Collaborative, Dynamic and Diversified User Profiling

Shangsong Liang[1,2]
[1]School of Data and Computer Science, Sun Yat-sen University, China
[2]Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 51006, China
liangshangsong@gmail.com

Abstract

In this paper, we study the problem of dynamic user profiling in the context of streams of short texts. Previous work on user profiling works with long documents, do not consider collaborative information, and do not diversify the keywords for profiling users' interests. In contrast, we address the problem by proposing a user profiling algorithm (UPA), which consists of two models: the proposed collaborative interest tracking topic model (CITM) and the proposed streaming keyword diversification model (SKDM). UPA first utilizes CITM to collaboratively track each user's and his followees' dynamic interest distributions in the context of streams of short texts, and then utilizes SKDM to obtain top-k relevant and diversified keywords to profile users' interests at a specific point in time. Experiments were conducted on a Twitter dataset and we found that UPA outperforms state-of-the-art non-dynamic and dynamic user profiling algorithms.

Introduction

To capture users' dynamic interests underlying their posts in microblogging platforms such as Twitter is of importance to the success of further design of applications that cater for users of such platforms, such as dynamic user clustering (Liang et al. 2017a; 2017b). In this paper, we study the problem of *user profiling for streaming short documents (Balog et al. 2012; Liang 2018; Liang et al. 2018): collaboratively identifying users' dynamic interests and tracking how they evolve over time in the context of streaming short texts*. Our goal is to infer users' and their collaborative topic distributions over time and dynamically profile their interests with a set of diversified keywords in the context of streaming short texts.

The first user profiling model was proposed in (Balog et al. 2007), where a set of relevant keywords were identified for each user in a static collection of long documents and the dynamics of users' interests were ignored. Recent work realize the importance of capturing users' dynamic interests over time and a number of temporal profiling algorithms have been proposed for streams of long documents. However, previous work on user profiling suffer from the following problems: (1) They work with streams of long documents rather than short documents and made the assumption that the content of documents is rich enough to infer users' dynamic interests. (2) They ignore any collaborative information, such as friends' messages when inferring users' interests at a specific point in time. (3) They just simply retrieve a list of top-k keywords as a user's profile that may be semantically similar to each other and thus redundant.

Accordingly, in this paper, to address the aforementioned drawbacks in the previous work, we propose a User Profiling Algorithm in the context of streams of short documents, abbreviated as **UPA**, which is *collaborative, dynamic and diversified*. UPA consists of two proposed models– a Collaborative Interest Tracking topic Model, abbreviated as **CITM**, and a Streaming Keyword Diversification Model, abbreviated as **SKDM**. UPA algorithm first utilizes our proposed CITM to track the changes of users' dynamic interests in the context of streams of short documents. It then utilizes our proposed SKDM to produce top-k diversified keywords for profiling users' interests at a specific point in time.

Our CITM topic model works with streaming short texts and is a dynamic multinomial Dirichlet mixture topic model that is able to infer and track each user's ***dynamic*** interest distributions based not only on the user's posts but also the ***collaborative*** information, i.e., his followees' posts. Our hypothesis in CITM is that accounting for collaborative information is critical, especially for those users with limited activities, infrequent short posts, and thus sparse information. To perform the inference of users' interest distributions in streams of short documents, we propose a collapsed Gibbs sampling algorithm. Then, our SKDM model works with users' dynamic interest distributions produced by CITM and aims at retrieving a set of relevant and also ***diversified*** keywords for profiling users' interests at time t such that redundancy of the retrieved keywords can be avoided while still keeping relevant keywords to profile the users.

Our contributions are: (1) We propose a user profiling algorithm, UPA, to address the user profiling task in the context of streams of short documents. (2) We propose a topic model, CITM, that can collaboratively and dynamically track each user's and his followees' interests. (3) We propose a collapsed Gibbs sampling algorithm to infer users' and his followees' interest distributions. (4) We propose a streaming keyword diversification model, SKDM, to diversify the top-k keywords as users' profiling results at time t.

Related Work

User Profiling. User profiling has been gaining attention after the launch of user finding task at TREC 2005 enterprise track (Craswell, de Vries, and Soboroff 2005). Balog and de Rijke (2007) worked with a static long document corpus and modeled the profile of a user as a set of relevant keywords. Recent work were aware of the importance of temporal user profiling. Temporal profiling for long documents was first introduced in (Rybak, Balog, and Nørvåg 2014), where topical areas were organized in a predefined taxonomy and interests were represented as a weighted unchanged tree built by the ACM classification system. A probabilistic model was proposed in (Fang and Godavarthy 2014), where experts' academic publications were used to investigate how personal interests evolve over time. We follow most previous work, and retrieve top-k words as profile of a user's interests.

Topic Modeling. Topic models provide a suite of algorithms to discover hidden thematic structure in a collection of documents. A topic model takes a set of documents as input, and discovers a set of "latent topics"—recurring themes that are discussed in the collection—and the degree to which each document exhibits those topics (Blei, Ng, and Jordan 2003). Since the well-known topic models, PLSI (Probabilistic Latent Semantic Indexing) (Hofmann 1999) and LDA (Latent Dirichlet Allocation) (Blei, Ng, and Jordan 2003), were proposed, topic models with dynamics have been widely studied. These include Dynamic Topic Model (Blei and Lafferty 2006), Dynamic Mixture Model (Wei, Sun, and Wang 2007), Topic over Time (Wang and McCallum 2006), Topic Tracking Model (Iwata et al. 2009), and more recently, dynamic Dirichlet multinomial mixture topic model (Liang et al. 2017c), user expertise tracking topic model (Liang 2018) and user collaborative interest tracking topic model (Liang, Yilmaz, and Kanoulas 2018). To our knowledge, none of existing dynamic topic models has considered the problem of user profiling for short texts that utilizes collaborative information to infer topic distributions.

Problem Formulation

We follow most of the previous work (Balog and de Rijke 2007; Berendsen et al. 2013; Liang et al. 2018), and retrieve top-k words as profile of a user. Then, the problem we address in this paper is: given a set of users and streams of short documents generated by them, track their interests over time and dynamically identify a set of top-k relevant and diversified keywords to each of the users. The dynamic user profiling algorithm is essentially a function h that satisfies:

$$\mathbf{D}_t, \mathbf{u}_t \xrightarrow{h} \mathbf{W}_t,$$

where $\mathbf{D}_t = \{\ldots, \mathbf{d}_{t-2}, \mathbf{d}_{t-1}, \mathbf{d}_t\}$ represents the *stream* of short documents generated by the users $\mathbf{u}_t$ up to time t with $\mathbf{d}_t$ being the most recent set of short documents arriving at t, $\mathbf{u}_t = \{u_1, u_2, \ldots, u_{|\mathbf{u}_t|}\}$ represents a set of users appearing in the stream up to time t, with u_i being the i-th user in $\mathbf{u}_t$ and $|\mathbf{u}_t|$ being the total number of users in the user set, and $\mathbf{W}_t = \{\mathbf{w}_{t,u_1}, \mathbf{w}_{t,u_2}, \ldots, \mathbf{w}_{u_{t,|\mathbf{u}_t|}}\}$ represents all users' profiling results at t with $\mathbf{w}_{t,u_i} = \{w_{t,u_i,1}, w_{t,u_i,2}, \ldots, w_{t,u_i,k}\}$ being the profiling result, i.e., the top-k diversified keywords, for user u_i at t. We assume that the length of a document d in $\mathbf{D}_t$ is no more than a predefined small length (e.g., 140 characters in Twitter).

