists of 3,000 authors, 16,308 papers and 23,194 co-author links. The number of authors is relatively small, because we want to select the authors with a considerable number of papers and who have regularly published papers in different conference venues. By following [24], we define six research areas: Artificial Intelligence (AI), Databases (DB), Data Mining (DM), Graphics, Vision and HCI (GV), Networks, Communications and Performance (NC) and Natural Language Processing (NLP). We then classify 25 major computer science conferences into one of the above six areas.[7] With the above categorization, we can easily identify the corresponding research area of a research paper, i.e. the role that its authors play. We further compute the statistics of users with different numbers of research areas: 1504 (only 1 area), 1104 (2 areas), 316 (3 areas), 63 (4 areas), 13 (5 areas), 0 (6 areas). It is easy to see that quite a few authors publish papers in multiple areas, and as suggested in [24], [25], the users engaged in multiple research areas are likely to be "structural holes".

7.2 Experimental Setup

To apply our model, we first extract the abstract of a paper as a document. Then we group abstracts by authors. If a paper has multiple authors, we associate with each of the authors with the same paper abstract. After these processing steps, each author is associated with a document set consisting of all her published paper abstracts and each document is also assigned with a role label (i.e. research area). We consider the co-author relation as a type of social interactions and the weight is set to be the number of jointly-published papers. It is also possible to consider other types of interactions such as citations. But for simplicity, we only consider co-author relations in our multi-role rrPLSA model here.

7.3 Perplexity Comparison

For perplexity measurement, we randomly split the documents of an author into two equal parts: a training set and a test set. We train the models on the training set and evaluate the models' predictive power on the test set. Here we only compare our models with the best performing baseline, NetPLSA, as has been shown in previous sections. The results of perplexity with topic number varying between 20 and 60 are listed in Table 9. Our observations accord with the earlier findings on both the Twitter and cQA dataset that rrPLSA outperforms rPLSA, which in turn performs better than NetPLSA.

TABLE 9
Perplexity comparison on the scientific paper dataset with topics varying from 20 to 60 with a gap of 10.

Methods	20	30	40	50	60
NetPLSA	2133	2069	2025	2032	1968
rPLSA	2120	2024	1971	1933	1919
rrPLSA	2054	1960	1922	1895	1878

7.4 Expertise-Specific Topic Labeling

In this section, we introduce a novel application of expertise-specific topic labeling based on the results generated from our multi-role rrPLSA model. In particular, we want to produce some keywords which can best characterize topics in a researcher's expertise area based on her papers published. As a researcher may have multiple research interests, existing approaches based on simple keyword labeling [26] may mix up keywords from different topics or different expertise areas. Similar to the idea in [27], we propose to label researchers with topical keywords within each of her expertise areas. Specifically, given a researcher u, for one of her research areas r, we first identify top related topics $\mathcal{T}^{u,r}$ based on the conditional probabilities $P(t|u_{(r)})$. Then for each research topic $t \in \mathcal{T}^{u,r}$, we further generate a list of keywords $\mathcal{W}^{u,r,t}$ ranked by $n_r(u,w) \times P(z_{u_{(r)},w} = t)$, where $n_r(u,w)$ denotes the frequency that u used the word w in research area r, and $P(z_{u_{(r)},w} = t)$ is the posterior probabilities of word w generated by topic t according to the role-specific interest. The top n keywords from $\mathcal{W}^{u,r,t}$ are used as a label of topic t in user u's expertise area r.

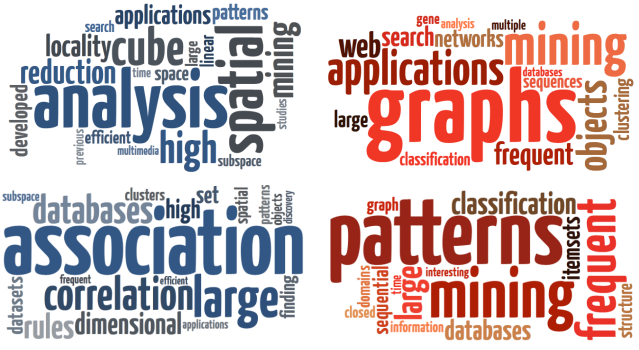
Here we select Chengxiang Zhai and Jiawei Han as case studies, both of them having papers published in multiple research areas. Figure 6 and Figure 7 list the keywords generated for the top two topics in two different expertise areas for Chengxiang Zhai and Jiawei Han respectively. It can be observed that the produced topical keywords characterize the main research topics of these two researchers well. For example, in NLP, Chengxiang Zhai focused on entity-related information retrieval and statistical language translations while in DM, he is more interested in pattern mining and topic models, For Jiawei Han, he has prominent interests on association and spatial analysis in DB while mainly works on graph mining and frequent pattern mining in DM. Note that different research areas may have similar topic coverage, e.g.

