Collaborative, Dynamic and Diversified User Profiling

Shangsong Liang\textsuperscript{1,2}
\textsuperscript{1}School of Data and Computer Science, Sun Yat-sen University, China
\textsuperscript{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 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 realizes 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:

\[
D_t, u_t \xrightarrow{h} W_t,
\]

where \( D_t = \{\ldots, d_{t-2}, d_{t-1}, d_t\} \) represents the stream of short documents generated by the users \( u_t \) up to time \( t \) with \( d_t \) being the most recent set of short documents arriving at \( t \), \( u_t = \{u_1, u_2, \ldots, u_{|u_t|}\} \) represents a set of users appearing in the stream up to time \( t \), with \( u_i \) being the \( i \)-th user in \( u_t \) and \( |u_t| \) being the total number of users in the user set, and \( W_t = \{w_{t,u_1}, w_{t,u_2}, \ldots, w_{t,u_{|u_t|}}\} \) represents all users’ profiling results at \( t \) with \( w_{t,u_i} = \{w_{t,u_i,1}, w_{t,u_i,2}, \ldots, w_{t,u_i,k}\} \) being the profiling result, i.e., the top- \( k \) diversified keywords, for user \( u_t \) at \( t \). We assume that the length of a document \( d \) in \( D_t \) 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 \( u \) at time period \( t \) can be represented as a multinomial distribution \( \theta_{t,u} \) over topics, where \( \theta_{t,u} = \{\theta_{t,u,z}\}_{z=1}^{Z} \) 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 \( \psi_{t,u} = \{\psi_{t,u,z}\}_{z=1}^{Z} \) with \( \psi_{t,u,z} \) being the interest score of user \( u \)’s followees \( f_{t,u} \) as a whole on topic \( z \) at \( t \). Here, \( f_{t,u} \) 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 \( \theta_{t,u} \) and his collaborative interests \( \psi_{t,u} \). (2) Given \( \theta_{t,u} \) and \( \psi_{t,u} \) 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 \( \theta_{t,u} = \{\theta_{t,u,z}\}_{z=1}^{Z} \) and his collaborative interest distribution \( \psi_{t,u} = \{\psi_{t,u,z}\}_{z=1}^{Z} \) 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 \( \theta_{t,u} \) 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 \( \alpha_t = \{\alpha_{t,z}\}_{z=1}^{Z} \), where we let the mean of the current distribution \( \theta_{t,u} \) depend on the mean of the previous distribution \( \theta_{t-1,u} \) as:

\[
P(\theta_{t,u} | \theta_{t-1,u}, \alpha_t) \propto \prod_{z=1}^{Z} \alpha_{t,z} \theta_{t-1,u,z}^{-1},
\]

where the precision value \( \alpha_{t,z} = \{\alpha_{t,u,z}\}_{u=1}^{u=|u_t|} \) 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 \( \psi_{t,u} \), we use
we propose a collapsed Gibbs sampling algorithm to infer their corresponding dynamic collaborative interest distributions of the previous distribution \( \phi_{t-1,u,z} \):  
\[
P(\phi_{t,z}|\phi_{t-1,z}, \gamma_t) \propto \prod_{v=1}^{V} \phi_{t,v}^{\gamma_{t,v}} \phi_{t-1,u,v}^{-1},
\]
where \( V \) is the total number of words in a vocabulary \( v = \{v_i\}_{i=1}^{V} \) and \( \gamma_t = \{\gamma_{t,v}\}_{v=1}^{V} \). For each document \( d \in d_t \), we can easily integrate out the uncertainty associated with multinomials \( \phi_{t,u,z} \) and \( \phi_{t,z} \).

**Algorithm 1:** Inference for our CITM model at time \( t \).

**Input:** Distributions \( \Theta_{t-1}, \Psi_{t-1} \) and \( \Phi_{t-1} \) at \( t-1 \); number of iterations \( N_{iter} \).

**Output:** Current distributions \( \Theta_t, \Psi_t \), and \( \Phi_t \).

1. Initialize topic assignments randomly for all documents in \( d_t \).

2. **for iteration** = 1 to \( N_{iter} \) **do**

   3. **for** user = 1 to \( |u_t| \) **do**

      4. **for** \( d \) = 1 to \( d_{t,u} \) **do**

       5. Draw \( z_{t,u,d} \) from (5)

       6. Update \( m_{t,u,z_{t,u,d}}, \alpha_{t,u,z_{t,u,d}} \) and \( n_{t,u,z_{t,u,d}} \).

       7. Update \( \alpha_t, \beta_t, \gamma_t \).

8. Compute the posterior estimates \( \Theta_t, \Psi_t \), and \( \Phi_t \).

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 \( \theta_{t,u}, \psi_{t,u} \), and \( \phi_{t,z} \).