User Profiling Algorithm

In this section, we detail our proposed User Profiling Algorithm (**UPA**) that consists of the proposed Collaborative Interest Tracking topic Model (**CITM**) and the proposed Streaming Keyword Diversification Model (**SKDM**).

Overview

We model users' interests in streams by latent topics. Therefore, the dynamic interests of each user $u \in \mathbf{u}_t$ at time period t can be represented as a multinomial distribution $\boldsymbol{\theta}_{t,u}$ over topics, where $\boldsymbol{\theta}_{t,u} = \{\theta_{t,u,z}\}_{z=1}^{Z}$ with $\theta_{t,u,z}$ being the interest score on topic z for user u at time period t and Z being the total number of latent topics. Similarly, the dynamic interests of each user's followees at t can be represented as a multinomial distribution $\boldsymbol{\psi}_{t,u} = \{\psi_{t,u,z}\}_{z=1}^{Z}$ with $\psi_{t,u,z}$ being the interest score of user u's followees $\mathbf{f}_{t,u}$ as a whole on topic z at t. Here, $\mathbf{f}_{t,u}$ denotes user u's all followees at t.

Our UPA algorithm consists of two main steps: (1) UPA first utilizes the proposed CITM to capture each user's dynamic interests $\boldsymbol{\theta}_{t,u}$ and his collaborative interests $\boldsymbol{\psi}_{t,u}$. (2) Given $\boldsymbol{\theta}_{t,u}$ and $\boldsymbol{\psi}_{t,u}$ having been inferred, UPA then utilizes SKDM to identify top-k relevant and diversified keywords for profiling the user u's dynamic interests at time period t.

Collaborative Interest Tracking Model

Modeling Interests over Time. We close follow the previous work in (Liang, Yilmaz, and Kanoulas 2018; Liang et al. 2017a), and aim at inferring each user's dynamic interest distribution $\boldsymbol{\theta}_{t,u} = \{\theta_{t,u,z}\}_{z=1}^{Z}$ and his collaborative interest distribution $\boldsymbol{\psi}_{t,u} = \{\psi_{t,u,z}\}_{z=1}^{Z}$ at t in the context of streams of short documents in our CITM. We provide CITM's graphical representation in Fig. 1.

To track the dynamics of a user u's interests, we assume that the mean of his current interests $\boldsymbol{\theta}_{t,u}$ at time period t is the same as that at $t-1$, unless otherwise newly arrived documents associated with the user u in the streams can be observed. With this assumption and following the previous work on dynamic topic models (Iwata et al. 2010; 2009; Wei, Sun, and Wang 2007), we use the following Dirichlet prior with a set of precision values $\boldsymbol{\alpha}_t = \{\alpha_{t,z}\}_{z=1}^{Z}$, where we let the mean of the current distribution $\boldsymbol{\theta}_{t,u}$ depend on the mean of the previous distribution $\boldsymbol{\theta}_{t-1,u}$ as:

$$P(\boldsymbol{\theta}_{t,u}|\boldsymbol{\theta}_{t-1,u}, \boldsymbol{\alpha}_t) \propto \prod_{z=1}^{Z} \theta_{t,u,z}^{\alpha_{t,u,z}\theta_{t-1,u,z}-1}, \qquad (1)$$

where the precision value $\alpha_{t,z} = \{\alpha_{t,u,z}\}_{u=1}^{|\mathbf{u}_t|}$ represents the persistency of users' interests, which is how saliency topic z is at time period t in contrast to that at $t-1$ for the users. As the distribution is a conjugate prior of the multinomial distribution, the inference is able to performed by Gibbs sampling (Liu 1994). Similarly, to track the dynamic changes of a user u's collaborative interest distribution $\boldsymbol{\psi}_{t,u}$, we use

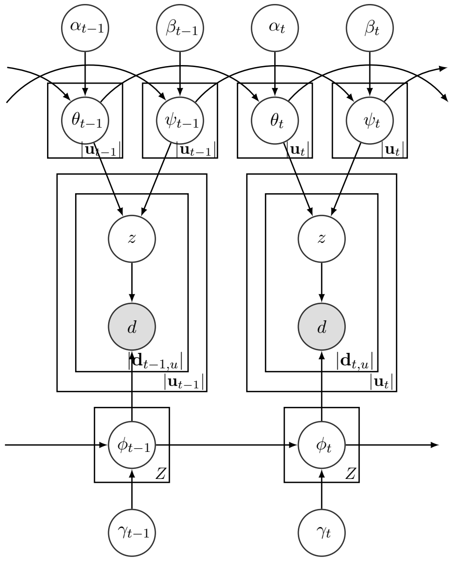

Figure 1: Graphical representation of our proposed CITM model. Shaded nodes represent observed variables.

the following Dirichlet prior with a set of precision values $\boldsymbol{\beta}_t = \{\beta_{t,z}\}_{z=1}^{Z}$, where the mean of the current distribution $\boldsymbol{\psi}_{t,u}$ evolves from that of the previous distribution $\boldsymbol{\beta}_{t-1,u}$:

$$P(\boldsymbol{\psi}_{t,u}|\boldsymbol{\psi}_{t-1,u},\boldsymbol{\beta}_t) \propto \prod_{z=1}^{Z} \psi_{t,u,z}^{\beta_{t,u,z}\psi_{t-1,u,z}-1}, \tag{2}$$

where the precision value $\beta_{t,z} = \{\beta_{t,u,z}\}_{u=1}^{|\mathbf{u}_t|}$ represents the persistency of users' collaborative interest, which is how saliency topic z is at time period t in contrast to that at $t-1$ for the users. In a similar way, to model the dynamic changes of the multinomial distribution of words specific to topic z, we assume a Dirichlet prior, in which the mean of the current distribution $\boldsymbol{\phi}_{t,z} = \{\phi_{t,z,v}\}_{v=1}^{V}$ evolves from the mean of the previous distribution $\boldsymbol{\phi}_{t-1,z}$:

$$P(\boldsymbol{\phi}_{t,z}|\boldsymbol{\phi}_{t-1,z},\boldsymbol{\gamma}_t) \propto \prod_{v=1}^{V} \phi_{t,z,v}^{\gamma_{t,z,v}\phi_{t-1,z,v}-1}, \tag{3}$$

where V is the total number of words in a vocabulary $\mathbf{v} = \{v_i\}_{i=1}^{V}$ and $\boldsymbol{\gamma}_t = \{\gamma_{t,v}\}_{v=1}^{V}$, with $\gamma_{t,v} = \{\gamma_{t,z,v}\}_{z=1}^{Z}$ representing the persistency of the word v in all topics at time t, a measure of how consistently the word belongs to the topics at t compared to that at $t-1$. Later in this subsection, we propose a collapsed Gibbs sampling algorithm to infer all users' dynamic interest distributions $\boldsymbol{\Theta}_t = \{\boldsymbol{\theta}_{t,u}\}_{u=1}^{|\mathbf{u}_t|}$, their corresponding dynamic collaborative interest distributions $\boldsymbol{\Psi}_t = \{\boldsymbol{\psi}_{t,u}\}_{u=1}^{|\mathbf{u}_t|}$, and the words' dynamic topic distributions $\boldsymbol{\Phi}_t = \{\boldsymbol{\phi}_{t,z}\}_{z=1}^{Z}$, and describe our update rules to obtain the optimal persistency values $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$.