6. http://academic.research.microsoft.com/

7. AI: IJCAI, AAAI, ICML, UAI, NIPS, AAMAS; DB: VLDB, SIGMOD, PODS, ICDE, ICDT, EDBT; DM: KDD, ICDM; GV: CVPR, ICCV; NC: SIGCOMM, SIGMETRICS, INFOCOM, MOBICOM; NLP: ACL, EMNLP, COLING, HLT, IJCNLP.

(a) Natural Language Processing (NLP) (b) Data Mining (DM)

Fig. 6. The labels of the top 2 topics in NLP and DM respectively for Professor Chengxiang Zhai.



(a) Database (DB) (b) Data Mining (DM)

Fig. 7. The labels of the top 2 topics in DB and DM respectively for Professor Jiawei Han.

DB and DM. This is mainly due to the classification rules of conferences in [24] are not fully orthogonal.

8 Related Work

Several previous studies have included author or user information when modeling documents using topic models. For example, the Author-Topic (AT) model [14] models a document with multiple authors as a distribution over topics that is a mixture of distributions associated with the authors. Built upon the AT model, the Author-Recipient-Topic (ART) model [28] conditions per-message topic distribution from emails jointly on both the author and individual recipients, rather than on individual authors. The Role-Author-Recipient-Topic (RART) model [28] extends from ART that it assumes an author can have multiple roles and a role is a persona represented by a topic distribution.

While our work is closely related to these studies, it is different from them in the following aspects. First, the aforementioned models assume either a single user specific interest [14] or a shared set of personas or roles among all the users [28]. "Roles" defined in these models are intrinsically different from the social roles defined in our models which essentially correspond to different social activities. As such, the role-based topical interests in our models are both role- and user-specific, which are opposed to the role-based topic distribution shared among all the users as in RART. In addition, ART explicitly characterizes the two roles in a specific relation and is not easy to generalize to multi-role modeling. RART cannot model user-level interactions but only role-level interactions. Furthermore, these models built upon sender-receiver relations assume that there are at least one "sender" and one "receiver" for any document, which is clearly not the case in our datasets. For example, there are many tweets which are never retweeted on Twitter.

In addition to various author/user topic models mentioned above, there has been some work incorporating underlying network structures into topic modeling analysis [17], [29], which showed regularization methods are very promising to deal with links or relationships. Our work is partly inspired by NetPLSA [17] but has three main differences: 1) our focus is to model role-specific interests of users and we propose a principled approach which incorporates SRT into the generative process of topic models; 2) we jointly consider users' roles when modeling social interactions; 3) we propose to use implicit interactions to leverage latent similarities of user interests. Recently, there is a considerable body of research which focuses on topic modeling on tweets [2], [30]. There have also been some studies analyzing users' retweeting behaviors when performing topical analysis in Twitter [31]. However, they only focus on finding topical authorities and do not model both users and topics.

9 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 social media content in a new perspective, i.e., Social Role Theory, and it reflects the key elements of SRT. Experimental results on the two Twitter datasets, two cQA datasets and a scientific paper dataset show that rrPLSA outperforms a few competitive baselines, including AT, srPLSA and NetPLSA. These findings confirm the feasibility of incorporating the Social Role Theory for social media content analysis and shed lights on future research directions of online content generation.

There are several possible directions to pursue for future work. In this paper, we only focus on roles corresponding to common social activities. 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 and MySpace, where users play different roles, too.

REFERENCES

[1] C. Wang and D. Blei, "Collaborative topic modeling for recommending scientific articles," in *KDD*, 2011.

[2] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis, "Discovering geographical topics in the twitter stream," in *WWW*, 2012.

[3] B. Biddle, "Recent development in role theory," *Annual review of sociology*, pp. 67–92, 1986.

[4] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *43rd HICSS*, 2010, pp. 1–10.

[5] S. Golder and J. Donath, "Social roles in electronic communities," *Internet Research*, vol. 5, pp. 19–22, 2004.

[6] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *SNA-KDD*, 2007.

[7] T. Hofmann, "Probabilistic latent semantic indexing," in *SIGIR*, 1991.

[8] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *WSDM*, 2010.

[9] M. J. Welch, U. Schonfeld, D. He, and J. Cho, "Topical semantics of twitter links," in *WSDM*, 2011.

[10] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, "Recommender systems with social regularization," in *WSDM*, 2011.