We provide an overview of our proposed collapsed Gibbs sampling algorithm in Algorithm 1, where we denote \( m_{t,u,z}, \alpha_{t,u,z} \), and \( n_{t,u,z} \) 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 \( P(\phi_{t,z} | \phi_{t-1,z}, \gamma_t, \theta_{t,u,z}, \psi_{t,u}) \) for all topics assigned to user \( u \) at \( t \), and \( z_{t,u,d} \) is the topic assigned to document \( d \) by user \( u \) at \( t \).
assigned to the document \( d \in \mathcal{D}_{t,u} \). For obtaining this conditional distribution used during sampling, we begin with the joint probability of the current document set, \( P(z_t, d_{t}, \Theta_{t-1}, \Psi_{t-1}, \Phi_{t-1}, u_t, \alpha_t, \beta_t, \gamma_t) \) at time \( t \):

\[
P(z_t, d_{t}, \Theta_{t-1}, \Psi_{t-1}, \Phi_{t-1}, u_t, \alpha_t, \beta_t, \gamma_t) = (1 - \lambda) P(z_t|d_t, \Theta_{t-1}, \Phi_{t-1}, u_t, \alpha_t, \beta_t, \gamma_t) + \lambda P(z_t|d_t, \Theta_{t-1}, \Phi_{t-1}, u_t, \alpha_t, \beta_t, \gamma_t)
\]

where \( \Gamma(\cdot) \) is a gamma function, \( \lambda \) is a free parameter that governs the linear mixture of a user’s interests and his followers’ interests, and the set of parameters \( \{\Gamma(\cdot)\} \) are defined as:

\[
\begin{align*}
\alpha_t &= m_t, z_t + \alpha_t, \beta_t, \gamma_t \quad \nu_t = n_t, z_t - \nu_t, \phi_t
\end{align*}
\]

where \( m_t, z_t \) and \( n_t, z_t \) are the number of documents \( d \) and the number of words \( w \) assigned in \( d \), \( \nu_t \) is the number of words assigned to \( d \) 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 \( \alpha_t, \beta_t \) and \( \gamma_t \) can be estimated by maximizing the joint distribution (4). We apply fixed-point iterations to obtain the optimal \( \alpha_t, \beta_t \) and \( \gamma_t \). By applying the two bounds in (Minka 2000), we can derive the following update rules of \( \alpha_t, \beta_t \) and \( \gamma_t \) for maximizing the joint distribution in our fixed-point iterations:

\[
\begin{align*}
\alpha_{t,u,z} &= \frac{(1 - \lambda) \alpha_{t,u,z} + \beta_t, \gamma_t, \lambda \sum_{j=1}^{N_t} \psi_{t,u,z} \nu_{t,u,z} - \Delta(\alpha_{t,u,z})}{\Delta(\sum_{j=1}^{N_t} \psi_{t,u,z} \nu_{t,u,z} - \Delta(\beta_t, \gamma_t))}
\end{align*}
\]

where \( \Delta(x) = \frac{d \log(\Gamma(x))}{dx} \) is a Digamma function. Our derivations of the update rules for \( \alpha_t, \beta_t \) and \( \gamma_t \) in (6) are analogous to those in (Liang, Yilmaz, and Kanoulas 2018; Liang et al. 2017c; 2017b).

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 \( \theta_{t,u} \), his collaborative interest distribution \( \psi_{t,u} \) and the

### Algorithm 2: SKDM model for generating top-\( k \) keywords for collaborative, dynamic, diversified user profiling

**Input:** Current distributions \( \Theta_t \) and \( \Phi_t \)

**Output:** All users’ profiling results at time \( t, W_t \)

1. for \( u = 1, \ldots, |u| \) do
   2. \( w_{t,u} \leftarrow \emptyset \)
   3. \( v \leftarrow v \)
   4. for \( z = 1, \ldots, Z \) do
      5. \( z_t \leftarrow \operatorname{argmax}_z q_t[z|t, u] \times P(v|t, z) + (1 - \mu - \nu) \times \text{tfidf}(v|t, u) \)
      6. \( w_{t,u} \leftarrow w_{t,u} \cup \{v\} \)
      7. \( z_t \leftarrow z_t^* \) /* remove \( z^* \) from \( w_{t,u} \)
   8. for \( z = 1, \ldots, Z \) do
      9. \( s'[z|t, u] \leftarrow s[z|t, u] + P(v'|t, u) \sum_{z} P(v'|t, z) \)