Assume that we know all users' interest distribution at time $t-1$, $\boldsymbol{\Theta}_{t-1}$, their collaborative interest distribution at time $t-1$, $\boldsymbol{\Psi}_{t-1}$, and the words' topic distribution, $\boldsymbol{\Phi}_{t-1}$. Then the proposed collaborative interest tracking model is essentially a generative topic model that depends on $\boldsymbol{\Theta}_{t-1}$, $\boldsymbol{\Psi}_{t-1}$ and $\boldsymbol{\Phi}_{t-1}$. For initialization and without loss of generalization, we let $\theta_{0,u,z} = 1/Z$, $\psi_{0,u,z} = 1/Z$ and $\phi_{0,z,v} = 1/V$ at $t = 0$. Let all the short documents posted by user u at time period t denote as $\mathbf{d}_{t,u}$. The generative process of our model for documents in stream at time t, is as follows,

Algorithm 1: Inference for our CITM model at time t.

```
Input  : Distributions $\boldsymbol{\Theta}_{t-1}$, $\boldsymbol{\Psi}_{t-1}$ and $\boldsymbol{\Phi}_{t-1}$ at $t-1$;
         Initialized $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$; Number of iterations
         $N_{iter}$.
Output: Current distributions $\boldsymbol{\Theta}_t$, $\boldsymbol{\Psi}_t$ and $\boldsymbol{\Phi}_t$.
Initialize topic assignments randomly for all documents
  in $\mathbf{d}_t$
for iteration = 1 to $N_{iter}$ do
    for user = 1 to $|\mathbf{u}_t|$ do
        for $d$ = 1 to $\mathbf{d}_{t,u}$ do
            Draw $z_{t,u,d}$ from (5)
            Update $m_{t,u,z_{t,u,d}}$, $o_{t,u,z_{t,u,d}}$ and $n_{t,z_{t,u,d},v}$
    Update $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$
Compute the posterior estimates $\boldsymbol{\Theta}_t$, $\boldsymbol{\Psi}_t$ and $\boldsymbol{\Phi}_t$.
```

i. Draw Z multinomials $\boldsymbol{\phi}_{t,z}$, one for each topic z, from a Dirichlet prior distribution $\boldsymbol{\gamma}_{t,z}\boldsymbol{\phi}_{t-1,z}$;

ii. For each user $u \in \mathbf{u}_t$, draw multinomials $\boldsymbol{\theta}_{t,u}$ and $\boldsymbol{\psi}_{t,u}$ from Dirichlet distributions with priors $\boldsymbol{\alpha}_{t,u}\boldsymbol{\theta}_{t-1,u}$ and $\boldsymbol{\beta}_{t,u}\boldsymbol{\psi}_{t-1,u}$, respectively;

iii. For each document $d \in \mathbf{d}_{t,u}$, draw a topic z_d based on the mixture of $\boldsymbol{\theta}_{t,u}$ and $\boldsymbol{\psi}_{t,u}$, and then for each word v_d in the document d:

   (a) Draw a word v_d from multinomial ϕ_{t,z_d}.

In the above generative process, given the documents in streams are short, and because most of the short documents are likely to talk about one single topic only (Yin and Wang 2014), *we let all the words in the same document* d *be drawn from the multinomial distribution associated with the same topic* z_d. See the graphical representation of CITM in Fig. 1.

Interest Distribution Inference. We propose a collapsed Gibbs sampling algorithm that is based on the basic collapsed Gibbs sampler (Griffiths and Steyvers 2004; Wallach 2006) to approximately infer the distributions in our CITM topic model. As shown in Fig. 1 and the generative process, we adopt a conjugate prior (Dirichlet) for the multinomial distributions, and thus we can easily integrate out the uncertainty associated with multinomials $\boldsymbol{\theta}_{t,u}$, $\boldsymbol{\psi}_{t,u}$ and $\boldsymbol{\phi}_{t,z}$.

We provide an overview of our proposed collapsed Gibbs sampling algorithm in Algorithm 1, where we denote $m_{t,u,z}$, $o_{t,u,z}$ and $n_{t,z,v}$ to be the number of documents assigned to topic z for user u, the number of documents assigned to topic z for user u's followees and the number of times word v assigned to topic z for user u at t, respectively. In the Gibbs sampling procedure, we need to calculate the conditional distribution $P(z_{t,u,d}|\mathbf{z}_{t,-(u,d)}, \mathbf{d}_t, \boldsymbol{\Theta}_{t-1}, \boldsymbol{\Psi}_{t-1}, \boldsymbol{\Phi}_{t-1}, \mathbf{u}_t, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t)$ at time t, where $\mathbf{z}_{t,-(u,d)}$ represents the topic assignments for all the documents in $\mathbf{d}_t$ except the document $d \in \mathbf{d}_{t,u}$ associated with user u at t, and $z_{t,u,d}$ is the topic assigned to the document $d \in \mathbf{d}_{t,u}$. For obtaining this conditional distribution used during sampling, we begin with the joint probability of the current document set, $P(\mathbf{z}_t, \mathbf{d}_t|\mathbf{\Theta}_{t-1}, \mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \mathbf{u}_t, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t)$ at time t:

$$\begin{aligned}
&P(\mathbf{z}_t, \mathbf{d}_t|\mathbf{\Theta}_{t-1}, \mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \mathbf{u}_t, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t) \qquad (4)\\
&= (1-\lambda)P(\mathbf{z}_t, \mathbf{d}_t|\mathbf{\Theta}_{t-1}, \mathbf{\Phi}_{t-1}, \mathbf{u}_t, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t)\\
&\quad +\lambda P(\mathbf{z}_t, \mathbf{d}_t|\mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \mathbf{u}_t, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t)\\
&= (1-\lambda)\prod_z \left(\frac{\mathit{\Gamma}(\sum_v(\varkappa_b))}{\prod_v \mathit{\Gamma}(\varkappa_b)}\frac{\prod_v \mathit{\Gamma}(\varkappa_a)}{\mathit{\Gamma}(\sum_v \varkappa_a)}\right)\cdot\prod_u \frac{\mathit{\Gamma}(\sum_z(\varkappa_2))}{\prod_z \mathit{\Gamma}(\varkappa_2)}\frac{\prod_z \mathit{\Gamma}(\varkappa_1)}{\mathit{\Gamma}(\sum_z \varkappa_1)}\\
&\quad +\lambda\prod_z \left(\frac{\mathit{\Gamma}(\sum_v(\varkappa_b))}{\prod_v \mathit{\Gamma}(\varkappa_b)}\frac{\prod_v \mathit{\Gamma}(\varkappa_a)}{\mathit{\Gamma}(\sum_v \varkappa_a)}\right)\cdot\prod_u \frac{\mathit{\Gamma}(\sum_z(\varkappa_4))}{\prod_z \mathit{\Gamma}(\varkappa_4)}\frac{\prod_z \mathit{\Gamma}(\varkappa_3)}{\mathit{\Gamma}(\sum_z \varkappa_3)},
\end{aligned}$$