[11] H. Ma, M. R. Lyu, and I. King, "Learning to recommend with trust and distrust relationships," in *RecSys*, 2009.

[12] X. Zhu, "Semi-supervised learning with graphs," Ph.D. dissertation, 2005.

[13] H. Kwak, C. Lee, H. Park, and S. B. Moon, "What is twitter, a social network or a news media?" in *WWW*, 2010.

[14] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. L. Griffiths, "Probabilistic author-topic models for information discovery," in *KDD*, 2004.

[15] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *SOMA*, 2010.

[16] X. W. Zhao, J. Wang, Y. He, J.-Y. Nie, and X. Li, "Originator or propagator?: incorporating social role theory into topic models for twitter content analysis," in *CIKM*, 2013.

[17] Q. Mei, D. Cai, D. Zhang, and C. Zhai, "Topic modeling with network regularization," in *WWW*, 2008.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, 2003.

[19] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," in *WWW*, 2011.

[20] T. La Fond and J. Neville, "Randomization tests for distinguishing social influence and homophily effects," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10, 2010.

[21] X.-L. Mao, Z.-Y. Ming, T.-S. Chua, S. Li, H. Yan, and X. Li, "Sshlda: A semi-supervised hierarchical topic model," in *EMNLP*, 2012.

[22] Y. Zhou, G. Cong, B. Cui, C. S. Jensen, and J. Yao, "Routing questions to the right users in online communities," in *ICDE*, 2009.

[23] B. Li and I. King, "Routing questions to appropriate answerers in community question answering services," in *CIKM*, 2010.

[24] T. Lou and J. Tang, "Mining structural hole spanners through information diffusion in social networks," in *WWW*, 2013.

[25] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li, "Rolx: Structural role extraction & mining in large graphs," in *KDD*, 2012.

[26] Z. Liu, X. Chen, and M. Sun, "Mining the interests of chinese microbloggers via keyword extraction," *Front. Comput. Sci China*, vol. 6, no. 1, pp. 76–87, Feb. 2012.

[27] X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.-P. Lim, and X. Li, "Topical keyphrase extraction from twitter," in *ACL-HLT*, 2011.

[28] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "Topic and role discovery in social networks," in *IJCAI*, 2005.

[29] H. Deng, J. Han, B. Zhao, Y. Yu, and C. X. Lin, "Probabilistic topic models with biased propagation on heterogeneous information networks," in *KDD*, 2011.

[30] D. Ramage, S. T. Dumais, and D. J. Liebling, "Characterizing microblogs with topic models," in *ICWSM*, 2010.

[31] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," in *WWW*, 2011.

Wayne Xin Zhao is currently a PhD student at the School of Electronic Engineering and Computer Science, Peking University, China. He received his BEng degree in Computer Science from Harbin Institute of Technology in 2008, China. His research interests are web text mining and natural language processing. He has published several referred papers in international conferences such as ACL, EMNLP, ECIR, CIKM and SIGIR.

Jinpeng Wang is currently a PhD student at the School of Electronic Engineering and Computer Science, Peking University, China. He received his BEng degree in Computer Science from China University of Geosciences in 2010, China. His research mainly focuses on social media content analysis, especially on commercial intent detection.

Yulan He is a Senior Lecturer at the School of Engineering and Applied Science, Aston University, UK. She received her PhD degree from Cambridge University working on statistical models to spoken language understanding. She has published over 100 papers with many in high-impact journals and top conferences. Her research interests include natural language processing, statistical modelling, text and data mining, sentiment analysis, and social media analysis.

Jian-Yun Nie is a professor at the Computer Science Department in Université de Montréal, Canada. He has published more than 150 research papers in information retrieval and natural language processing in journals and conferences. He has served as a general co-chair of the ACM-SIGIR conference in 2011. He is currently on the editorial board of seven international journals. He has been an invited professor and researcher at several universities and companies.

Ji-Rong Wen is a professor at the School of Information, Renmin University of China. Before that, he was a senior researcher and group manager of the Web Search and Mining Group at MSRA since 2008. He has published extensively on prestigious international conferences/journals and served as program committee members or chairs in many international conferences. He was the chair of the "WWW in China" track of the 17th World Wide Web conference. He is currently the associate editor of ACM Transactions on Information Systems (TOIS).

Xiaoming Li is a professor at the School of Electronic Engineering and Computer Science and the director of Institute of Network Computing and Information Systems in Peking University, China. He is a Senior member of IEEE and currently served as Vice president of China Computer Federation. His research interests include search engine and web mining, and web technology enabled social sciences.