### Streaming Keyword Diversification Model

After we obtain \( \theta_{t,u}, \psi_{t,u} \) and \( \phi_{t,z} \), 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 \( w_{t,u} \) with \( k \) empty seats (step 2 of Algorithm 2), and a set of candidate keywords (step 3), \( \tilde{v} \), which is the whole words \( v \) in the vocabulary, i.e., initially let \( \tilde{v} = v \). For each of the seats, it computes the quotient \( q_t[z|t, u] \) for each topic \( z \) given a user \( u \) at \( t \) by the Sainte-Lagué formula (step 9):

\[
q_t[z|t, u] = \frac{\delta_{t,u,z}}{\delta_{t,u,z} + \nu_t z_t^*} \times \psi_{t,u,z}, \quad \psi_{t,u,z} = \frac{m_t, z_t + \alpha_t, \beta_t, \gamma_t}{\sum_{z'} m_t, z_t + \alpha_t + \beta_t, \gamma_t}, \quad \text{and} \quad \delta_{t,u,z} = \lambda P(z|t, u) + \lambda P(z|t, f_{t,u}) \times \text{tfidf}(z|t, u)
\]

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 \( \delta_{t,u,z} = (1 - \lambda) P(z|t, u) + \lambda P(z|t, f_{t,u}) \) (step 5), and \( s[z|t, u] = 0 \) for all topics (step 6). Here \( P(z|t, u) \) and \( P(z|t, f_{t,u}) \) 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 \( P(z|t, f_{t,u}) \) by our CITM algorithm such that \( P(z|t, u) = \theta_{t,u,z} \) and \( P(z|t, f_{t,u}) = \psi_{t,u,z} \), i.e., we have:

\[
\delta_{t,u} = (1 - \lambda) \theta_{t,u} + \lambda \psi_{t,u},
\]
where $\delta_{t,u} = \{ \delta_{t,u,z} \}_{z=1}^{Z}$ is user $u$’s final interest distributions inferred based on his own and his collaborative information at time $t$. According to the Sainete-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): $v^* \leftarrow \arg \max_{v \in \mathcal{V}} \eta_1 \times \operatorname{tfidf}(v, t, u) \times P(v | t, z^*) + \eta_2 \times \left( \sum_{z \neq z^*} q_t[z | t, u] \times P(v | t, z) \right) + (1 - \eta_1 - \eta_2) \times \operatorname{tfidf}(v, t, u)$, 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 $\operatorname{tfidf}(v, t, u)$ is a time-sensitive term frequency-inverse document frequency function for user $u$ at $t$, which can be defined as:

$$\operatorname{tfidf}(v, t, u) = \operatorname{tf}(v | d_{t,u}) \times \operatorname{idf}(v | u, d_t),$$

where $\operatorname{tf}(v | d_{t,u}) = \frac{| \{d_{t,v,z} : z \in \mathcal{Z} \} |}{|d_{t,v}|}$ is a term frequency function that computes how many percents of the documents that contain the word $v$ in the whole document set $d_{t,u}$, and $\operatorname{idf}(v | u, d_t) = \log \frac{|d_{t,u}|}{| \{d_{t,v,z} : z \in \mathcal{Z} \} |}$ is an inverse document frequency function with $\epsilon$ being set to 1 to avoid the division-by-zero error. According to (8), if $v$ frequently appears in the document set $d_{t,u}$ generated by user $u$ but not frequently appears in the document set $d_t$ generated by all the users, $\operatorname{tfidf}(v, t, u)$ will return a high score. After the word $v^*$ is selected, SKDM adds $v^*$ as a result keyword to $w_{t,u}$, i.e., $w_{t,u} \leftarrow w_{t,u} \cup \{ v^* \}$ (step 12), removes it from the candidate word set $\mathcal{V}$, i.e., $\mathcal{V} \leftarrow \mathcal{V} \setminus \{v^*\}$ (step 13), and increases the “number” of seats occupied by each of the topics $z$ by its normalized relevance to $v^*$ as (step 15):

$$s_{z,t,u} \leftarrow s_{z,t,u} + \frac{P(v^* | t, u)}{\sum_{z' \neq z} P(v^* | t, z')}.$$ 

The process (steps 7 to 15) repeats until we get $k$ diversified keywords. The order in which a keyword is appended to $w_{t,u}$ determines its ranking for the profiling. After the process is done, we obtain a set of diversified keywords $w_{t,u}$ 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 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. 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’ followers. The average length of the tweets is 12 words.

We use this dataset as our stream of short documents. We obtain two categories of Ground Truths: one for evaluating Relevance-oriented (RGT) performance and another for evaluating Diversity-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 Goadway 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.$^8$


$^3$Information of the categories is available from http://dmnotools.net.