where $\mathit{\Gamma}(\cdot)$ is a gamma function, λ is a free parameter that governs the linear mixture of a user's interests and his followees' interests, and the set of parameters $\varkappa$ are defined as: $\varkappa_1 = m_{t,u,z} + \alpha_{t,z}\overline{\theta}$, $\varkappa_2 = \alpha_{t,u,z}\overline{\theta}$, $\varkappa_3 = o_{t,u,z} + \beta_{t,z}\overline{\psi}$, $\varkappa_4 = \beta_{t,u,z}\overline{\psi}$, $\varkappa_a = n_{t,z,v} + \gamma_{t,v}\overline{\phi}$, and $\varkappa_b = \gamma_{t,z,v}\overline{\phi}$. Here, we let $\overline{\theta}$, $\overline{\psi}$ and $\overline{\phi}$ abbreviate for $\theta_{t-1,u,z}$, $\psi_{t-1,u,z}$ and $\phi_{t-1,z,v}$, respectively. Based on the above joint distribution (4) and using the chain rule, we can obtain the following conditional distribution conveniently for the proposed Gibbs sampling (step 5 of Algorithm 1) as the following:

$$\begin{aligned}
&P(z_{t,u,d} = z|\mathbf{z}_{t,-(u,d)}, \mathbf{d}_t, \mathbf{\Theta}_{t-1}, \mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \mathbf{u}_t, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t) =\\
&\left((1-\lambda)\frac{m_{t,u,z} + \alpha_{t,u,z}\overline{\theta} - 1}{\sum_{z=1}^{Z}(m_{t,u,z} + \alpha_{t,u,z}\overline{\theta}) - 1} + \right.\\
&\quad\left.\lambda\frac{o_{t,u,z} + \beta_{t,u,z}\overline{\psi} - 1}{\sum_{z=1}^{Z}(o_{t,u,z} + \beta_{t,u,z}\overline{\psi}) - 1}\right)\\
&\times\frac{\prod_{v\in d}\prod_{j=1}^{N_{d,v}}(n_{t,z,v,-(u,d)} + \gamma_{t,z,v}\overline{\phi} + j - 1)}{\prod_{i=1}^{N_d}(n_{t,z,-(u,d)} + i - 1 + \sum_{v=1}^{V}\gamma_{t,z,v}\overline{\phi})}, \qquad (5)
\end{aligned}$$

where N_d, $N_{d,v}$, $\mathbf{z}_{t,-(u,d)}$, $n_{t,z,v,-(u,d)}$ and $n_{t,z,-(u,d)}$ are the length of document d, the number of word v appearing in d, topic assignments for all documents except the document d from user u at t, the number of word v assigned to topic z in all documents except the one from user u at t, and the number of documents assigned to z in all documents except the one from user u at t, respectively. At each iteration during the sampling (steps 2 to 7 of Algorithm 1), the precision parameters $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$ can be estimated by maximizing the joint distribution (4). We apply fixed-point iterations to obtain the optimal $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$. By applying the two bounds in (Minka 2000), we can derive the following update rules of $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$ for maximizing the joint distribution in our fixed-point iterations:

$$\begin{aligned}
\alpha_{t,u,z} &\leftarrow \frac{(1-\lambda)\alpha_{t,u,z}\left(\mathit{\Delta}(m_{t,u,z} + \alpha_{t,u,z}\overline{\theta}) - \mathit{\Delta}(\alpha_{t,u,z}\overline{\theta})\right)}{\mathit{\Delta}(\sum_{z=1}^{Z} m_{t,u,z} + \alpha_{t,u,z}\overline{\theta}) - \mathit{\Delta}(\sum_{z=1}^{Z}\alpha_{t,u,z}\overline{\theta})},\\
\beta_{t,u,z} &\leftarrow \frac{\lambda\beta_{t,u,z}\left(\mathit{\Delta}(o_{t,u,z} + \beta_{t,u,z}\overline{\psi}) - \mathit{\Delta}(\beta_{t,u,z}\overline{\psi})\right)}{\mathit{\Delta}(\sum_{z=1}^{Z} o_{t,u,z} + \beta_{t,u,z}\overline{\psi}) - \mathit{\Delta}(\sum_{z=1}^{Z}\beta_{t,u,z}\overline{\psi})}, \qquad (6)\\
\gamma_{t,z,v} &\leftarrow \frac{\gamma_{t,z,v}\left(\mathit{\Delta}(n_{t,z,v} + \gamma_{t,z,v}\overline{\phi}) - \mathit{\Delta}(\gamma_{t,z,v}\overline{\phi})\right)}{\mathit{\Delta}(\sum_{v=1}^{V} n_{t,z,v} + \gamma_{t,z,v}\overline{\phi}) - \mathit{\Delta}(\sum_{v=1}^{V}\gamma_{t,z,v}\overline{\phi})},
\end{aligned}$$

where $\mathit{\Delta}(x) = \frac{\partial\log\mathit{\Gamma}(x)}{x}$ is a Digamma function. Our derivations of the update rules for $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$ in (6) are analogous to those in (Liang, Yilmaz, and Kanoulas 2018; Liang et al. 2017c; 2017b).

Algorithm 2: SKDM model for generating top-k keywords for collaborative, dynamic, diversified user profiling.

```
Input  : Current distributions $\mathbf{\Theta}_t$ and $\mathbf{\Phi}_t$
Output: All users' profiling results at time $t$, $\mathbf{W}_t$
1  for $u = 1, \ldots, |\mathbf{u}_t|$ do
2      $\mathbf{w}_{t,u} \leftarrow \varnothing$                    /* $\mathbf{w}_{t,u} \in \mathbf{W}_t$ */
3      $\widetilde{\mathbf{v}} \leftarrow \mathbf{v}$
4      for $z = 1, \ldots, Z$ do
5          $\delta_{t,u,z} \leftarrow (1-\lambda)P(z|t,u) + \lambda P(z|t,\mathbf{f}_{t,u})$
6          $s_{z|t,u} \leftarrow 0$
7      for all positions in the ranked list $\mathbf{w}_{t,u}$ do
8          for $z = 1, \ldots, Z$ do
9              $qt[z|t,u] = \frac{\delta_{t,u,z}}{2s_{z|t,u}+1}$
10         $z^* \leftarrow \arg\max_z qt[z|t,u]$
11         $v^* \leftarrow \arg\max_{v\in\widetilde{\mathbf{v}}}\ \eta_1 \times qt[z^*|t,u] \times$
             $P(v|t,z^*) + \eta_2\sum_{z\neq z^*} qt[z|t,u] \times$
             $P(v|t,z) + (1-\eta_1-\eta_2) \times \mathrm{tfidf}(v|t,u)$
12         $\mathbf{w}_{t,u} \leftarrow \mathbf{w}_{t,u} \cup \{v^*\}$      /* append $v^*$ to
             $\mathbf{w}_{t,u}$ */
13         $\widetilde{\mathbf{v}} \leftarrow \widetilde{\mathbf{v}}\backslash\{v^*\}$   /* remove $v^*$ from $\widetilde{\mathbf{v}}$ */
14         for $z = 1, \ldots, Z$ do
15             $s_{z|t,u} \leftarrow s_{z|t,u} + \frac{P(v^*|t,u)}{\sum_{z'=1}^{Z} P(v^*|t,z')}$
```

After the Gibbs sampling is done, with the fact that Dirichlet distribution is conjugate to multinomial distribution, we can conveniently infer each user's interest distribution $\boldsymbol{\theta}_{t,u}$, his collaborative interest distribution $\boldsymbol{\psi}_{t,u}$ and the words' topic distribution $\boldsymbol{\phi}_{t,z}$ at t, respectively as: $\theta_{t,u,z} = \frac{m_{t,u,z}+\alpha_{t,u,z}}{\sum_{z'=1}^{Z} m_{t,u,z'}+\alpha_{t,u,z'}}$, $\psi_{t,u,z} = \frac{o_{t,u,z}+\beta_{t,u,z}}{\sum_{z'=1}^{Z} o_{t,u,z'}+\beta_{t,u,z'}}$, and $\phi_{t,z,v} = \frac{n_{t,z,v}+\gamma_{t,z,v}}{\sum_{v'=1}^{V} n_{t,z,v'}+\gamma_{t,z,v'}}$.

Streaming Keyword Diversification Model

After we obtain $\boldsymbol{\theta}_{t,u}$, $\boldsymbol{\psi}_{t,u}$ and $\boldsymbol{\phi}_{t,z}$, inspired by PM-2 diversification method (Dang and Croft 2012), we closely follow the work in (Liang et al. 2017c; 2018) and propose a streaming keyword diversification model (i.e., Algorithm 2), SKDM. To generate top-k diversified keywords for each user u at t, SKDM starts with an empty keyword set $\mathbf{w}_{t,u}$ with k empty seats (step 2 of Algorithm 2), and a set of candidate keywords (step 3), $\widetilde{\mathbf{v}}$, which is the whole words $\mathbf{v}$ in the vocabulary, i.e., initially let $\widetilde{\mathbf{v}} = \mathbf{v}$. For each of the seats, it computes the quotient $qt[z|t,u]$ for each topic z given a user u at t by the Sainte-Laguë formula (step 9): $qt[z|t,u] = \frac{\delta_{t,u,z}}{2s_{z|t,u}+1}$, where $\delta_{t,u,z}$ is the final probability of the user u has interest on topic z at t and is set to be $\delta_{t,u,z} = (1-\lambda)P(z|t,u) + \lambda P(z|t,\mathbf{f}_{t,u})$ (step 5), and $s_{z|t,u}$ is the "number" of seats occupied by topic z (in initialization, let $s_{z|t,u} = 0$ for all topics (step 6)). Here $P(z|t,u)$ and $P(z|t,\mathbf{f}_{t,u})$ are the probabilities of user u's own and his collaborative interest on topic z at t, respectively. Obviously, we can obtain $P(z|t,u)$ and $P(z|t,\mathbf{f}_{t,u})$ by our CITM algorithm such that $P(z|t,u) = \theta_{t,u,z}$ and $P(z|t,\mathbf{f}_{t,u}) = \psi_{t,u,z}$, i.e., we have:

$$\boldsymbol{\delta}_{t,u} = (1-\lambda)\boldsymbol{\theta}_{t,u} + \lambda\boldsymbol{\psi}_{t,u}, \qquad (7)$$

where $\boldsymbol{\delta}_{t,u} = \{\delta_{t,u,z}\}_{z=1}^{Z}$ is user u's final interest distributions inferred based on his own and his collaborative information at time t. According to the Sainte-Laguë method, seats should be awarded to the topic with the largest quotient in order to best maintain the proportionality of the result list. Therefore, our SKDM assigns the current seat to the topic z^* with the largest quotient (step 10). The keyword to fill this seat is the one that is not only relevant to topic z^* but to other topics and should be specific to the user, and thus we propose to obtain the keyword v^* for user u's profiling at t as (step 11): $v^* \leftarrow \arg\max_{v \in \widetilde{\mathbf{v}}} \eta_1 \times qt[z^*|t,u] \times P(v|t,z^*) + \eta_2 \times \left(\sum_{z \neq z^*} qt[z|t,u] \times P(v|t,z)\right) + (1 - \eta_1 - \eta_2) \times \mathrm{tfidf}(v|t,u)$, where $0 \leq \eta_1, \eta_2 \leq 1$ are two free parameters that satisfy $0 \leq \eta_1 + \eta_2 \leq 1$, $P(v|t,z)$ is the probability that v is associated with topic z at time t and thus can be set to be $P(v|t,z) = \phi_{t,z,v}$, and $\mathrm{tfidf}(v|t,u)$ is a time-sensitive term frequency-inverse document frequency function for user u at t, which can be defined as:

$$\mathrm{tfidf}(v|t,u) = \mathrm{tf}(v|\mathbf{d}_{t,u}) \times \mathrm{idf}(v|u,\mathbf{d}_t), \tag{8}$$

where $\mathrm{tf}(v|\mathbf{d}_{t,u}) = \frac{|\{d \in \mathbf{d}_{t,u} : v \in d\}|}{|\mathbf{d}_{t,u}|}$ is a term frequency function that computes how many percents of the documents that contain the word v in the whole document set $\mathbf{d}_{t,u}$, and $\mathrm{idf}(v|u,\mathbf{d}_t) = \log \frac{|\mathbf{d}_t|}{|\{d \in \mathbf{d}_t : v \in d\}| + \epsilon}$ is an inverse document frequency function with ϵ being set to 1 to avoid the division-by-zero error. According to (8), if v frequently appears in the document set $\mathbf{d}_{t,u}$ generated by user u but not frequently appears in the document set $\mathbf{d}_t$ generated by all the users, $\mathrm{tfidf}(v|t,u)$ will return a high score. After the word v^* is selected, SKDM adds v^* as a result keyword to $\mathbf{w}_{t,u}$, i.e., $\mathbf{w}_{t,u} \leftarrow \mathbf{w}_{t,u} \cup \{v^*\}$ (step 12), removes it from the candidate word set $\widetilde{\mathbf{v}}$, i.e., $\widetilde{\mathbf{v}} \leftarrow \widetilde{\mathbf{v}} \backslash \{v^*\}$ (step 13), and increases the "number" of seats occupied by each of the topics z by its normalized relevance to v^* as (step 15): $s_{z|t,u} \leftarrow s_{z|t,u} + \frac{P(v^*|t,u)}{\sum_{z'=1}^{Z} P(v^*|t,z')}$. The process (steps 7 to 15) repeats until we get k diversified keywords. The order in which a keyword is appended to $\mathbf{w}_{t,u}$ determines its ranking for the profiling. After the process is done, we obtain a set of diversified keywords $\mathbf{w}_{t,u}$ that profile the user u at t.

Experimental Setup

Research Questions

The research questions guiding the remainder of the paper are: **(RQ1)** How does UPA perform for user profiling compared to state-of-the-art methods? **(RQ2)** How does the contribution of the proposed interest tracking topic model, CITM, to the overall performance of UPA? **(RQ3)** What is the contribution of the collaborative information for user profiling? **(RQ4)** What is the impact of the length of the time intervals, $t_i - t_{i-1}$, in UPA?