$^8$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 $\theta_{t,u}$ 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
fallowees as his collaborative interest distribution. Thus, unlike (7), in these baselines we use the mixture interests $\delta_{i,u} = (1 - \lambda)\theta_{i,u} + \lambda \sum_{t' \in \mathcal{F}_{t,u}} \theta_{t',u}$ for representing each user’s final interest distribution with $\theta_{i,u}$ 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, $\lambda, \eta_1$ and $\eta_2$, 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 $\psi_{t,u}$ inferred by our CITM model to the profiling, we use another baseline denoted as UPA$_{avg}$, in which $\delta_{i,u}$ is set to be $(1 - \lambda)\theta_{i,u} + \lambda \sum_{t' \in \mathcal{F}_{t,u}} \theta_{t',u}$ with $\theta_{i,u}$ 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 Strohminger 2015), and diversity-oriented metrics, Pre-IA@$k$ (Intent-Aware Pre@$k$) (Agrawal et al. 2009), $\alpha$-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@$k$, $\alpha$-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$_{avg}$, 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$_{avg}$ 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$_{avg}$ $>$ GSDMM $>$ ToT $>$ TTM $>$ DTM $>$ AuthorT $>$ LDA $>$ PLM $>$ tfidf, 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$_{avg}$, outperform all the baselines. (4) In most cases, UPA $>$ UPA$_{avg}$ holds, which confirms that the collaborative information inferred by the proposed topic model, CITM, does help to improve the profiling performance.

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$_{avg}$ and the baselines is that UPA utilizes our CITM to track users’ dynamic

| Table 1: Relevance performance of UPA, UPA$_{avg}$ and the baselines using time periods of each month. Statistically significant differences between UPA$_{avg}$ and GSDMM, and between UPA and UPA$_{avg}$ are marked in the upper right hand corner of UPA$_{avg}$’s and UPA scores, respectively. Statistical significance is tested using a two-tailed paired t-test and is denoted using ▲ for $\alpha = .01$, and ▲ for $\alpha = .05$. |
|------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                  | Pre-S | NDCG-S | MRR | MAP | Pre-S | NDCG-S | MRR | MAP |
| 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 ▲ |
| ToT              | .301 ▲ | .276 ▲ | .728 ▲ | .156 ▲ | .440 ▲ | .426 ▲ | .882 ▲ | .228 ▲ |
| GSDMM            | .321 ▲ | .301 ▲ | .746 ▲ | .163 ▲ | .452 ▲ | .437 ▲ | .891 ▲ | .236 ▲ |
| UPA              | .399 ▲ | .399 ▲ | .860 ▲ | .211 ▲ | .501 ▲ | .490 ▲ | .946 ▲ | .274 ▲ |

| Table 2: Diversification performance of UPA, UPA$_{avg}$ and the baselines using time periods of every month. Notational conventions for the statistical significances are as in Table 1. |
|------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                  | Pre-IA | $\alpha$-ND | MRR | MAP | Pre-IA | $\alpha$-ND | MRR | MAP |
| 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 ▲ |
| ToT              | .180 ▲ | .208 ▲ | .509 ▲ | .221 ▲ | .282 ▲ | .351 ▲ | .751 ▲ | .162 ▲ |
| GSDMM            | .194 ▲ | .228 ▲ | .525 ▲ | .237 ▲ | .304 ▲ | .368 ▲ | .780 ▲ | .173 ▲ |
| UPA$_{avg}$      | .238 ▲ | .265 ▲ | .572 ▲ | .252 ▲ | .362 ▲ | .421 ▲ | .808 ▲ | .216 ▲ |
| UPA              | .266 ▲ | .302 ▲ | .623 ▲ | .266 ▲ | .395 ▲ | .452 ▲ | .814 ▲ | .231 ▲ |
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, \( \text{UPA}_{\text{avg}} \) outperforms all the topic models, i.e., GSDMM, ToT, TTM, DTM AuthorIT 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.

\textbf{Contribution of Collaborative Interests}

Here we turn to answer research question \textbf{RQ3}. We vary the parameter \( \lambda \) that governs how much the collaborative information, \( \psi_{t,u} \), are utilized for profiling. A larger \( \lambda \) 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 \( \lambda \) from 0 to 0.6, i.e., giving more weight to the collaborative information, the performance of all the models gradually improves, with \( \text{UPA} \) still outperforming \( \text{UPA}_{\text{avg}} \) and GSDMM. This, again, illustrates that integrating collaborative information into the models helps to improve the performance. Moreover, as shown in Fig. 2, \( \text{UPA} \) that utilizes collaborative interests outperforms \( \text{UPA}_{\text{avg}} \) that simply utilizes the average of the followees’ interests as its collaborative interests, which once again demonstrates that the inferred collaborative interests in \( \text{UPA} \) is effective.

\textbf{Impact of Time Period Length}

Finally, we answer research question \textbf{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, \( \text{UPA} \) and \( \text{UPA}_{\text{avg}} \) 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 \( \text{UPA} \), \( \text{UPA}_{\text{avg}} \) 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 \( \text{UPA} \) and \( \text{UPA}_{\text{avg}} \) 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, \( \text{UPA} \) always outperforms \( \text{UPA}_{\text{avg}} \) 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.

\textbf{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 conducted 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


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