Dataset

We work with a dataset collected from Twitter.[1] It contains 1,375 active randomly selected users and their tweets posted from the beginning of their registrations up to May 31, 2015. According to the statistics, most of the users are being followed by 2 to 50 followers. In total, we have 7.52 million tweets with timestamps including those from users' followees'. The average length of the tweets is 12 words.

We use this dataset as our stream of short documents. We obtain two categories of **G**round **T**ruths: one for evaluating **R**elevance-oriented (RGT) performance and another for evaluating **D**iversity-oriented (DGT) performance. To create the RGT ground truth, we split the dataset into 5 different partitions of time periods, i.e., a week, a month, a quarter, half a year and a year, respectively. For each Twitter user at every specific time period, an annotator was asked to generate a ranked list of top-k relevant keywords (k were decided by the annotators) as the user's profile. In total, 68 annotators took part in the labelling with each of them labelled about 5 Twitter user for these 5 different partitions. To create the ground truth for diversity evaluation, DGT, as it is expensive to manually obtain aspects of the keywords from annotators, we cluster the relevant keywords with their embeddings[2] into 15 categories [3] by k-means (MacQueen 1967). Relevant keywords within a cluster are regarded as being relevant to the same aspect in the DGT ground truth.

Baselines

We make comparisons between our UPA and the following state-of-the-art baseline algorithms: (1) **tfidf.** It simply utilizes (8), i.e., the content of users' documents to retrieve top-k keywords as profiles for the users. (2) **Predictive Language Model (PLM).** It models the dynamics of personal interests via a probabilistic language model (Fang and Godavarthy 2014). (3) **Latent Dirichlet Allocation (LDA).** This model (Blei, Ng, and Jordan 2003) infers topic distributions specific to each document via the LDA model. (4) **Author Topic model (AuthorT).** This model (Rosen-Zvi et al. 2004) infers topic distributions specific to each user in a static dataset. (5) **Dynamic Topic Model (DTM).** This dynamic model (Blei and Lafferty 2006) utilizes a Gaussian distribution for inferring topic distribution of long documents in streams. (6) **Topic over Time model (ToT).** This dynamic model (Wang and McCallum 2006) normalizes timestamps of long documents in a collection and then infers topics distribution for each document. (7) **Topic Tracking Model (TTM).** This dynamic model (Iwata et al. 2009) captures the dynamic topic distributions of long documents arriving at time t in streams of long documents. (8) **GSDMM.** This is a Gibbs Sampling-based Dirichlet Multinomial Mixture model that assigns one topic for each short document in a static collection (Yin and Wang 2014).

For fair comparisons, the topic model baselines, GSDMM, TTM, ToT, DTM and LDA, use both each user's interests $\boldsymbol{\theta}_{t,u}$ and their collaborative interests for profiling. As these baselines can not directly infer collaborative interest distributions, we use the average interests of the user's

[1] Crawled from https://dev.twitter.com/.

[2] Publicly available from https://nlp.stanford.edu/projects/glove/.

[3] Information of the categories is available from http://dmoztools.net.

followees as his collaborative interest distribution. Thus, unlike (7), in these baselines we use the mixture interests $\boldsymbol{\delta}_{t,u} = (1-\lambda)\boldsymbol{\theta}_{t,u} + \lambda \frac{1}{|\mathbf{f}_{t,u}|}\sum_{u' \in \mathbf{f}_{t,u}} \boldsymbol{\theta}_{t,u'}$ for representing each user's final interest distribution with $\boldsymbol{\theta}_{t,u}$ being inferred by the corresponding baseline topic models. The baselines, tfidf, PLM and AuthorT, are static profiling algorithms, while the others are dynamic. Again, for fair comparisons, UPA and all the other topic models use our SKDM algorithm to obtain the top-k keywords. We set the number of topics $Z = 20$ in all the topic models. For tuning parameters, λ, η_1 and η_2, we use a 70%/20%/10% split for our training, validation and test sets, respectively. The train/validation/test splits are permuted until all users were chosen once for the test set. We repeat the experiments 10 times and report the average results.

For further analysis of the contribution of collaborative interests $\boldsymbol{\psi}_{t,u}$ inferred by our CITM model to the profiling, we use another baseline denoted as UPA$_{\text{avg}}$, in which $\boldsymbol{\delta}_{t,u}$ is set to be $(1-\lambda)\boldsymbol{\theta}_{t,u} + \lambda \frac{1}{|\mathbf{f}_{t,u}|}\sum_{u' \in \mathbf{f}_{t,u}} \boldsymbol{\theta}_{t,u'}$ with $\boldsymbol{\theta}_{t,u}$ being inferred by CITM. Note that we still denote the proposed profiling algorithm using (7) as UPA.

Evaluation Metrics

We use standard relevance-oriented evaluation metrics, Pre@k (Precision at k), NDCG@k (Normalized Discounted Cumulative Gain at k), MRR@k (Mean Reciprocal Rank at k), and MAP@k (Mean Average precision at k) (Croft, Metzler, and Strohman 2015), and diversity-oriented metrics, Pre-IA@k (Intent-Aware Pre@k) (Agrawal et al. 2009), α-NDCG@k (Clarke et al. 2008), MRR-IA@k (Agrawal et al. 2009), MAP-IA@k (Agrawal et al. 2009). We also propose semantic versions of the original metrics, denoted as Pre-S@k, NDCG-S@k, MRR-S@k, MAP-S@k, Pre-IA-S@k, α-NDCG-S@k, MRR-IA-S@k, and MAP-IA-S@k, respectively. Here the only difference between the original metrics and the corresponding semantic ones is the way to compute the relevance score of a retrieval keyword v^* to ground truth keyword v_{gt}. For original metrics, we let the relevance score be 1 if and only if $v^* = v_{gt}$, otherwise be 0; whereas for the semantic versions, we let the relevance score be the cosine similarity between the word embedding vectors of v^* and v_{gt}. Since we usually choose not too many keywords to describe a user's profile, we compute the scores at depth 10, i.e., let $k = 10$. For all the metrics we abbreviate $M@k$ as M, where M is one of the metrics.

Results and Discussions

In this section, we analyse our experimental results.

Overall Performance

We start by answering research question **RQ1**. The following findings can be observed from Tables 1 and 2: (1) In terms of both relevance and diversity, all the topic model-based profiling algorithms, i.e., UPA, UPA$_{\text{avg}}$, GSDMM, ToT, TTM, DTM, AuthorT and LDA, outperform traditional algorithms, i.e., PLM and tfidf, which demonstrates that topic modeling does help to profile users' interests. (2) UPA and UPA$_{\text{avg}}$ outperform all the baseline models in terms of relevance and diversity on all the metrics, which confirms the effectiveness of the proposed user profiling algorithm for the task. (3) The ordering of the methods, UPA $>$ UPA$_{\text{avg}}$ $>$ GSDMM $>$ ToT $\sim$ TTM $\sim$ DTM $\sim$ AuthorT $\sim$ LDA $>$ PLM $>$ tifdf, is mostly consistent across the two ground truths and on the relevance and diversity evaluation metrics. Here A $>$ B denotes that method A statistically significantly performs better than method B and A $\sim$ B denotes that we did not observe a significant difference between A and B. This, once again, confirms that UPA and its averaged version, UPA$_{\text{avg}}$, outperform all the baselines. (4) In most cases, UPA $>$ UPA$_{\text{avg}}$ holds, which confirms that the collaborative information inferred by the proposed topic model, CITM, does help to improve the profiling performance.

Table 1: *Relevance* performance of UPA, UPA$_{\text{avg}}$ and the baselines using time periods of each month. Statistically significant differences between UPA$_{\text{avg}}$ and GSDMM, and between UPA and UPA$_{\text{avg}}$ are marked in the upper right hand corner of UPA$_{\text{avg}}$'s and UPA scores, respectively. Statistical significance is tested using a two-tailed paired t-test and is denoted using $^\blacktriangle$ for $\alpha = .01$, and $^\triangle$ for $\alpha = .05$.

	Pre	NDCG	MRR	MAP	Pre-S	NDCG-S	MRR-S	MAP-S
tfidf	.254	.229	.375	.135	.409	.392	.853	.203
PLM	.273	.239	.668	.140	.417	.398	.870	.212
LDA	.281	.252	.674	.142	.424	.407	.878	.217
AuthorT	.288	.260	.674	.145	.429	.408	.897	.220
DTM	.295	.270	.694	.153	.436	.419	.883	.226
TTM	.301	.276	.728	.156	.440	.426	.882	.228
ToT	.312	.283	.744	.158	.445	.428	.884	.230
GSDMM	.321	.301	.746	.163	.452	.437	.891	.236
UPA$_{\text{avg}}$	.367$^\blacktriangle$	.361$^\blacktriangle$	.840$^\blacktriangle$	.195$^\blacktriangle$	.483$^\blacktriangle$	.468$^\blacktriangle$	.939$^\blacktriangle$	.262$^\blacktriangle$
UPA	.399$^\blacktriangle$	.398$^\blacktriangle$	.860$^\blacktriangle$	.211$^\blacktriangle$	.501$^\blacktriangle$	.490$^\blacktriangle$	.946$^\triangle$	.274$^\blacktriangle$

Table 2: *Diversification* performance of UPA, UPA$_{\text{avg}}$ and the baselines using time periods of every month. Notational conventions for the statistical significances are as in Table 1.

	Pre	α-ND	MRR	MAP	Pre	α-ND	MRR	MAP
	-IA	CG	-IA	-IA	-IA-S	CG-S	-IA-S	-IA-S
tfidf	.157	.187	.480	.185	.257	.325	.725	.150
PLM	.162	.192	.487	.187	.265	.332	.742	.152
LDA	.171	.203	.493	.192	.272	.338	.744	.155
AuthorT	.174	.205	.505	.195	.276	.343	.748	.157
DTM	.177	.206	.507	.197	.279	.347	.748	.159
TTM	.180	.208	.509	.221	.282	.351	.751	.162
ToT	.182	.213	.513	.225	.290	.355	.754	.170
GSDMM	.194	.228	.525	.237	.304	.368	.780	.173
UPA$_{\text{avg}}$	.238$^\blacktriangle$	.265$^\blacktriangle$	.597$^\blacktriangle$	.252$^\blacktriangle$	.362$^\blacktriangle$	.421$^\blacktriangle$	.808$^\blacktriangle$	.216$^\blacktriangle$
UPA	.266$^\blacktriangle$	.302$^\blacktriangle$	.623$^\blacktriangle$	.266$^\blacktriangle$	.395$^\blacktriangle$	.452$^\blacktriangle$	.814$^\triangle$	.231$^\blacktriangle$

Additionally, Table 3 shows the top six keywords of an example user's dynamic profile with time being five quarters from April 2014 to May 2015. As shown in the table, the diversified keywords generated by UPA are semantically closer to those from the ground truth compared to those generated by the baseline, GSDMM, which again demonstrates the effectiveness of the proposed UPA algorithm.

Contribution of CITM

We now turn to answer research question **RQ2**. Recall that the only difference between our UPA/UPA$_{\text{avg}}$ and the baselines is that UPA utilizes our CITM to track users' dynamic

Table 3: Top six keywords of an example user's dynamic profile with the time being five quarters from April 2014 to May 2015. The keywords from the DGT ground truth, generated from GSDMM and UPA are presented for the user, respectively.

	Apr. 2014 to Jun. 2014	Jul. 2014 to Sep. 2014	Oct. 2014 to Dec. 2014	Jan. 2015 to Mar. 2015	Apr. 2015 to May 2015
Ground Truth	Apple Java iPhone Python ApplePay OjectiveC	Apple Git iPad OjectiveC AppleEvent Python	AppleEvent LininProfile openEducation iOS NatsTwitter Education	Microblog Students LinkedInProfile ArtsEducation FB AfterSchool	SocialMedia Education NatsTwitter ConnectedLearning FB Courses
GSDMM	Apple Computer iPhone Science Java Technology	Apple Company University Technology iPad Language	Apple Christmas LinkedIn Education iOS Friends	Online Education Students Website Degree Presentation	Courses Online Presentation Digital Learning Education
UPA	Apple Java iPhone Programming CPlusPlus Computer	Apple Programming iPad Git Event Python	Apple LinkedIn Education iOS Twitter Education	LinkedIn Students Microblog Education FB Art	Education Media Learning FB Courses Twitter

interests and then our SKDM to diversify the keywords, whereas other topic models utilize different topic models to obtain users' interests and then the SKDM for keyword diversification. As in Tables 1 and 2, UPA/UPA$_{\text{avg}}$ outperforms all the topic models, i.e., GSDMM, ToT, TTM, DTM AuthorT and LDA, which illustrates that the proposed topic model, CITM, does be effective and has significant contribution to the performance of our user profiling algorithm.

Contribution of Collaborative Interests

Here we turn to answer research question **RQ3**. We vary the parameter λ that governs how much the collaborative information, $\psi_{t,u}$, are utilized for profiling. A larger λ indicates more collaborative information is utilized for the profiling.

Fig. 2 shows the performance on the relevance and diversity evaluation metrics (use Precision and Pre-IA as representative metrics only), where we use the best baseline, GSDMM, as a representative. When we increase λ from 0 to 0.6, i.e., giving more weight to the collaborative information, the performance of all the models gradually improves, with UPA still outperforming UPA$_{\text{avg}}$ and GSDMM. This, again, illustrates that integrating collaborative information into the models helps to improve the performance. Moreover, as shown in Fig. 2, UPA that utilizes collaborative interests outperforms UPA$_{\text{avg}}$ that simply utilizes the average of the followees' interests as its collaborative interests, which once again demonstrates that the inferred collaborative interests in UPA is effective.

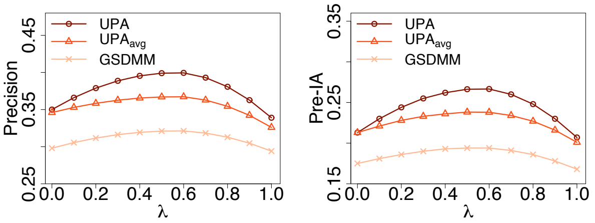

Figure 2: Relevance and diversity performance of UPA, UPA$_{\text{avg}}$ and GSDMM on representative metrics, Precision and Pre-IA, with varying scores of λ, respectively.

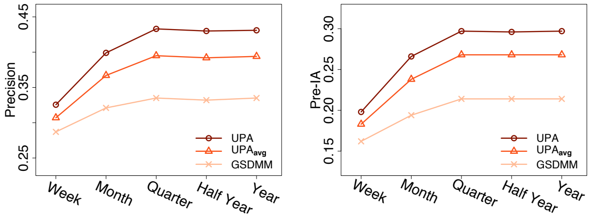

Figure 3: Relevance and diversity performance of UPA, UPA$_{\text{avg}}$ and GSDMM on time periods of a week, a month, a quarter, half a year, and a year, respectively.

Impact of Time Period Length

Finally, we answer research question **RQ4**. We compare the performance for different time periods, a week, a month, a quarter, half a year and a year, respectively, using the two ground truths, RGT and DGT, on the representative relevance and diversity metrics, Precision and Pre-IA, in Fig. 3.

As is shown in Fig. 3, UPA and UPA$_{\text{avg}}$ beat the baselines for time periods of all lengths, which illustrates that our proposed user profiling algorithm works better than the state-of-the-art ones for dynamic user profiling regardless of period length. The performance of UPA, UPA$_{\text{avg}}$ and the best baseline, GSDMM, improves significantly on all the metrics when the period length increases from a week to a quarter, whereas it reaches a plateau as the time periods further increase from a quarter to a year. In all the cases UPA and UPA$_{\text{avg}}$ significantly outperform the best baseline, GSDMM. These findings illustrate the fact that the performance of the proposed algorithms is robust and is able to maintain significant improvements over the state-of-the-art non-dynamic and dynamic algorithms. In addition, UPA always outperforms UPA$_{\text{avg}}$ on all the metrics and all the different period lengths, which once again illustrates that the collaborative interest distribution inferred by the proposed CITM model helps to enhance the user profiling performance.

Conclusions

We have studied the problem of collaborative, dynamic and diversified user profiling in the context of streams of short texts. To tackle the problem, we have proposed a streaming profiling algorithm, UPA, that consists of two models: the proposed collaborative interest tracking topic model, CITM, and the proposed streaming keyword diversification model, SKDM. Our CITM tracks the changes of users' and their followees' interest distribution in streams of short texts, a sequentially organized corpus of short texts, and our SKDM diversifies the top-k keywords for profiling users' dynamic interests. To effectively infer users' and their followees' dynamic interest distribution in our CITM model, we have proposed a collapsed Gibbs sampling algorithm, where during the sampling one single topic is assigned to a document to address the textual sparsity problem. We have conduced experiments on a Twitter dataset. We evaluated the performance of our UPA and the baseline algorithms using two categories of ground truths on both the original metrics and the proposed semantic versions of the metrics. Experimental results show that our UPA is able to profile users' dynamic interests over time for streams of short texts. In the future, we intend to utilize auxiliary resources such as Wikipedia articles that the entities in the short documents link to for further improvement of user profiling.

References

Agrawal, R.; Gollapudi, S.; Halverson, A.; and Ieong, S. 2009. Diversifying search results. In *WSDM*, 5–14.

Balog, K., and de Rijke, M. 2007. Determining expert profiles (with and application to expert finding). In *IJCAI*, 2657–2662.

Balog, K.; Bogers, T.; Azzopardi, L.; de Rijke, M.; and van den Bosch, A. 2007. Broad expertise retrieval in sparse data environments. In *SIGIR*, 551–558.

Balog, K.; Fang, Y.; de Rijke, M.; Serdyukov, P.; and Si, L. 2012. Expertise retrieval. *Found. Trends Inf. Retr.* 6:127–256.

Berendsen, R.; Rijke, M.; Balog, K.; Bogers, T.; and Bosch, A. 2013. On the assessment of expertise profiles. *JAIST* 64(10):2024–2044.

Blei, D. M., and Lafferty, J. D. 2006. Dynamic topic models. In *ICML*, 113–120.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Clarke, C. L. A.; Kolla, M.; Cormack, G. V.; Vechtomova, O.; Ashkan, A.; Büttcher, S.; and MacKinnon, I. 2008. Novelty and diversity in information retrieval evaluation. In *SIGIR*, 659–666.

Craswell, N.; de Vries, A. P.; and Soboroff, I. 2005. Overview of the TREC 2005 enterprise track. In *TREC'05*, 1–7.

Croft, W. B.; Metzler, D.; and Strohman, T. 2015. *Search engines: Information retrieval in practice*. Addison-Wesley Reading.

Dang, V., and Croft, W. B. 2012. Diversity by proportionality: An election-based approach to search result diversification. In *SIGIR*, 65–74.

Fang, Y., and Godavarthy, A. 2014. Modeling the dynamics of personal expertise. In *SIGIR*, 1107–1110.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *PNAS* 101:5228–5235.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *SIGIR*, 50–57.

Iwata, T.; Watanabe, S.; Yamada, T.; and Ueda, N. 2009. Topic tracking model for analyzing consumer purchase behavior. In *IJCAI*, volume 9, 1427–1432.

Iwata, T.; Yamada, T.; Sakurai, Y.; and Ueda, N. 2010. Online multiscale dynamic topic models. In *KDD*, 663–672. ACM.

Liang, S.; Ren, Z.; Yilmaz, E.; and Kanoulas, E. 2017a. Collaborative user clustering for short text streams. In *AAAI*, 3504–3510.

Liang, S.; Ren, Z.; Zhao, Y.; Ma, J.; Yilmaz, E.; and Rijke, M. D. 2017b. Inferring dynamic user interests in streams of short texts for user clustering. *ACM Trans. Inf. Syst.* 36(1):10:1–10:37.

Liang, S.; Yilmaz, E.; Shen, H.; Rijke, M. D.; and Croft, W. B. 2017c. Search result diversification in short text streams. *ACM Trans. Inf. Syst.* 36(1):8:1–8:35.

Liang, S.; Zhang, X.; Ren, Z.; and Kanoulas, E. 2018. Dynamic embeddings for user profiling in twitter. In *KDD*, 1764–1773.

Liang, S.; Yilmaz, E.; and Kanoulas, E. 2018. Collaboratively tracking interests for user clustering in streams of short texts. *IEEE Transactions on Knowledge and Data Engineering*.

Liang, S. 2018. Dynamic user profiling for streams of short texts. In *AAAI*, 5860–5867.

Liu, J. S. 1994. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Am. Stat. Assoc.* 89(427):958–966.

MacQueen, J. B. 1967. Some methods for classification and analysis of multivariate observations.

Minka, T. 2000. Estimating a dirichlet distribution.

Rosen-Zvi, M.; Griffiths, T.; Steyvers, M.; and Smyth, P. 2004. The author-topic model for authors and documents. In *UAI*, 487–494.

Rybak, J.; Balog, K.; and Nørvåg, K. 2014. Temporal expertise profiling. In *ECIR*, 540–546.

Wallach, H. M. 2006. Topic modeling: beyond bag-of-words. In *ICML*, 977–984.

Wang, X., and McCallum, A. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *KDD*, 424–433.

Wei, X.; Sun, J.; and Wang, X. 2007. Dynamic mixture models for multiple time-series. In *IJCAI*, 2909–2914.

Yin, J., and Wang, J. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *KDD*, 233–242.